International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 08, August 2026
Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026

SPEECH EMOTION RECOGNITION USING A HYBRID CNN–SVM FRAMEWORK WITH THREE-CHANNEL MEL–DELTA–DELTA² SPECTROGRAM FEATURES

[1]Meghana Kuruba
PG Scholar
Department of Electronics and Communication Engineering
JNTUA College of Engineering, Anantapur
Andhra Pradesh, India
Email: meghanakuruba47@gmail.com

[2]Dr. P. Ramana Reddy
Senior Professor
Department of Electronics and Communication Engineering
JNTUA College of Engineering, Anantapur
Andhra Pradesh, India
Email: prr.ece@jntua.ac.in

Abstract: Speech Emotion Recognition (SER) has become a key area of study in affective computing.It allo ws smart systems to understand human emotions by analyzin g spoken words, which is useful in how people interact with computers., healthcare, virtual assistants, and smart communication systems. Accurate recognition of emotions remains challenging due to variations in speech characteristics, speaker dependency, and acoustic conditions. This paper presents a hybrid SER framework that combines deep transfer learning with machine learning to improve emotion classification performance. Speech recordings are first preprocessed through mono conversion, resampling, and amplitude normalization. Each audio sample is converted into a three-channel representation that includes Mel spectrogram, delta, and delta-delta features.. This representation preserves both spectral and temporal information. A pre-trained ResNet18 network is fine-tuned to extract meaningful deep features, which are then classified using a multi-class Support Vector Machine (SVM) with a radial basis function kernel. The proposed framework is tested on the RAVDESS emotional speech dataset, which includes six different emotion categories. The experimental results show that the overall classification accuracy is 93.04%, and the precision, recall, and F1-score are 0.94, 0.93, and 0.93 respectively. A MATLAB-based graphical user interface (GUI) is also developed to provide an end-to-end platform for audio preprocessing, spectrogram visualization, and real-time emotion prediction. Overall, the proposed hybrid CNN–SVM framework demonstrates reliable emotion recognition while maintaining computational efficiency, making it suitable for practical affective computing applications.

Keywords— Speech Emotion Recognition, Deep Learning, Transfer Learning, ResNet18, Support Vector Machine, Mel Spectrogram, RAVDESS, MATLAB, Affective Computing.

I. INTRODUCTION TO SER

Human speech conveys not only linguistic information but also emotional cues that reflect a speaker's psychological state. Automatic Speech Emotion Recognition (SER) aims to identify these emotional states from speech signals and has become an important component of affective computing. Reliable emotion recognition enables more natural human–computer interaction and has applications in intelligent virtual assistants, healthcare monitoring, call-center analytics, online education, driver assistance systems, and social robotics.

Recent improvements in deep learning have greatly enhance d SER performance by enabling the automatic learning of dis criminative representations from speech signals. Convolutional Neural Networks (CNNs) have demonstrated remarkable capability in extracting hierarchical features from spectrogram representations of speech. However, conventional CNN classifiers generally rely on fully connected layers, which may not always provide optimal discrimination when the extracted deep features exhibit complex decision boundaries. Traditional machine learning classifiers, particularly Support Vector Machines (SVMs), have shown excellent generalization ability for high-dimensional feature spaces and can effectively complement deep feature extractors.

The RAVDESS dataset has become one of the most commonly used benchmark datasets for assessing S peech Emotion Recognition (SER) algorithms. This is because it includes professionally recorded emotional speech with well-balanced emotion categories and high-quality recordings.Nevertheless, distinguishing acoustically similar emotions such as fear, sadness, and disgust remains challenging due to overlapping spectral characteristics and speaker variability.

To tackle these challenges, this work introduces a hybrid CNN–SVM framework that combines transfer learning with conventional machine learning. Speech recordings are converted into three-channel Mel–Delta–Delta² spectrogram images that preserve both spectral and temporal dynamics. A specially adjusted ResNet18 network is used to get detaile d and clear features from the speech. Then, a multi-class SVM is used to classify the detected emotions into different catego ries. The proposed framework is further integrated into a MATLAB-based graphical user interface (GUI), enabling

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 08, August 2026
Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026

audio preprocessing, spectrogram visualization, and interactive emotion prediction.

