Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 10 Issue 01
2025

Research Article
https://doi.org/10.56979/1001/2025/1146

Bilingual Spectral Emotion Learning Through Patch-Encoded VGG-16 Features and a Full Vision Transformer Pipeline

Yogeshkumar Prajapati[1], Priyesh Gandhi[2], and Sheshang Degadwala[3*]

[1]Research Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India.
[2]Provost, Sigma University, Vadodara, Gujarat, India.
[3]Professor and Head, Department of Computer Engineering, Sigma University, Vadodara, Gujarat, India.
*Corresponding Author: Sheshang Degadwala. Email: sheshang13@gmail.com

Received: September 20, 2025 **Accepted:** November 25, 2025

Abstract: The research introduces a new bilingual speech emotion recognition system which integrates patch-encoded VGG-16 spectral cues with a whole Vision Transformer pipeline to learn the affective cues on English and Gujarati speech. Mel-spectrograms are initially inputted into a frozen VGG-16 backbone to obtain high-level spatial spectral features, and these features are then split into regular patches and transformed into an embedding space to be used to represent them in a transformer-based global attention model. It is tested on four reference English emotional speech datasets, including RAVDESS, CREMA-D, SAVEE, and TESS, with the results of accuracy 99% for all. To evaluate robustness on non-controlled data, a hand-collected bilingual corpus of student recordings was created, where the model was able to assess English and Gujarati speech with accuracy on 90% and 88% percent respectively. Such findings show that the convolutional spectral extraction with contextual learning by transformers is an effective way of modeling cross-lingual emotional differences and outperforms traditional convolution-only or transformer-only models. The bilingual results also show that the model can be used to achieve stable performance with languages that have different phonetics and prosody and thus is applicable to scalable and inclusive emotion-sensitive speech technologies in practice through interactive assistants, call-center analytics and affect sensitive human-machine interfaces.

Keywords: Speech Emotion Recognition; Spectrogram Analysis; VGG-16 Feature Extraction; Vision Transformer Modeling; Bilingual English–Gujarati Dataset

1. Introduction

Speech Emotion Recognition (SER) has been a vital part of the contemporary intelligent systems in that machines can deduce the human affective statuses directly by hearing them. The increasing use of SER in applications including affect conscious virtual assistants, medical monitoring, call-center automation, and intelligent tutoring systems has driven a great deal of study into the development of effective and precise emotion classification models. As the transition of the handcrafted acoustic features was to the deep learning-based representations, a few studies have shown to achieve notable enhancements in the reliability and generalizability of SER systems. In recent literature, the concept of both data augmentation and deep feature extraction has become relevant to realize a high level of robustness in real-time and diverse acoustic settings [1]. Furthermore, the advancement of biologically inspired and previously crossbreed algorithms has also expanded SER potential with better discrimination in the case of emotional states in noisy conditions [2].

Spectrogram analysis with deep learning models like CNNs has enjoyed a lot of use and has been very effective when it comes to modeling spatial-spectral patterns of emotions [3]. A few SER systems to address realistic deployment applications have focused on efficiency in computation by compressing audio and using lightweight audio processing methods [4]. In the case of a language that is underrepresented, data augmentation techniques were found to improve performance dramatically and overcome the lack of data [5]. Hybrid networks that combine CNN and LSTM layers are also investigated and it has been shown that a combination of spatial and time-based learning enhances the accuracy of emotion recognition [6]. Another development suggests improved embedding spaces and low-level acoustic descriptions to reinforce emotional feature representation [7], and assessments of the raw waveform modeling in order to avoid reliance on manual feature engineering [8].

Similar progress in parallel multi-microphone processing, token-semantic representations, and hierarchical audio transformers indicates the growing move to attention-based modeling [9]. MFCCs as well as CNNs and recurrent networks have also been used as ensemble architectures and demonstrated impressive results in diverse emotional datasets [10]. Also, in-depth evaluations reveal that efficient data augmentation has been a primary determinant of the deep learning in SER tasks [11]. Resource-constrained environments have also been studied with handcrafted feature lightweight ensemble models [12]. The efficiency of convolutional learning as a method of SER is also proved by studies that concentrate on the mel-spectrograms and CNNs [13].

Machine learning-based multilingual SER systems have been used to overcome the necessity of language, accent, and emotional style generalized models [14]. Accent diversity, which creates performance variance, was also another motivation that drove cross-accent SER frameworks that increase strength within groups of speakers [15]. Further applications of LSTM attention/CNN hybrid models have shown higher emotional discrimination in highly acoustical conditions [16], and graph-based audio representations have been suggested to structure emotional modeling [17]. Cross-attention based networks have enhanced multimodal and multi-resolution feature extraction methods [18] and combined CNN architectures also demonstrated the ability to capture hierarchical emotional information [19].

The latest reviews point to the important gaps in the data sets diversity, methods of features extraction and cross-language generalization [20]. Architectures that build relationships between mel-spectrogram regions have demonstrated significant progress in learning context-dependent patterns of emotion via transformer-based architecture [21]. Vision Transformer-based SER systems also emphasize the usefulness of self-attention all over the world to understand spectrograms [22], and multi-axis transformer models also show good learning in multi-scale emotional representations [23]. Temporal frequency correlation techniques, positional modeling techniques, and knowledge transfer techniques are also useful in improving the accuracy [24]. Explainable frameworks in deep transfer learning have further enhanced interpretability as well as performance on SER datasets [25].

Regardless of these improvements, there is still the under-researched multilingual and low-resource language, like Gujarati. The difference in phonetics, prosody, emotional expression and recording conditions pose more problems to multilingual SER. To fill these gaps, the given work suggests a bilingual SER model, which combines patch-encoded VGG-16 spectrogram features with a full Vision Transformer pipeline and is tested on both established English datasets and a newly created English-speaking Gujarati protestant speech corpus.

2. Related Works

The development of deep learning and spectrogram-based analysis has seen a significant advancement of Speech Emotion Recognition (SER). The first SER methods used mainly handcrafted acoustic characteristics like MFCC, pitch, energy, and prosodic indicators, which constituted the foundations of the classical machine learning models. Barhoumi et al. [1] showed the potential of using a combination of the traditional spectral features and data augmentation to enhance the stability of recognition in different acoustic conditions. Later literature took the approach of biologically inspired algorithms and multi-stage signal processing to improve emotional separability, with CNN technique [2] demonstrating significant resilience in noisy conditions.

The most popular spectrogram feature extraction methodology is now the deep convolutional neural networks (CNNs). In Author et al. [3], CNN-based systems were effective in fine-grained spatial-spectral variations in emotional speech. Being lightweight and efficient frameworks like those suggested in Author et al. [4], placed an emphasis on the minimization of computational complexity in real-time SER. In low-resource language fields, augmentation-based gains were made by means of domain expansion methods

[5], whereas hybrid CNNLSTM architectures allowed to make better temporal spatial predictions regarding emotional clues [6]. Embedding space and low-level acoustic descriptors were also optimized, and feature discriminability was reinforced [7], as well as raw waveform modeling presented in Kilimci et al. [8] eliminated manual feature engineering. Multilingual SER experiments with raw waveforms [8] have been shown to be strongly performing without the postprocessing of spectrograms; they are however sensitive to channel distortions and tend to consume large amounts of data. Conversely, our hybrid architecture with spectrogram provides more stability in different microphone-related conditions and small sample bilingual data, as the extracted spectral abstractions through VGG-16 are processed by Transformer to provide more accurate results.

The attention mechanisms have been a major force in recent advances. Semantic token representations [9], CNNRNN pipeline and augmentation-based training protocols [10] have continued to enhance generalization over datasets with multi-microphone processing. Lightweight handcrafted-feature ensembles [12] and mel-spectrogram CNN frameworks [13] can also be applied to resource-constrained applications. The necessity of multilingual SER solutions inspired the cross-language models discussed in Author et al. [14], and speech diversity cross-accent discussed in XYZ framework [15]. Multimodal and multi-resolution learning was also improved by hybrid LSTM-attention systems [16], graph-based emotional representations [17], and dual-stream cross-attention networks [18]. Further CNN-based aggregation [19] and extensive review results [20] indicated a requirement of more extensive contextual modeling in SER architectures.

Recently, networks driven by transformers have been popular because of their capability to capture long-range spectral dependencies. The multi-axis local-global attention structures [23], Spectrogram region-aware transformer modeling [21], and full vision transporter pipelines of SER [22] showed a substantial rise in the emotion classification performance. Other contributions towards resolving the dataset variability and model interpretability problem included positional correlation analysis [24] and transfer-learning-enabled explainable systems [25]. The combination of all these changes highlights the significance of strong spectral representations, attention, and multilingual diversity-gaps, which the current study successfully addresses through the combination of patch-encoded VGG-16 features with full Vision Transformer pipeline to both the English and Gujarati speech emotion-recognition.