Unlike conventional transfer-learning-based SER models that directly employ the softmax classifier of a CNN, the proposed framework separates feature extraction from decision making by combining a fine-tuned ResNet18 network with a multi-class RBF-kernel Support Vector Machine. This design enables robust feature learning while allowing the classifier to construct effective decision boundaries within the learned feature space. Also, a three-channel Mel–Delta–Delta² spectrogram is used.
This helps keep both the sound patterns and the timing information in
the speech, making the emotion detection more accurate.

Main Contributions

The key contributions of this work can be summarized as follows:

- A hybrid SER framework integrating ResNet18 deep feature extraction with a multi-class SVM classifier.
- A three-channel Mel–Delta–Delta² spectrogram representation that captures complementary spectral and temporal information.
- A MATLAB-based GUI that supports audio preprocessing, spectrogram visualization, emotion ranking, confidence estimation, and interactive prediction.
- Comprehensive evaluation on the RAVDESS dataset using accuracy, precision, recall, F1-score, and confusion matrix analysis.
- Experimental validation demonstrating competitive recognition performance with an overall accuracy of 93.04%.

• II. RELATED WORK

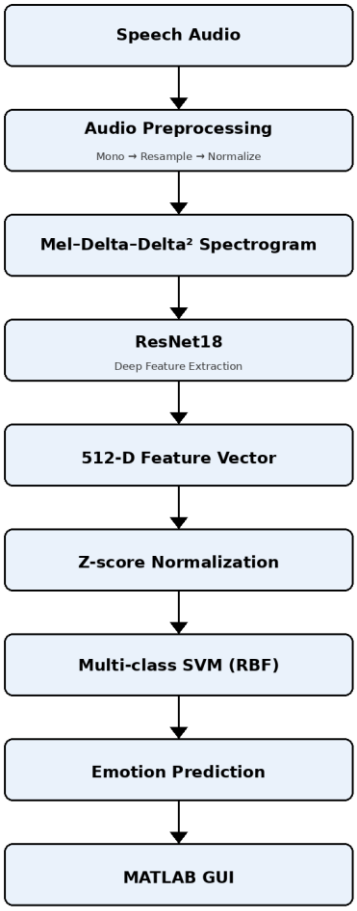
- Speech Emotion Recognition (SER) has attracted significant attention due to its applications in affective computing, intelligent human–computer interaction, healthcare, and conversational systems. Early SER approaches relied on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, zero-crossing rate, and spectral descriptors, which were classified using conventional machine learning algorithms including Support Vector Machines (SVMs), k-Nearest Neighbours (KNN), and Random Forests are used. Although these methods achieved reasonable performance, their dependence on manually engineered features limited their ability to capture complex emotional characteristics.
- Recent advances in deep learning have significantly improved SER performance by automatically learning hierarchical feature representations from speech signals. Convolutional Neural Networks (CNNs) have become popular for processing spectrogram-based speech representations because they effectively learn local time-frequency patterns associated with different emotions. Transfer learning using pre-trained models such as ResNet18, VGG16, DenseNet, and EfficientNet has further enhanced recognition accuracy while reducing training time and computational requirements.
- Several researchers have explored hybrid frameworks that combine deep feature extraction with traditional machine learning classifiers. In these methods, CNN models are used to learn discriminative features from spectrogram images, while classifiers such as SVM perform the final emotion prediction. Such hybrid models often demonstrate improved generalization because SVMs are effective in handling high-dimensional feature spaces extracted by deep networks.
- Despite these advances, distinguishing acoustically similar emotions remains challenging due to variations in speaking style, recording conditions, and speaker characteristics. Moreover, many existing studies primarily utilize single-channel spectrogram representations, which may not fully capture both spectral and temporal variations present in emotional speech.
- ***Research Gap:*** Existing SER studies generally focus on either end-to-end deep learning classifiers or conventional machine learning approaches. Comparatively fewer studies investigate the integration of transfer learning-based deep feature extraction with SVM classification using a three-channel Mel–Delta–Delta² spectrogram representation. Furthermore, many published works provide limited discussion of practical deployment or interactive visualization.
- To address this gap, the proposed framework transforms speech signals into three-channel Mel–Delta–Delta² spectrogram images, extracts deep features using a fine-tuned ResNet18 network, and performs multi-class emotion classification through an SVM classifier. The framework is further integrated into a MATLAB-based graphical user interface that enables preprocessing, visualization, and real-time emotion prediction.

Reference	Method	Dataset	Accuracy	Limitation
Ref. [1]	CNN	RAVDESS	88.6%	Limited feature representation
Ref. [2]	CNN + LSTM	RAVDESS	91.2%	Higher computational complexity
Ref. [3]	ResNet18	RAVDESS	92.5%	Fully connected classifier only
Ref. [4]	CNN + SVM	TESS	90.8%	Evaluated on a single dataset
Proposed	**ResNet18 + SVM**	**RAVDESS**	**93.04%**	**MATLAB GUI for interactive prediction**

• TABLE I. SUMMARY OF RELATED SER APPROACHES COMPARED WITH THE PROPOSED METHOD.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 08, August 2026
Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026



- **Fig. 1. Overall architecture of the proposed Hybrid CNN–SVM Speech Emotion Recognition framework. (Source: Authors).**

- **III. METHODOLOGY**

- ***A. Overview of the Proposed Framework***

- The proposed Speech Emotion Recognition (SER) framework integrates deep transfer learning with machine learning to accurately classify human emotions from speech signals. The framework consists of five sequential stages: audio preprocessing, spectrogram generation, deep feature extraction, feature normalization, and multi-class emotion classification. Figure 1 illustrates the overall architecture of the proposed system.

- Initially, the input speech signal is preprocessed by converting it into a mono-channel waveform, resampling it to 16 kHz, and normalizing the signal amplitude. The preprocessed audio is then transformed into a three-channel spectrogram representation composed of the Mel spectrogram, Delta spectrogram, and Delta-Delta spectrogram, preserving both spectral and temporal information.

- A pre-trained ResNet18 network is fine-tuned to extract a 512-dimensional deep feature vector from the generated spectrogram images. These features are standardized and provided to a multi-class Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel for emotion classification. Finally, the predicted emotion and confidence score are displayed through a MATLAB-based graphical user interface (GUI).

- ***B. RAVDESS Dataset***

- The proposed framework is evaluated using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), one of the most widely used benchmark datasets for Speech Emotion Recognition. The database contains professionally recorded speech samples from multiple speakers expressing different emotional states. For this study, six emotion categories were selected: Angry, Happy, Sad, Fear, Neutral, and Disgust. The dataset was divided into 80% training and 20% testing while maintaining class balance to ensure fair evaluation.

- ***C. Audio Preprocessing***

- Before feature extraction, each speech recording undergoes three preprocessing operations to improve data consistency.

- **1) Mono Conversion:** If the audio recording contains multiple channels, it is converted into a single-channel waveform by averaging the channels.

  - $x_{mono}(n) = [x_L(n) + x_r(n)] / 2$

- where $x_L(n)$ and $x_r(n)$ denote the left and right audio channels, respectively.

- **2) Resampling:** All recordings are resampled to 16 kHz to maintain a consistent sampling frequency across the dataset and reduce computational complexity.

- **3) Amplitude Normalization:** The speech waveform is normalized to reduce amplitude variations between speakers.

  - $x_{norm}(n) = x(n) / max(|x(n)|)$

- This process makes sure that the features we take out are not changed by how loud the recording is.