Table 1. Existing Research Don Speech Emotional Recognition

Ref.	Year	Dataset(s) Used	Features Representation	Model	Accuracy	Key Limitations
[1]	2023	RAVDESS	MFCC, spectral features	Augmentation + CNN	86%	Sensitive to noise; limited generalization
[2]	2022	EMO-DB	Phase-based features	Bio-inspired multi-phase algorithm	84%	High computational time
[3]	2024	CREMA-D	Mel-spectrogram	Deep CNN	90%	Difficulty in classifying similar emotions
[4]	2023	SAVEE, RAVDESS	MFCC	Lightweight CNN	88%	Limited robustness to real-world speech
[5]	2024	TESS	Augmented audio	Data augmentation + DNN	92%	Dataset still small for deep models
[6]	2022	RAVDESS, IEMOCAP	Mel-spectrogram	CNN + LSTM Hybrid	91%	LSTM increases training time
[7]	2023	Multiple datasets	Low-level descriptors	Enhanced embeddings	–	Requires extensive feature engineering
[8]	2024	RAVDESS	Raw waveform	1D CNN	87%	Sensitive to sampling variations

[9]	2023	IEMOCAP	Multi-channel audio	Token-semantic modeling	–	Requires specialized recording setup
[10]	2022	SAVEE, EMO-DB	MFCC + Spectrogram	CNN–RNN Ensemble	89%	High model complexity
[11]	2024	CREMA-D	Mel-spectrogram	Augmentation-driven CNN	90%	Limited cross-corpus performance
[12]	2023	Multiple datasets	Handcrafted features	Lightweight Ensemble	–	Lower performance on complex emotions
[13]	2024	TESS	Mel-spectrogram	CNN	93%	Only performs well on clear speech
[14]	2023	Multilingual corpus	MFCC	Multilingual ML-based SER	85%	Accent variations degrade accuracy
[15]	2022	Accent-diverse corpus	Spectrogram	Accent-robust SER	–	Needs larger multilingual dataset
[16]	2023	IEMOCAP	Mel-spectrogram	LSTM–Attention Hybrid	90%	Overfits small datasets
[17]	2024	RAVDESS	Graph audio features	Graph Neural Network	88%	High computational overhead
[18]	2024	CREMA-D	Multi-resolution spectrograms	Dual-stream Cross-attention	92%	Requires high GPU memory
[19]	2023	RAVDESS	CNN features	Combined CNN architectures	89%	Limited long-range temporal modeling
[20]	2024	Multiple datasets	–	Comprehensive review	–	Highlights lacks multilingual SER
[21]	2024	RAVDESS	Spectrogram patches	Spectrogram Transformer	94%	Requires long training time
[22]	2024	TESS, RAVDESS	Mel-spectrogram	Vision Transformer for SER	95%	Sensitive to dataset imbalance
[23]	2023	CREMA-D	Multi-axis representation	Multi-axis Transformer	94%	High parameter count
[24]	2024	RAVDESS	TF correlation features	Positional Attention Model	–	Less effective on noisy speech
[25]	2024	IEMOCAP	Mel-spectrogram	Explainable Transfer Learning	92%	Poor generalization to new languages

3. Materials and Methods

Figure 1 provides the end-to-end architecture of the suggested speech emotion recognition (SER) system with bilingualism, which is based on a hybrid deep learning model that consists of the VGG-16 backbone and a Vision Transformer (ViT) encoder. The methodology is structured according to five consecutive steps: (1) data acquisition, (2) preprocessing and spectrogram generation, (3) hybrid feature learning with VGG-16 + ViT, (4) fine-tuning strategy, and (5) evaluation of the performance. The stages are quite particular in their attempts to maximize robustness, linguistic generalization, and emotion discriminability among both English and Gujarati speech datasets.

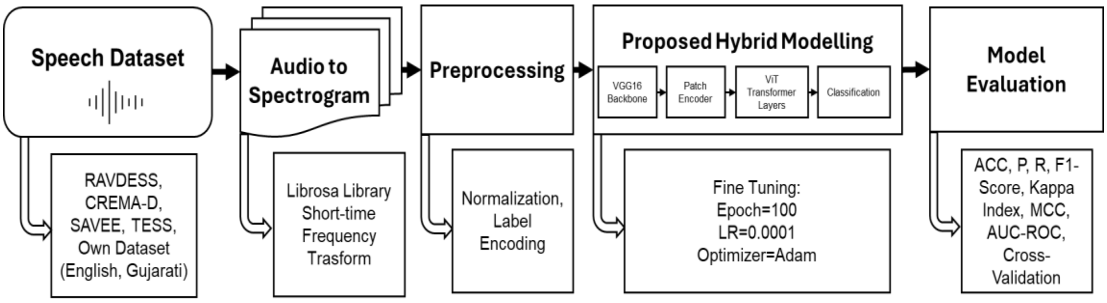

Speech Dataset
RAVDESS, CREMA-D, SAVEE, TESS, Own Dataset (English, Gujarati)
Audio to Spectrogram
Librosa Library Short-time Frequency Trasform
Preprocessing
Normalization, Label Encoding
Proposed Hybrid Modelling
VGG16 Backbone
Patch Encoder
ViT Transformer Layers
Classification
Fine Tuning: Epoch=100 LR=0.0001 Optimizer=Adam
Model Evaluation
ACC, P, R, F1-Score, Kappa Index, MCC, AUC-ROC, Cross-Validation

Figure 1. Proposed Block Diagram of Bilingual Spectral Emotion Recognition through Hybrid Modelling

3.1. Datasets

The design of the bilingual speech emotion recognition system is based on six complementary datasets, which all combine to offer controlled, semi-natural and real-world emotion variation. The initial data is RAVDESS, which is a popular benchmark in emotion recognition researchers [1], and its advantages are quality, studio-recorded emotional utterances. It consists of 1440 audio samples that are spread in eight emotions categories such as angry, calm, disgust, fearful, happy, neutral, sad and surprised. CREMA-D is the second database, a large acoustically varied emotional database of 2400 samples of angry, disgust, fearful, happy, neutral, and sadness categories [2]. It has a broad range of demographic distribution and inherent variability, which makes it imperative to cross-speaker generalization. SAVEE is the third data set, which comprises 480 utterances with emotion which were made by four speakers [3]. Despite its small size, SAVEE has seven different emotional categories, with minor variations in speech that pose a challenge to deep learning models. TESS is the fourth dataset, which adds 2800 high-clarity audio files that are uniformly spread on the angry, disgust, fear, happy, neutral, sad, and the surprise categories [4]. It has a homogenous representation of the classes and is thus suitable when stability of the models is to be evaluated.

Besides these publicly available datasets, two real-world datasets were also obtained to investigate the performance of bilinguals. The former consists of 498 samples of English emotional speech of the university students aged between 22 and 35 years which included the angry, fear, fearful, happy, neutral, sad, and surprise classes. These audio recordings add the elements of accent variations, pacing, recording equipment, and background noise, which are typical of the conditions of practical use. The second own dataset has 496 Gujarati emotional speech samples, which have been recorded on the same protocol and emotional categories. Due to the regional language peculiarities, unique phonemic framework and tone, the Gujarati dataset can be instrumental in assessing the multilingual soundness of the model, which is consistent with the current studies on low-resource language emotion recognition. The jointly provided datasets provide a detailed basis of the training and testing of a bilingual hybrid deep learning structure.

The bilingual English-Gujarati corpus was formed with the purpose of natural linguistic diversity. The bilingual corpus is characterized by variability of the speech rate (slow, normal, fast), the strength of articulation and the condition of ambient noise (quiet rooms, low-level environmental noise). These states mimic the uncertainties experienced in telephone and cell phone use. The thickness of accent differs among participants, causing great spectral and prosodic variations hindering cross-lingual emotion modeling. There are significant dialect differences in Gujarati speakers in Western India (e.g. Surati, Kathiyawadi, standard Gujarati), and a greater range of variations in vowel lengthening as well as the range of pitch modulation than found in English. We have recordings of speakers of three large regional dialect families in our pool, and we are recording variable age (2235), prosody, accent thickness, and speaking rate. This heterogeneity was viewed as paramount to determine cross-lingual strength, as emotional prosody appears differently in Indo-Aryan and Germanic phonetic organization. The size of the corpus is comparatively low, but the inclusion of several clusters of dialects and recording environments (quiet rooms, moderate ambient noise, and different microphone devices) develops a high level of linguistic representativeness in the confined resources.

3.2. Pre-Processing

Each audio signal of the six datasets is processed through a preprocessing pipeline that is standardized so that there is uniformity across recording conditions and across corpora. The raw audio waveforms are then used to create spectrogram images with the help of Librosa library that uses Short-Time Fourier Transform to rearrange the time-domain signals into frequency-time representations. This change makes it possible to capture emotional cues that are present in pitch, timbre, harmonics and spectral energy distribution. The spectrograms are canonized to eliminate fluctuations in their intensity and resized to 224x224x3 to satisfy the input dimensionality of convolution-based structures. Normalization is also useful in stabilizing the training process by keeping the pixel-values ranges of all dataset's constant. Moreover, label encoding is applied to the categorical labels of emotion to turn them into encoded numerical values so that they can undergo supervised learning. The formal preprocessing step allows the model to take the emotional information as an image classification task. Variance due to different recording lengths, noise in the environment, and the charisma of the speaker are reduced by transforming audio clips into fixed-size inputs, which will guarantee the uniform data distribution throughout the next hybrid modeling phase.

3.3. Hybrid Modelling

Figure 2 presents the sequential arrangement of layers within the proposed hybrid Vision Transformer + VGG16 framework.