- ***D. Mel–Delta–Delta² Spectrogram Generation***

- Instead of using only a conventional Mel spectrogram, the proposed method constructs a three-channel representation consisting of the Mel spectrogram, Delta spectrogram, and Delta-Delta spectrogram. The Mel spectrogram represents the spectral energy of speech, whereas the Delta and Delta-Delta components illustrate the changes in speech over time. The Mel-frequency mapping is given by

  - $f\_mel = 2595 \times log_{10}(1 + f / 700)$

- where f is the linear frequency. The three feature maps are combined to form an RGB-like image, as summarized in Table II.

RGB Channel	Feature
Red	Mel Spectrogram
Green	Delta Spectrogram
Blue	Delta-Delta Spectrogram

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 08, August 2026
Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026

TABLE II. CHANNEL ASSIGNMENT FOR THE THREE-CHANNEL SPECTROGRAM IMAGE.

- Each spectrogram image is resized to 224 × 224 pixels, matching the input size required by ResNet18.

- ***E. Deep Feature Extraction Using ResNet18***

- A pre-trained ResNet18 model is employed as the backbone feature extractor. Transfer learning enables the network to utilize representations learned from large-scale image datasets while adapting them to spectrogram-based speech analysis. Instead of using the original fully connected classification layer, deep features are extracted from the pool5 layer. The resulting feature vector is

  - $F = [f_1, f_2, ..., f_{512}]$

- where each element represents a learned discriminative feature. These features encode high-level spectral patterns useful for emotion recognition while significantly reducing manual feature engineering.

- ***F. Feature Normalization***

- Prior to classification, each extracted feature is standardized using z-score normalization:

  - $z = (x - \mu) / \sigma$

- where μ is the mean of the training features and σ is the standard deviation. Normalization improves convergence and ensures equal contribution of all feature dimensions during classification.

- ***G. Multi-Class SVM Classification***

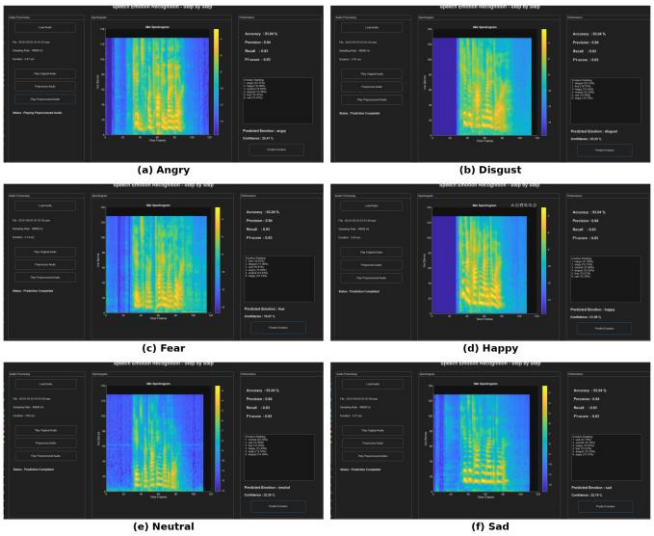
- The normalized feature vectors are classified using a multi-class Support Vector Machine (SVM) with an RBF kernel. MATLAB's Error-Correcting Output Codes (ECOC) framework is adopted to extend binary SVM classification to multiple emotion classes. The RBF kernel is defined as

  - $K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)$

- where x_i and x_j denote two feature vectors and γ is the kernel parameter. The final predicted emotion corresponds to the class with the highest decision score.

- ***H. MATLAB-Based GUI***

- To facilitate practical usage, the proposed framework is integrated into a MATLAB graphical user interface (GUI). The GUI provides an end-to-end environment where users can load a speech recording, perform audio preprocessing, display the Mel spectrogram, predict the emotional category, visualize the confidence score, display emotion ranking, and view overall performance metrics. This interface enables interactive evaluation without requiring command-line execution.