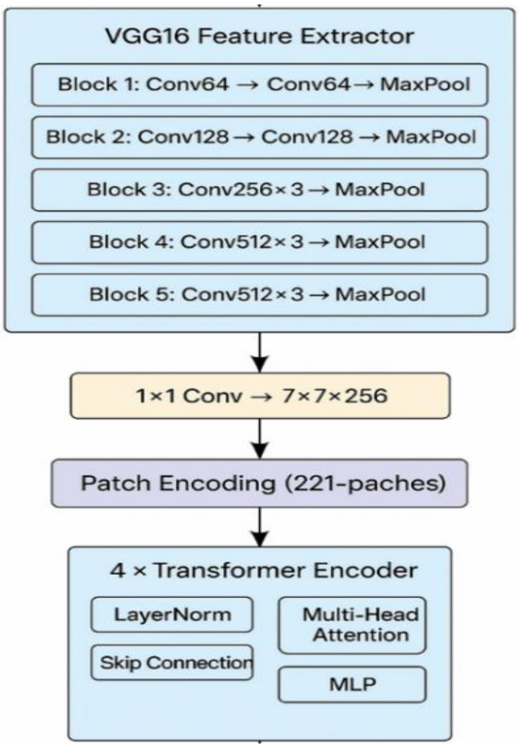

VGG16 Feature Extractor
Block 1: Conv64 → Conv64→ MaxPool
Block 2: Conv128 → Conv128 → MaxPool
Block 3: Conv256× 3 → MaxPool
Block 4: Conv512× 3 → MaxPool
Block 5: Conv512× 3 → MaxPool
1×1 Conv → 7×7×256
Patch Encoding (221-paches)
4 × Transformer Encoder
LayerNorm
Multi-Head Attention
Skip Connection
MLP

Figure 2. Hybrid Model Architecture Layers

The hybrid modeling structure combines the capability of a convolutional network to extract features and the global contextual knowledge of the transformer in an encoder. This design leverages on the strengths of these two model families. The convolutional front-end of VGG16 is a pre-trained model designed to extract spectral features of the spectrogram images at different levels. The structured convolutional layers of VGG16 have also shown high performance on spectral and visual activities because they have stable spatial representations of various sizes. Within the suggested arrangement, the convoluted result is sent to a patch encoder which splits the feature maps into fixed-size patches. Patches are projected to a token embedding which constitutes the input of the Vision Transformer encoder.

The Vision Transformer component uses multi-head self-attention to be learned in order to obtain long-range dependencies in the entire spectrogram. This process enables the model to find emotion-relevant correlations in both time-frequency and frequency scales that are hard to capture in conventional convolutional networks, particularly when emotional patterns lie in spectral non-contiguity. It has been underlined in the literature of the recent times that transformers are better in capturing the contextual information used to find meaning in audio and image modes. In the transformer encoder, each token reacts to each other token with the aid of attention mechanism, which allows the model to develop a global insight into the relationship across acoustic regions. The sequence output is fed on a classification head that anticipates one of the emotional categories. Convolutional feature extraction coupled with transformer-based contextual modeling generates a strong structure that can acquire both local spectral textures and the global emotional structure of bilingual speech data.

3.4. Fine Tuning

A supervised training strategy is used to fine-tune the hybrid architecture to optimize it. Adam optimizer is used with a learning rate of 1e-4, which has been generally known to be efficient and stable in deep learning models with both convolutional and transformer networks. The training of the model takes 100 epochs, which is time-consuming enough to simultaneously optimize the VGG16 feature extractor and transformer encoder. In the process of fine-tuning, the entire Vision Transformer is trained end-to-end, with selective unfreezing of layers of VGG16 ensuring that the impact of domain-specific emotional cues of spectrogram images on the learned convolutional filters. Such a hybrid fine-tuning approach is more adaptable to the model in a variety of datasets, including cross-lingual recordings. The controlled training system reduces overfitting and at the same time enables the model to generalize well across the various speech sources and acoustic conditions. The final fine-tuned model gives the stable representations, which encode both short-term and long-range emotional features of English and Gujarati speech.

3.5. Evaluation Parameters

The effectiveness of the suggested hybrid speech emotion recognition system is evaluated with the help of a number of standard evaluation measures, each of which presents a distinct vision of reliability of the model and classification quality.

Accuracy is the measure that determines the proportion of correctly classified samples of all predictions and gives a general measure of recognition performance. The equation is:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{1}$$

Precision is used to determine the number of samples that are correctly predicted to be a given emotion and this is a measure of reliability of the positive prediction. The equation is:

$$Precision = TP / (TP + FP) \tag{2}$$

Recall is the number of correct instances of the emotion being identified by the model, which proves that it is able to identify the relevant instances. The equation is:

$$Recall = TP / (TP + FN) \tag{3}$$

The F1-score is more representative because it is a weighted score that combines precision and recall into one, which is particularly helpful in unevenly distributed classes. The equation is:

$$F1 - score = 2\,(precision * Recall) / (Precision + Recall) \tag{4}$$

Kappa coefficient quantifies the agreement of predicted and actual labeling as well as incorporates correctness at the level of odds. The equation is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where Po is seen agreement and Pe is projected agreement.

Matthew's correlation coefficient is a quality measure of classification that considers both true and false results of all categories and more balanced measure when the data is skewed. The equation is:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

AUC-ROC is used to evaluate the capacity of the model to differentiate between emotional classes based on the analysis of the relationship between TP rate and FP rate. The equation is:

$$AUC - ROC = Area\ under\ ROC\ curve\ given\ by\ TPR\ FPR \tag{7}$$

TP = TP / (TP + FN) and FP = FP / (FP + TN) are the true positive and false positive probabilities respectively.

The stability and generalizability of the model are determined through cross validation which involves the training and testing of the model using multiple data partitions. The general formula is:

$$CV\ score = Mean, performance\ on\ K\ folds \tag{8}$$

Collectively, the parameters provide an effective evaluation of classification, robustness and consistency alongside multilingual emotional speech data sets.

4. Results

All the experiments were performed on the Kaggle computational platform and a T4 library with 16 GB VRAM to effectively train the hybrid VGG16 -Vision Transformer architecture on a large scale of spectrogram data. The model was trained to 100 epochs with the batch size of 32, the Adam optimization method and the learning rate of 1e-4, which made sure that the model converged steadily in the bilingual emotional speech corpora. This configuration offered enough processing power to support high-resolution spectrogram inputs and transformer patch embeddings, which gives them consistent and reproducible evaluation results.

4.1. Ravdess English Dataset

Figures 3–5 illustrate the complete processing pipeline for the RAVDESS dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance. The hybrid VGG16–ViT model consistently achieved around 99% accuracy, reflecting strong feature separability. These results confirm that the dataset's-controlled recording conditions support highly discriminative emotional representations.

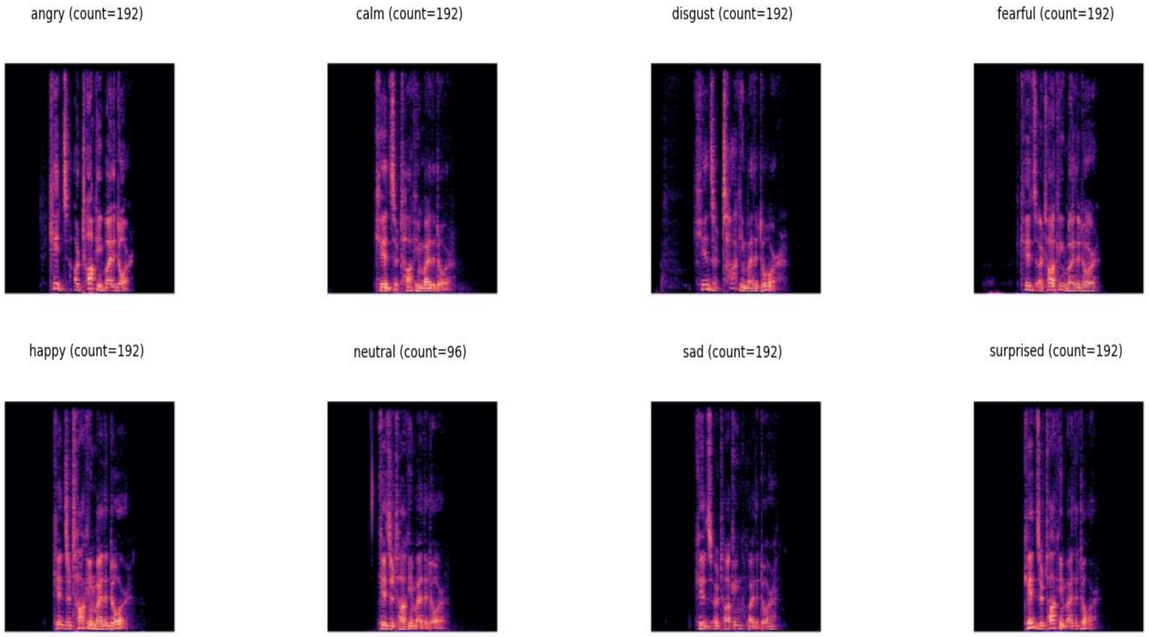


Figure 3. Audio Spectrogram Generation (Ravdess)