- **Fig. 4. MATLAB GUI showing the Mel spectrogram, emotion ranking, and predicted emotion for six representative RAVDESS test recordings: (a) angry, (b) disgust, (c) fear, (d) happy, (e) neutral, (f) sad. (Source: Authors).**

I. Design Rationale

The proposed framework was designed by combining transfer learning with a conventional machine learning classifier to balance recognition accuracy and computational efficiency. Instead of training a deep neural network from scratch, a pre-trained ResNet18 model was selected because it has demonstrated strong feature extraction capability while requiring relatively fewer computational resources than deeper architectures.

Rather than directly employing the final fully connected classification layer of ResNet18, the extracted 512-dimensional deep feature vectors are supplied to a multi-class Support Vector Machine (SVM). This design separates feature learning from decision making, allowing the CNN to focus on learning robust representations while enabling the SVM to construct effective decision boundaries within the learned feature space.

The proposed three-channel Mel–Delta–Delta² spectrogram representation was adopted to preserve complementary acoustic information. The Mel spectrogram captures the frequency distribution of speech, whereas the Delta and Delta² components provide first-order and second-order temporal variations that describe dynamic changes in emotional speech. Integrating these complementary representations improves the discriminative capability of the extracted deep features.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 08, August 2026
Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026

Overall, the proposed design combines transfer learning, hybrid classification, and spectrogram-based feature representation to achieve reliable speech emotion recognition while maintaining a relatively simple computational pipeline.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Experimental Setup

The proposed Speech Emotion Recognition (SER) framework was implemented in MATLAB R2025b. A pre-trained ResNet18 network was employed as the backbone model for deep feature extraction, while a multi-class The Support Vector Machine using a Radial Basis Function kernel was used to do the last step of classfying em otions.. The experiments were conducted using the RAVDESS emotional speech dataset, where the speech recordings were converted into Mel–Delta–Delta² spectrogram images prior to feature extraction.

The dataset was divided into 80% training and 20% testing to ensure unbiased model evaluation. All spectrogram images were resized to 224 × 224 × 3, which matches the input requirements of ResNet18. During training, the extracted 512-dimensional deep feature vectors were standardized using z-score normalization before being supplied to the SVM classifier.

Parameter	Value
Dataset	RAVDESS
Number of Emotions	6
Image Size	224 × 224 × 3
Feature Extractor	ResNet18
Feature Dimension	512
Classifier	Multi-class SVM
Kernel Function	RBF
Training Ratio	80%
Testing Ratio	20%
Software	MATLAB R2025b

TABLE III. EXPERIMENTAL CONFIGURATION PARAMETERS.

B. Performance Evaluation Metrics

The proposed framework was evaluated using four widely adopted classification metrics: Accuracy, Precision, Recall, and F1-score.

Accuracy measures the percentage of correctly classified speech samples.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision evaluates the proportion of correctly predicted samples among all predicted samples.

$$Precision = TP / (TP + FP)$$

Recall represents the capability of the classifier to correctly identify the actual emotional samples.

$$Recall = TP / (TP + FN)$$

The F1-score is found by taking the harmonic mean of Precision and Recall

$$F1 = (2 \times Precision \times Recall) / (Precision + Recall)$$

In this formula, TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

C. Experimental Results

The proposed hybrid CNN–SVM framework achieved high classification performance on the RAVDESS dataset. The combination of transfer learning-based feature extraction and SVM classification effectively captured discriminative emotional characteristics while maintaining good generalization capability.

Metric	Value
Accuracy	93.04%
Precision	0.94
Recall	0.93
F1-score	0.93

TABLE IV. OVERALL PERFORMANCE OF THE PROPOSED FRAMEWORK ON RAVDESS.

The obtained results demonstrate that the proposed framework accurately recognizes emotional speech across multiple emotion categories. The high precision indicates reliable predictions with fewer false positives, whereas the high recall confirms that most emotional samples were correctly identified. The F1-score further validates the balanced classification performance.