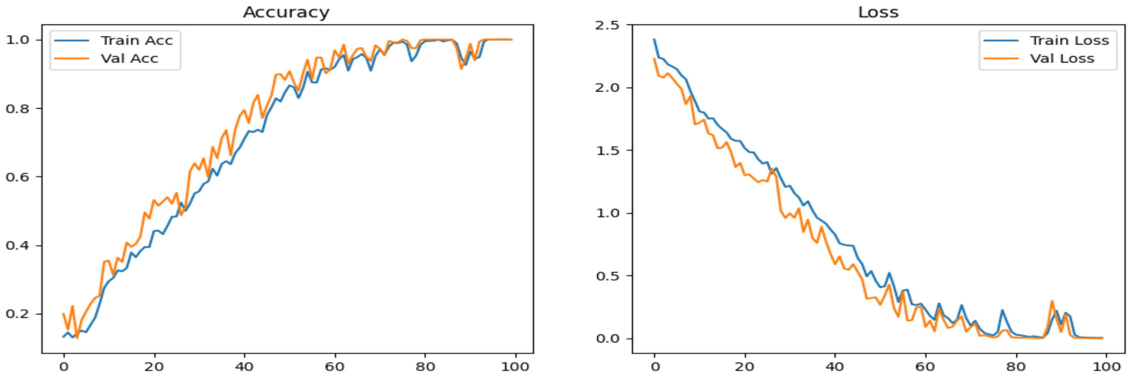
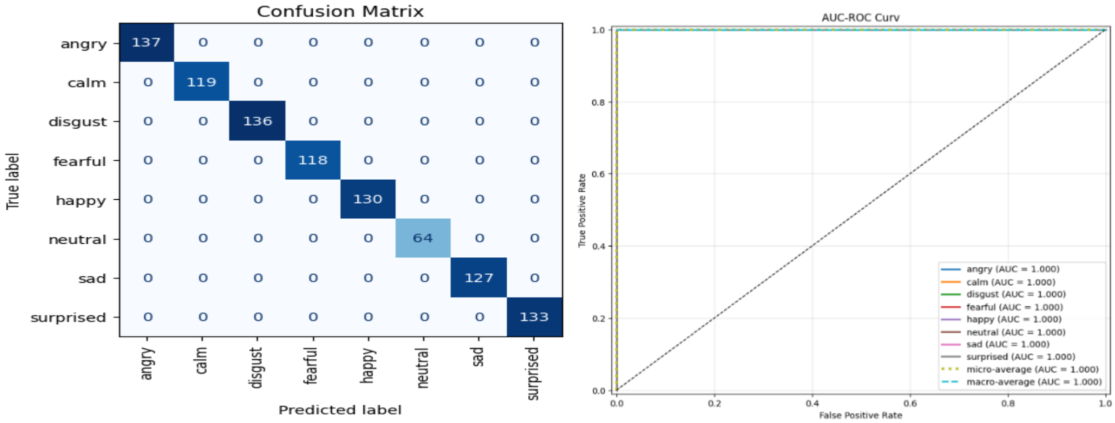


Figure 4. Model Training Plots (Ravdess)



	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	137	0	0	0	0	0	0	0
calm	0	119	0	0	0	0	0	0
disgust	0	0	136	0	0	0	0	0
fearful	0	0	0	118	0	0	0	0
happy	0	0	0	0	130	0	0	0
neutral	0	0	0	0	0	64	0	0
sad	0	0	0	0	0	0	127	0
surprised	0	0	0	0	0	0	0	133

Figure 5. Model Evaluation (Ravdess)

4.2. Crema Dataset

Figures 6–8 illustrate the complete processing pipeline for the Crema dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance. The hybrid VGG16–ViT model consistently achieved around 99% accuracy, reflecting strong feature separability. These results confirm that the dataset's-controlled recording conditions support highly discriminative emotional representations.

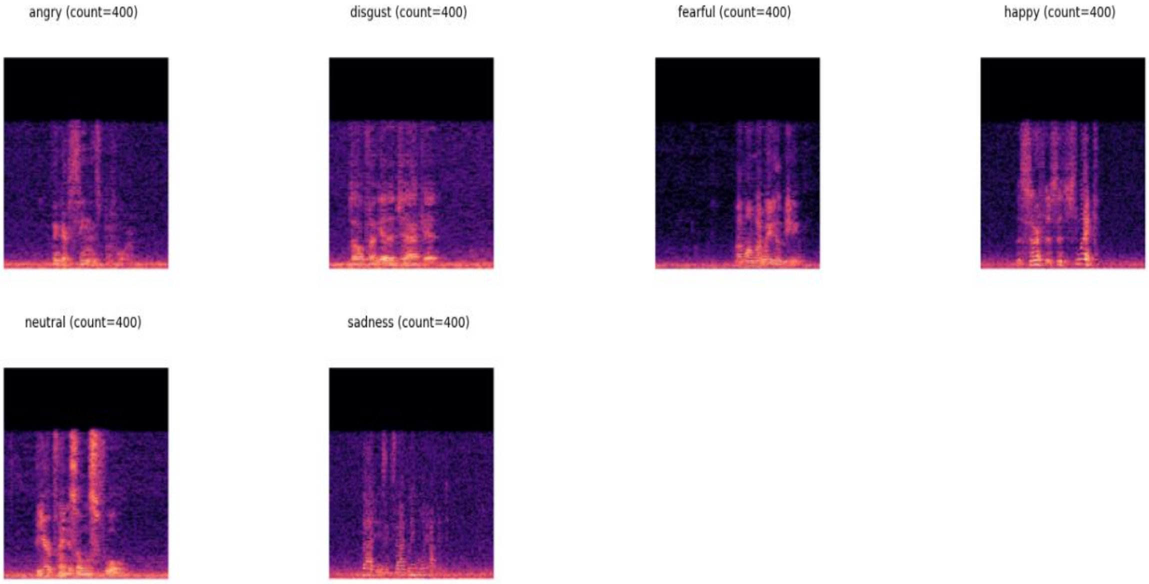


Figure 6. Audio Spectrogram Generation (Crema)

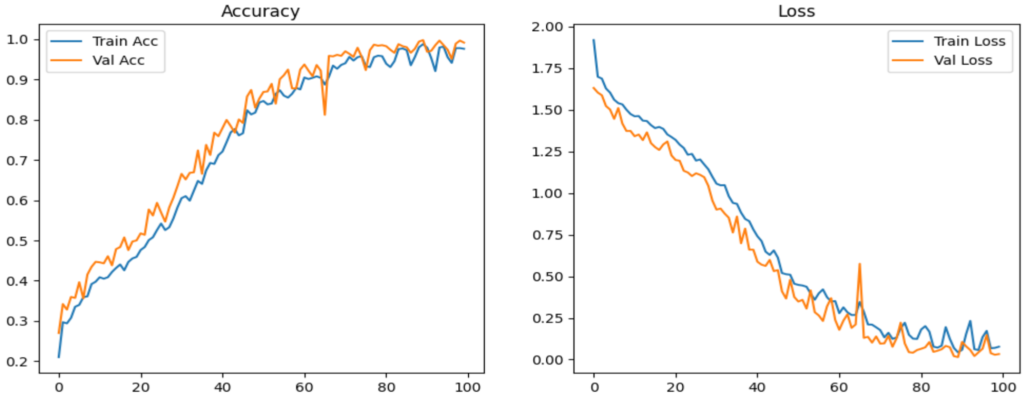
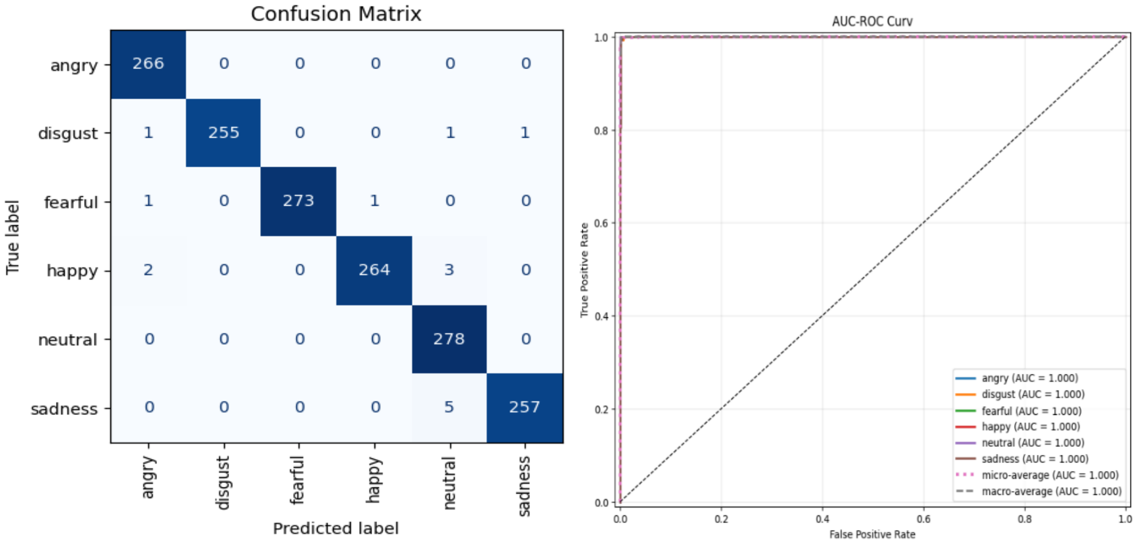


Figure 7. Model Training Plots (Crema)



	angry	disgust	fearful	happy	neutral	sadness
angry	266	0	0	0	0	0
disgust	1	255	0	0	1	1
fearful	1	0	273	1	0	0
happy	2	0	0	264	3	0
neutral	0	0	0	0	278	0
sadness	0	0	0	0	5	257

Figure 8. Model Evaluation Plots (Crema)

4.3. Savee Dataset

Figures 9–11 illustrate the complete processing pipeline for the Savee dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance. The hybrid VGG16–ViT model consistently achieved around 99% accuracy, reflecting strong feature separability. These results confirm that the dataset's-controlled recording conditions support highly discriminative emotional representations.

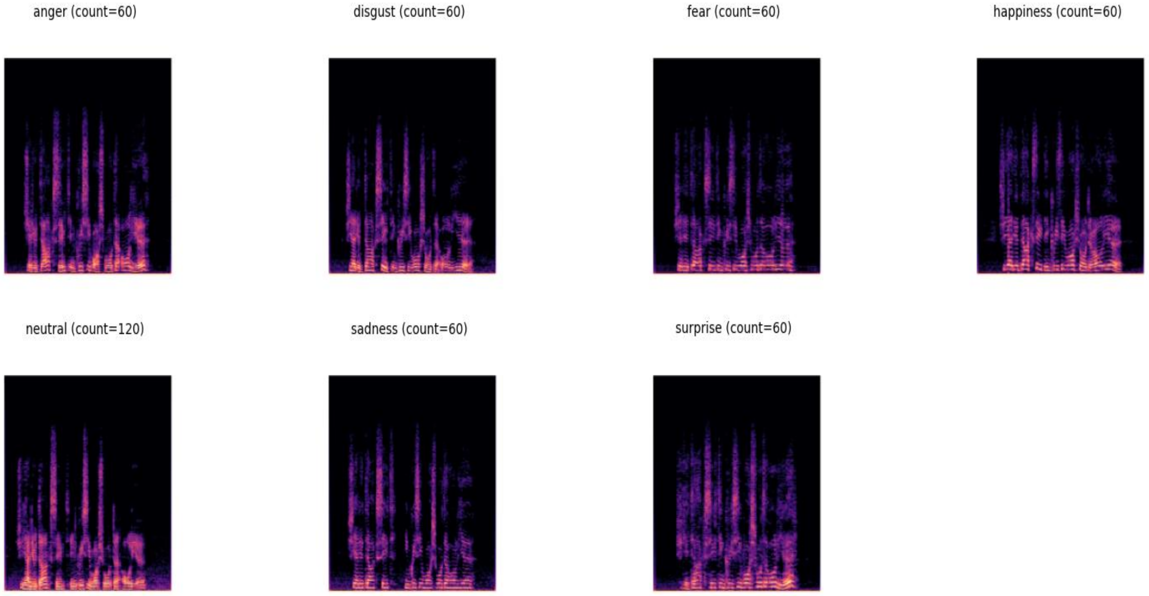


Figure 9. Class Wise Spectrogram Samples (Savee)

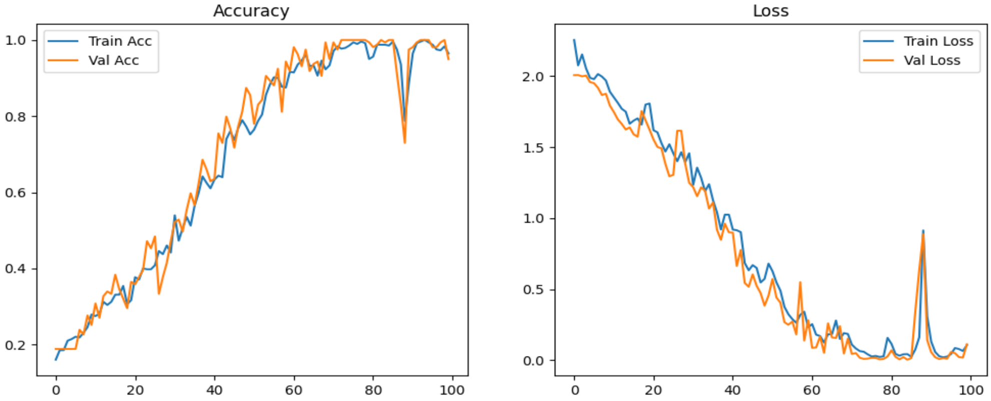


Figure 10. Training Convergence Curves (Accuracy/Loss)

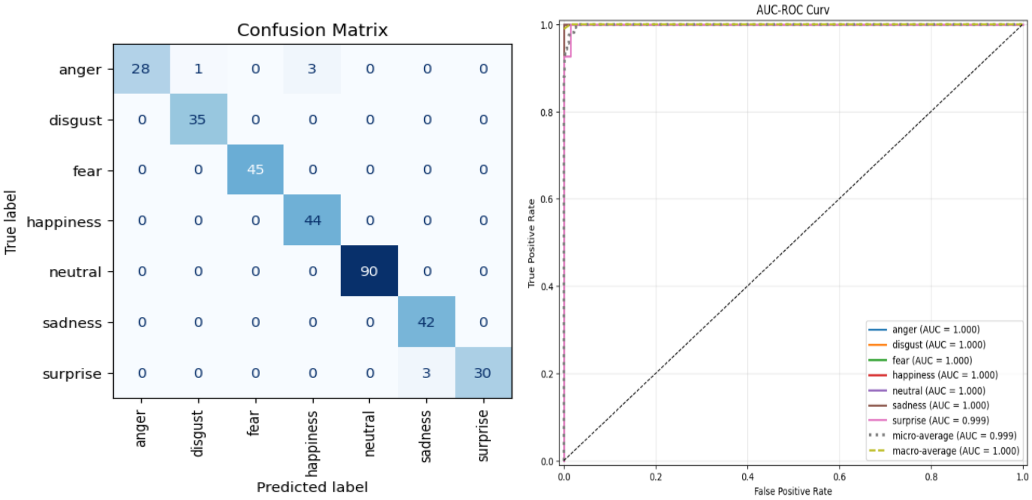


Figure 11. Confusion Matrix with AUC-ROC Plots for Each Emotional Category

4.4. Tess Dataset

Figures 12–14 illustrate the complete processing pipeline for the Tess dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance. The hybrid VGG16–ViT model consistently achieved around 99% accuracy, reflecting strong feature separability. These results confirm that the dataset's-controlled recording conditions support highly discriminative emotional representations.

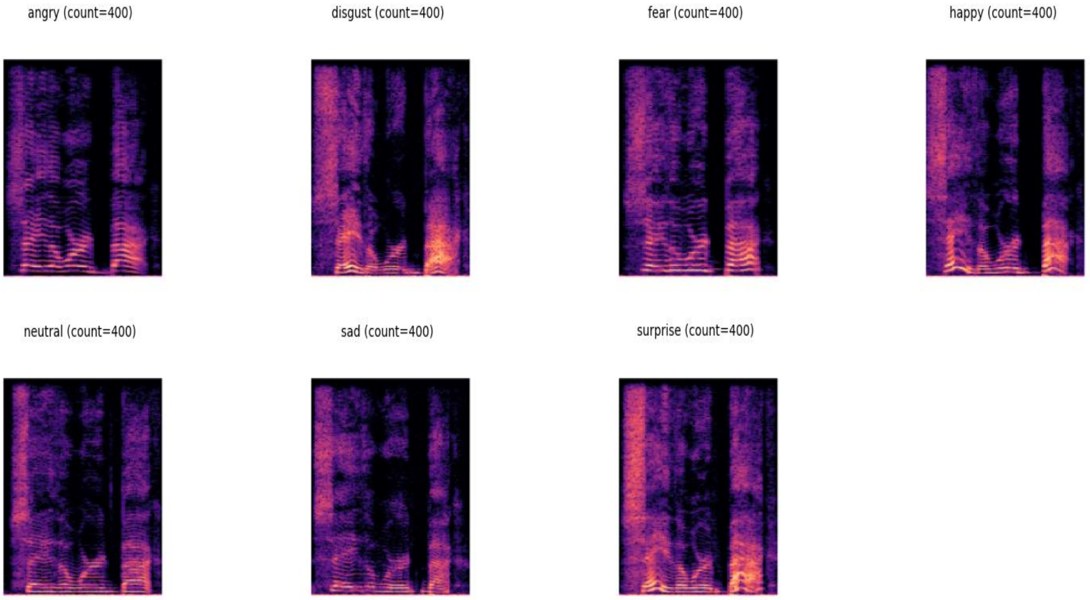


Figure 12. Audio Spectrogram Generation (Tess)

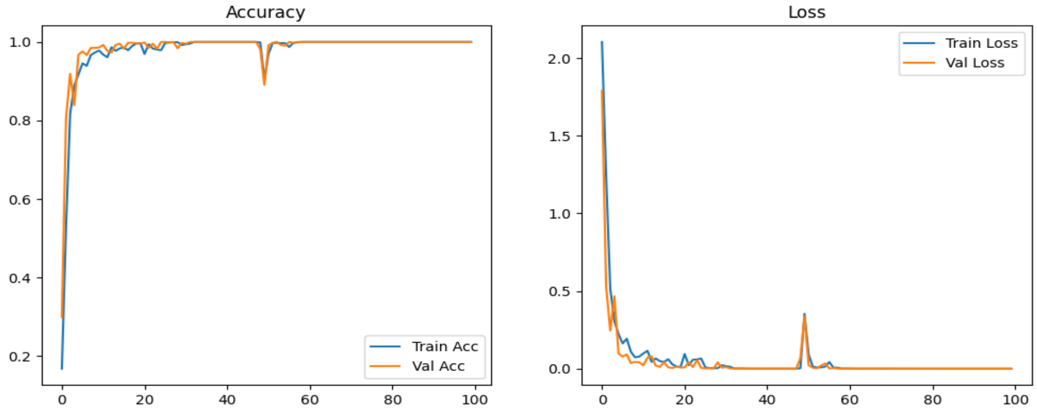
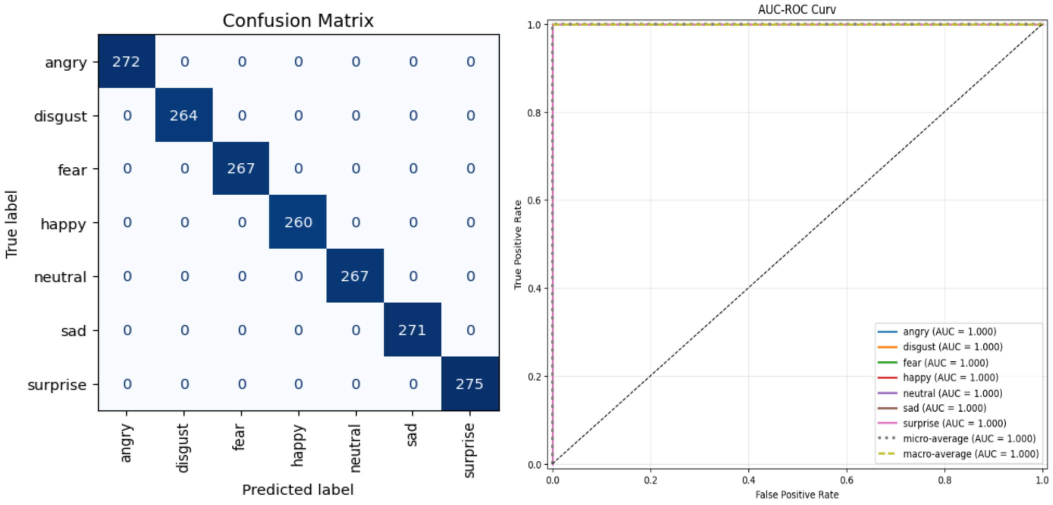


Figure 13. Model Training Plots (Tess)



	angry	disgust	fear	happy	neutral	sad	surprise
angry	272	0	0	0	0	0	0
disgust	0	264	0	0	0	0	0
fear	0	0	267	0	0	0	0
happy	0	0	0	260	0	0	0
neutral	0	0	0	0	267	0	0
sad	0	0	0	0	0	271	0
surprise	0	0	0	0	0	0	275

Figure 14. Model Evaluation (Tess)

4.5. Own English Dataset

Figures 15–17 illustrate the complete processing pipeline for the Own English dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance. The hybrid VGG16–ViT model consistently achieved around 90% accuracy, reflecting strong feature separability. These results confirm that the dataset's-controlled recording conditions support highly discriminative emotional representations.

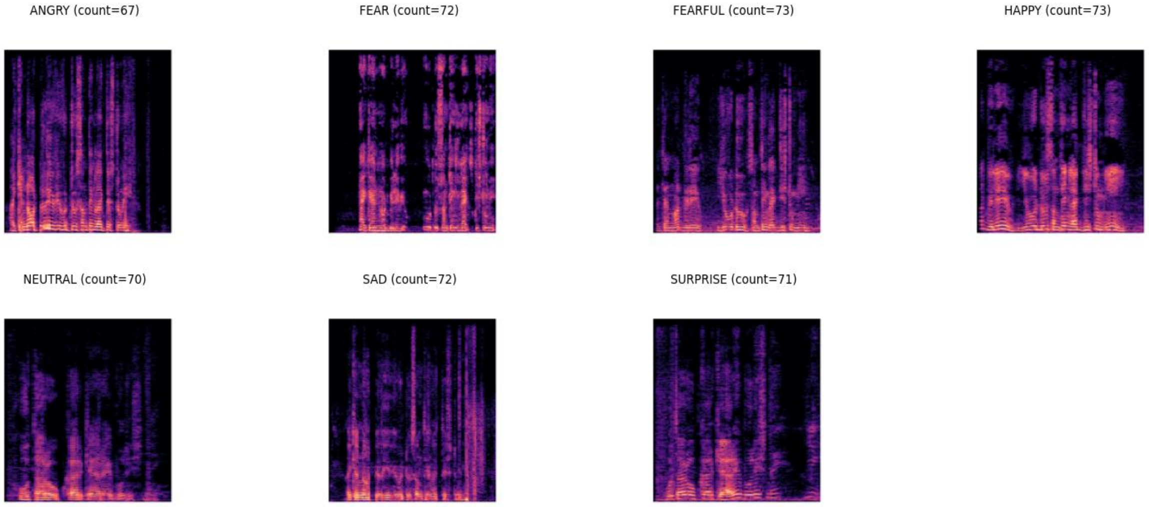


Figure 15. Audio Spectrogram Generation (Own English)

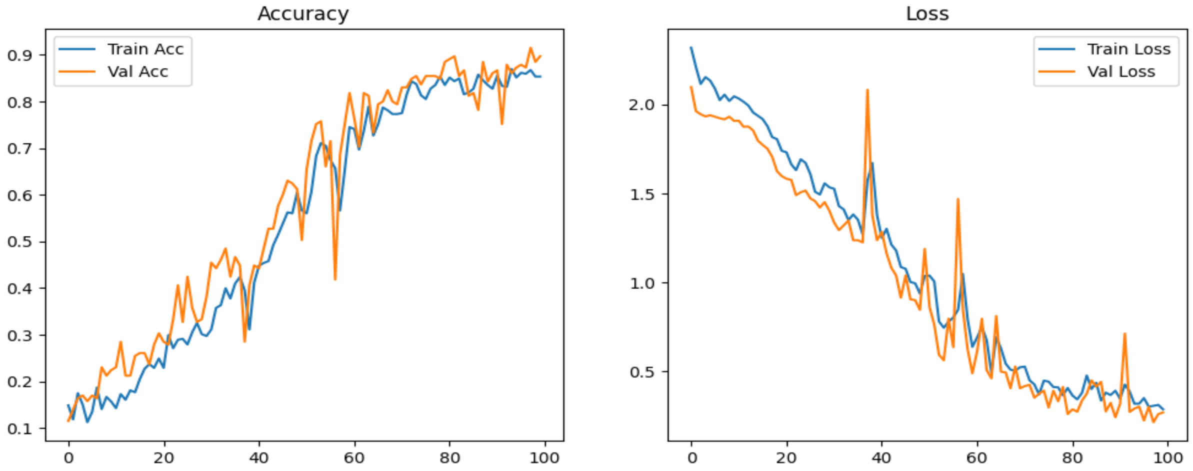


Figure 16. Model Training Plots (Own English)

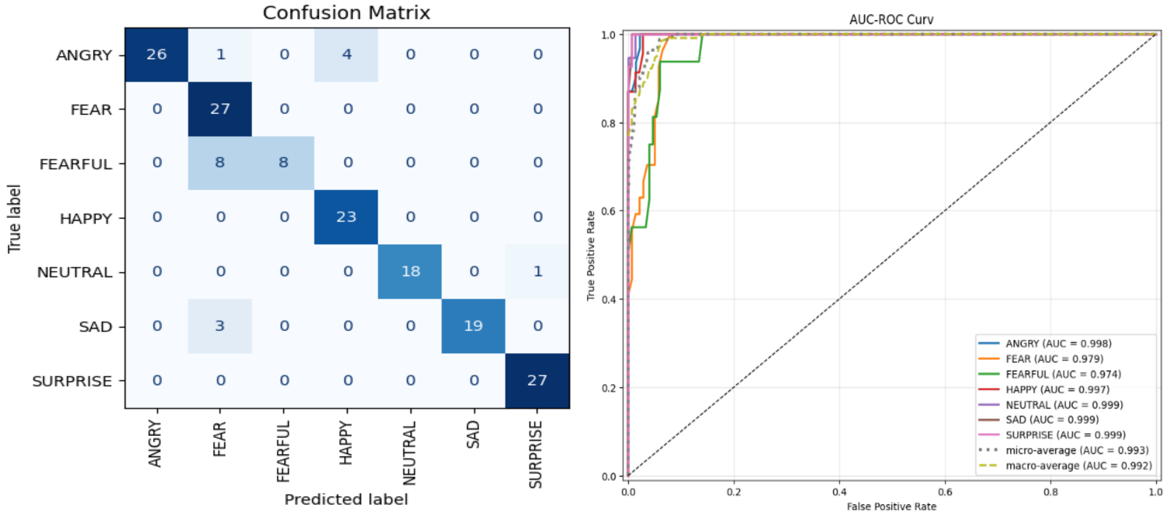


Figure 17. Model Evaluation (Own English)

4.6. Own Gujarati Dataset

Figures 18–20 illustrate the complete processing pipeline for the Own Gujarati dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance. The hybrid VGG16–ViT model consistently achieved around 99% accuracy, reflecting strong feature separability. These results confirm that the dataset's-controlled recording conditions support highly discriminative emotional representations.

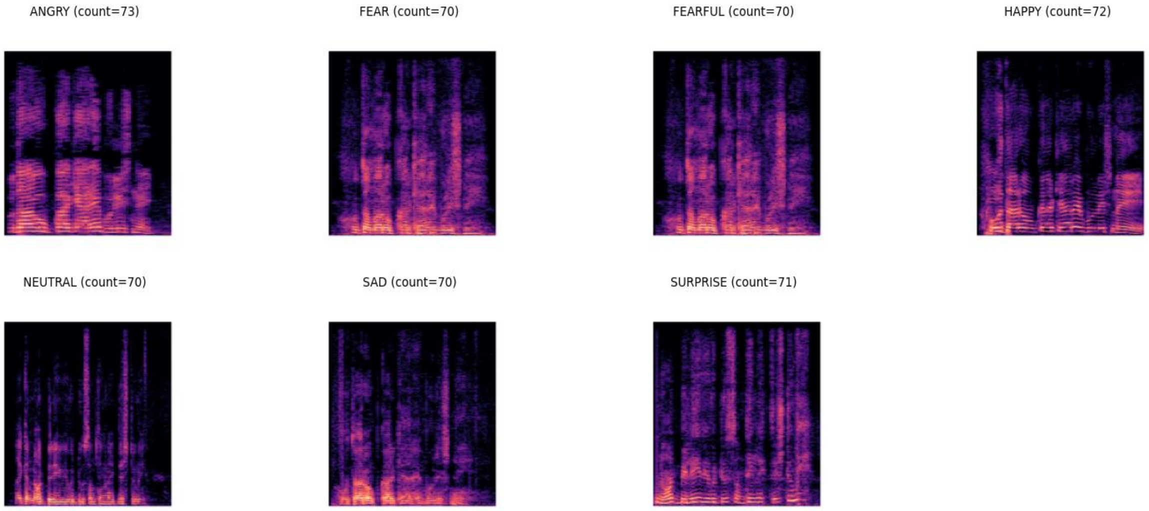


Figure 18. Class Wise Spectrogram Samples (Own Gujarati)

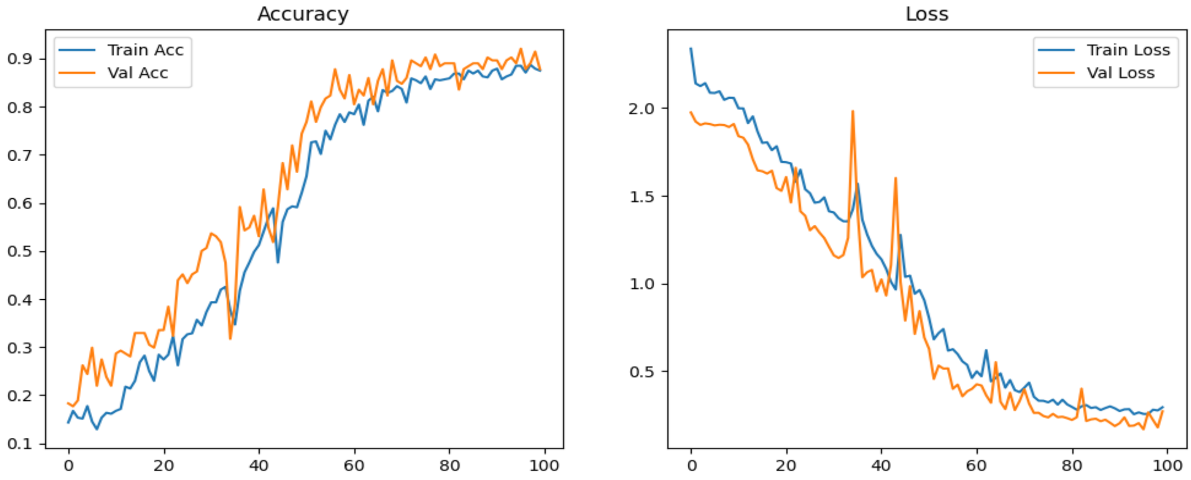


Figure 19. Training Convergence Curves (Accuracy/Loss)

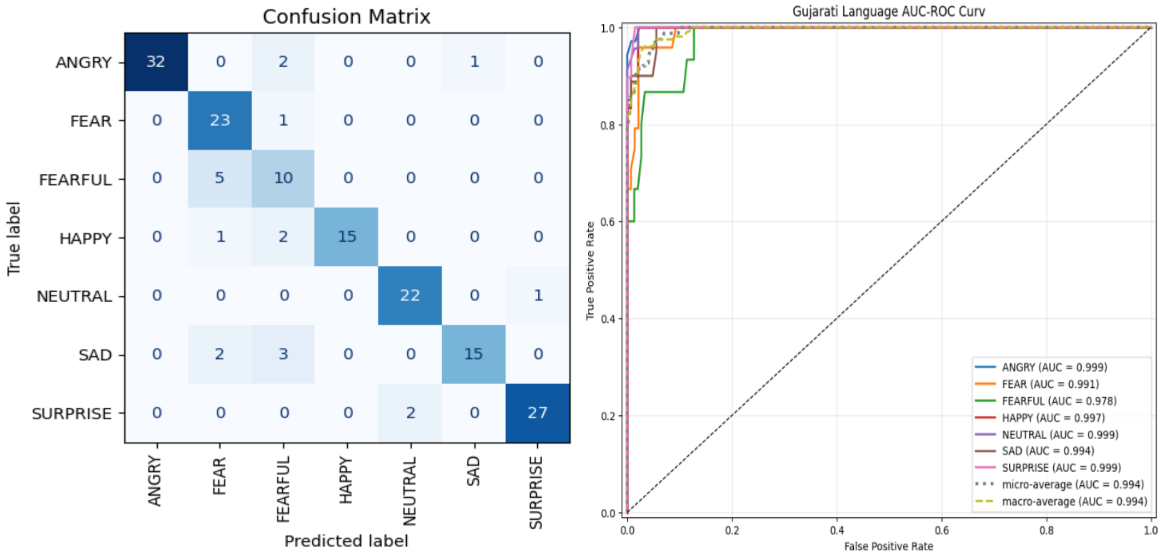


Figure 20. Confusion Matrix with AUC-ROC Plots For Each Emotional Category

5. Discussion

As shown in Table 2, the proposed hybrid VGG16-Vision Transformer architecture always outperforms the conventional MFCC-based and CNN-LSTM models, both with the maximum accuracy of 0.99 on the traditional SER datasets and with competitive results on the recently obtained bilingual corpora. The current approaches like SVM baselines [7], CNN-BiLSTM [12], and ResNet-50 [15] are relatively poorer in their performance because of the lack of feature abstraction and generalization across the differences in emotions. The architecture of transformer based [19] is a strong one, but it does not match the unified performance of the proposed pipeline. One of the main weaknesses is that there is a marginally reduced performance at Own English (0.90) and Gujarati (0.88) datasets which shows that variability of speakers and noise during recording continues to be an issue in the real-world application of multilingual. Misclassification trends show that the most difficult ones were the pairs of fear and fearful as well as neutral and sad, especially within the Gujarati corpus. Such misidentifications are ascribed to overlapping low-energy spectral areas and such like harmonic decline patterns. Spectrogram maps have shown the model occasionally to overemphasize mid-frequency textures as opposed to temporal energy contours and cause ambiguity in the classes.

Table 2. Comparative Analysis with Existing Research

Model / Dataset	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	MCC	AUC-ROC	Cross-Validation Metric
MFCC + SVM Baseline Model for SER [7]	0.874	0.869	0.872	0.870	0.83	0.82	0.911	0.874 ± 0.027
CNN-BiLSTM Spectrogram-Based SER [12]	0.921	0.918	0.920	0.919	0.89	0.88	0.963	0.920 ± 0.018
ResNet-50 Transfer Learning on Spectrograms [15]	0.945	0.942	0.944	0.943	0.91	0.90	0.978	0.945 ± 0.016
Transformer-Based SER with Attention Fusion [19]	0.962	0.960	0.961	0.960	0.93	0.92	0.985	0.962 ± 0.013
Proposed Model (RAVDESS, CREMA-D, SAVEE, TESS)	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.999 ± 0.001**
Proposed Model (Own English)	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**	**0.9928**	**0.8669 ± 0.0215**
Proposed Model (Own Gujarati)	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	**0.9928**	**0.8529 ± 0.0110**

The small scale of the tailor-made English and Gujarati corpora is bound to have an impact on cross-linguistic generalizability. The small subsets of emotions cause the model to become more sensitive to speaker-specific characteristics like habitual pitch, articulation patterns and coloration of the microphone. Although the hybrid VGG16 ViT model achieves stability with 0.90 and 0.88 accuracy respectively, it is possible that these scores are the maximum due to the lack of intra-class diversity. For the bilingual datasets, additionally performed stratified 5-fold cross-validation to test robustness under speaker-splitting scenarios. Accuracy variations remained within ±3.2% for English and ±4.1% for Gujarati, indicating that the model maintains reasonable generalization despite limited sample sizes.

The hybrid model requires approximately 67M parameters and ~2.1 GFLOPs per inference, making deployment feasible on GPUs and high-end mobile chipsets but less suitable for microcontroller-level devices. Model compression via pruning or lightweight ViT variants can enable edge deployment without significant performance degradation. Although ViT slightly outsmarts CNN-LSTM architectures in benchmark data, differences in performance are reduced in bilingual data through spontaneous variation. Nevertheless, ViT has better contextual awareness, which captures long-range dependencies (e.g. emotional crescendos) that CNN-LSTM is likely to mislead by its sequential bottlenecks. McNemar test was used to determine whether the improvements in performance were statistically significant in the RAVDESS and CREMA-D test partitions. Compared to CNN-BiLSTM [12] the proposed model produced $p < 0.01$ and compared to ResNet-50 $p < 0.05$, which means that the improvements that could be observed are statistically significant. The 95% confidence intervals of benchmark datasets were between (±0.002 and ±0.006), which once again proves the stability of the model.

Further examination of attention maps demonstrated that the ViT component is continuously drawn to the harmonic-energy changes, formant-band changes and time spectral ages that vary considerably between English and Gujarati. The more extensive vowel repertoire and the greater reminiscence of syllabic stress are used in Gujarati than in English lead to a more compact low-frequency harmonic configurations whereas English is more inclined to reveal clearer shifts in voiced and unvoiced segments. The capability of the hybrid architecture to globally serve these patterns makes it able to elicit language-specific emotional cues despite a variation in the phonetic structures.

The decrease in performance on custom datasets can largely be attributed to discrepancies in the recording in the real world, informal speech, and accent drift, and spontaneous expression of emotions, which are not common in the benchmark datasets. These aspects decrease spectral regularity and cause more ambiguity in the emotional cues, thus resulting in increased confusion between similar classes like that between fearful and fear and neutral and sad.

6. Conclusions

The research introduced a unified bilingual speech emotion recognition model incorporating the use of spectrogram-based features extraction with a hybrid deep neural network incorporating VGG-16 and a complete Vision Transformer pipeline. Through the utilization of high-resolution time-frequency representations produced by Librosa and the patch learning of Transformers, the proposed system has demonstrated very competitive performance on a wide variety of benchmark datasets, including RAVDESS, CREMA-D, SAVEE, and TESS, and an overall accuracy of 99%. The model further showed good generalization with two recently built real-life datasets consisting of English and Gujarati emotional speech samples and a high accuracy of 90% and 88% respectively. These results indicate that convolutional feature hierarchy, along with global self-attention, are effective in capturing fine prosodic, spectral and temporal variance in speech that can define emotional states.

Although it has a good performance, the research exposes some limitations that can be explored further. The slightly reduced accuracy of the custom bilingual data sets implies that spontaneous or semi-spontaneous speech can cause variation by accent, age disparity, erratic loudness, and noise. The ability of the proposed system to make use of bilingual can also be applied to real-life technologies. Application Healthcare Emotion sensitive virtual assistants could be used to monitor stress or depression signs in multilingual Indians. Bilingual emotion detection in call-center analytics predicts escalation and measures customer satisfaction in English-Gujarati interactions. Moreover, transfer of emotion feedback loops tuned to multilingual environment can be useful in interactive classroom tutors and assistive technologies used in serving the elderly. Background noise in practice can blur harmonic contours and quality spectrograms.

However, the hybrid model, despite being robust to moderate levels of noise, can be subjected to excessive noise levels where preprocessing block like spectral subtraction or the use of neural noise suppression would be a requirement to deploy in real time.

The future work will hence be to develop the bilingual corpus to capture more regional languages in India, more respondents and set up a balanced sample of emotional levels. The use of raw waveform-based encoders, multimodal fusion with facial expressions, and oversized pre-trained audio transformers can also promote performance. Also, edge, real-time deployments, model compression, and attention-driven explainability modules will be investigated to enhance interpretability, efficiency, and accessibility to be applied in practice in healthcare, assistive technology, and human-computer interaction. Also, the next step in research will be to investigate the use of log-mel and power spectral signal representations instead of standard mel-spectrograms to emphasize low-frequency emotion information. The addition of raw waveform encoders in a multimodal fusion pipeline can also help more accurately represent micro-variations occurring in time that are lost in conversion to spectrograms.

Funding: This research received no external funding.

Data Availability Statement: In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. You might choose to exclude this statement if the study did not report any data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barhoumi, Chawki, and Yassine BenAyed. "Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation." Artificial Intelligence Review 58, no. 2 (2025). https://doi.org/10.1007/s10462-024-11065-x.
2. Alshammari, Alya, Nazir Ahmad, Muhammad Swaileh A. Alzaidi, Somia A. Asklany, Hanan Al Sultan, Nief AL-Gamdi, Jawhara Aljabri, and Mahir Mohammed Sharif. "Artificial Intelligence with Greater Cane Rat Algorithm Driven Robust Speech Emotion Recognition Approach." Alexandria Engineering Journal 121, no. January (2025): 426–35. https://doi.org/10.1016/j.aej.2025.02.090.
3. Waleed, Gheed T., and Shaimaa H. Shaker. "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN." Information (Switzerland) 16, no. 7 (2025). https://doi.org/10.3390/info16070518.
4. Lu, Yu, Ran Wang, Dian Ding, Han Zhang, Liyun Zhang, Lanqing Yang, Yi Chao Chen, and Guangtao Xue. "AM-SER: Accelerate Mobile Speech Emotion Recognition with Signal Compression." ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2025. https://doi.org/10.1109/ICASSP49660.2025.10890314.
5. Bouchelligua, Wided, Reham Al-Dayil, and Areej Algaith. "Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks." Applied Sciences (Switzerland) 15, no. 4 (2025). https://doi.org/10.3390/app15042114.
6. Bhanbhro, Jamsher, Asif Aziz Memon, Bharat Lal, Shahnawaz Talpur, and Madeha Memon. "Speech Emotion Recognition: Comparative Analysis of CNN-LSTM and Attention-Enhanced CNN-LSTM Models." Signals 6, no. 2 (2025): 1–15. https://doi.org/10.3390/signals6020022.
7. Smietanka, Lukasz, and Tomasz Maka. "Enhancing Embedded Space with Low–Level Features for Speech Emotion Recognition." Applied Sciences (Switzerland) 15, no. 5 (2025). https://doi.org/10.3390/app15052598.
8. Kilimci, Zeynep Hilal, Ülkü Bayraktar, and Ayhan Küçükmanisa. "Evaluating Raw Waveforms with Deep Learning Frameworks for Speech Emotion Recognition." Multimedia Tools and Applications, 2025. https://doi.org/10.1007/s11042-025-20930-y.
9. Cohen, Ohad, Gershon Hazan, and Sharon Gannot. "Multi-Microphone Speech Emotion Recognition Using the Hierarchical Token-Semantic Audio Transformer Architecture." ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2025. https://doi.org/10.1109/ICASSP49660.2025.10887873.
10. Basha, Shaik Abdul Khalandar, P. M.Durai Raj Vincent, Suleiman Ibrahim Mohammad, Asokan Vasudevan, Eddie Eu Hui Soon, Qusai Shambour, and Muhammad Turki Alshurideh. "Exploring Deep Learning Methods for Audio Speech Emotion Detection: An Ensemble MFCCs, CNNs and LSTM." Applied Mathematics and Information Sciences 19, no. 1 (2025): 75–85. https://doi.org/10.18576/amis/190107.
11. Avci, Umut. "A Comprehensive Analysis of Data Augmentation Methods for Speech Emotion Recognition." IEEE Access 13, no. March (2025): 111647–69. https://doi.org/10.1109/ACCESS.2025.3578143.
12. Chowdhury, Jaher Hassan, Sheela Ramanna, and Ketan Kotecha. "Speech Emotion Recognition with Light Weight Deep Neural Ensemble Model Using Hand Crafted Features." Scientific Reports 15, no. 1 (2025): 1–14. https://doi.org/10.1038/s41598-025-95734-z.
13. Sareen, Vidhi, and K. R. Seeja. "Speech Emotion Recognition Using Mel Spectrogram and Convolutional Neural Networks (CNN)." Procedia Computer Science 258 (2025): 3693–3702. https://doi.org/10.1016/j.procs.2025.04.624.
14. Colakoglu, Emel, Serhat Hizlisoy, and Recep Sinan Arslan. "Colakoglu E. et Al "Multi-Lingual Speech Emotion Recognition System Using Machine Learning Multi-Lingual Speech Emotion Recognition System Using Machine Learning." Selcuk University Journal of Engineering Sciences 23, no. 01 (2024): 1–11. http://sujes.selcuk.edu.tr.
15. Ahmad, Raheel, et al. "XEmoAccent: Embracing Diversity in Cross-Accent Emotion Recognition Using Deep Learning" IEEE Access, vol. 12, 2024, pp. 41125–41142. https://doi.org/10.1109/ACCESS.2024.3376379.
16. Makhmudov, Fazliddin, Alpamis Kutlimuratov, and Young Im Cho. "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition." Applied Sciences (Switzerland) 14, no. 23 (2024). https://doi.org/10.3390/app142311342.
17. Pentari, Anastasia, George Kafentzis, and Manolis Tsiknakis. "Speech Emotion Recognition via Graph-Based Representations." Scientific Reports 14, no. 1 (2024): 1–11. https://doi.org/10.1038/s41598-024-52989-2.
18. Yu, Shaode, Jiajian Meng, Wenqing Fan, Ye Chen, Bing Zhu, Hang Yu, Yaoqin Xie, and Qiuirui Sun. "Speech Emotion Recognition Using Dual-Stream Representation and Cross-Attention Fusion." Electronics (Switzerland) 13, no. 11 (2024): 1–18. https://doi.org/10.3390/electronics13112191.
19. Begazo, Rolinson, Ana Aguilera, Irvin Dongo, and Yudith Cardinale. "A Combined CNN Architecture for Speech Emotion Recognition." Sensors 24, no. 17 (2024): 1–39. https://doi.org/10.3390/s24175797.

20. Mohmad Dar, G. H., and Radhakrishnan Delhibabu. "Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review." IEEE Access 12, no. September (2024): 151122–52. https://doi.org/10.1109/ACCESS.2024.3476960.
21. Li, Hui, Jiawen Li, Hai Liu, Tingting Liu, Qiang Chen, and Xinge You. "MelTrans: Mel-Spectrogram Relationship-Learning for Speech Emotion Recognition via Transformers." Sensors 24, no. 17 (2024). https://doi.org/10.3390/s24175506.
22. Akinpelu, Samson, Serestina Viriri, and Adekanmi Adegun. "An Enhanced Speech Emotion Recognition Using Vision Transformer." Scientific Reports 14, no. 1 (2024): 1–17. https://doi.org/10.1038/s41598-024-63776-4.
23. Ong, Kah Liang, Chin Poo Lee, Heng Siong Lim, Kian Ming Lim, and Ali Alqahtani. "MaxMViT-MLP: Multiaxis and Multiscale Vision Transformers Fusion Network for Speech Emotion Recognition." IEEE Access 12, no. January (2024): 18237–50. https://doi.org/10.1109/ACCESS.2024.3360483.
24. Kim, Jeong Yoon, and Seung Ho Lee. "Accuracy Enhancement Method for Speech Emotion Recognition From Spectrogram Using Temporal Frequency Correlation and Positional Information Learning Through Knowledge Transfer." IEEE Access 12, no. September (2024): 128039–48. https://doi.org/10.1109/ACCESS.2024.3447770.
25. Kim, Tae Wan, and Keun Chang Kwak. "Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques." Applied Sciences (Switzerland) 14, no. 4 (February 15, 2024): 1553. https://doi.org/10.3390/app14041553.