Fig. 2 shows the training progress of the ResNet18 backbone during the fine-tuning stage, prior to feature extraction for the SVM classifier. The network reached a validation accuracy of 73.68% after 35 epochs (1,995 iterations), with training carried out on a single-CPU resource in 282 minutes 23 seconds. The relatively stable convergence of both the accuracy and loss curves indicates that the fine-tuned backbone learned discriminative spectrogram features without significant overfitting, providing a reliable foundation for the downstream SVM classification stage.

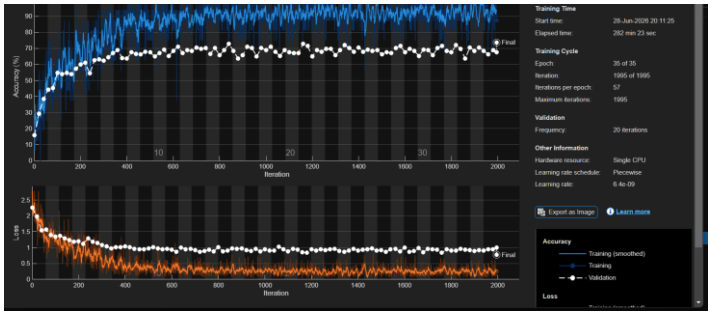


Fig. 2. Training and validation accuracy/loss curves of the ResNet18 backbone over 35 epochs. (Source: Authors).

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 08, August 2026
Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026

D. Confusion Matrix Analysis

Fig. 3 (to be inserted by the authors) presents the confusion matrix of the proposed Hybrid CNN–SVM framework. The confusion matrix provides a detailed visualization of the classification performance for each emotion category. The diagonal elements represent correctly classified samples, whereas the off-diagonal elements indicate misclassifications. Most speech samples are concentrated along the diagonal, demonstrating that the prposed framework accurately distinguishes the six emotional class.

Fig. 3. Confusion matrix of the proposed Hybrid CNN–SVM framework on the RAVDESS test set. (Source: Authors)

A small number of misclassifications occur between acoustically similar emotions, particularly those sharing comparable prosodic characteristics. Despite these overlaps, the overall classification accuracy of 93.04% indicates that the proposed deep feature representation effectively captures discriminative emotional information.

E. Error Analysis

Although the proposed framework achieved an overall classification accuracy of **93.04%**, a small number of misclassifications were observed between emotions that exhibit similar acoustic characteristics. The confusion matrix indicates that emotions such as **fear** and **sadness** occasionally overlap because both are characterized by relatively lower speech intensity and similar spectral energy distributions. Likewise, limited confusion between **anger** and **disgust** can occur due to comparable variations in pitch and energy.\ These observations indicate that the proposed hybrid framework successfully captures discriminative emotional characteristics despite the acoustic similarity between certain emotion categories.

Despite these challenges, the majority of predictions were concentrated along the diagonal elements of the confusion matrix, indicating consistent recognition across the evaluated emotion classes. The obtained results demonstrate that the hybrid ResNet18–SVM architecture effectively learns discriminative speech representations while maintaining balanced precision and recall across different emotional categories.

These observations suggest that incorporating complementary temporal information through the Mel–Delta–Delta² representation contributes to improved discrimination of emotional speech compared with relying solely on static spectral features.

To further validate real-time operation, six representative recordings from the RAVDESS test set were processed through the developed MATLAB GUI. For each recording, the Mel spectrogram was generated, the emotion probabilities were ranked, and the predicted emotion together with its confidence score was recorded, as summarized in Table VI.

Test File (RAVDESS)	True Emotion	Predicted Emotion	Confidence (%)	Result
03-01-05-01-01-01.wav	Angry	Angry	22.41	Correct
03-01-07-01-01-02.wav	Disgust	Disgust	23.23	Correct
03-01-06-01-01-03.wav	Fear	Fear	19.97	Correct
03-01-03-01-01-04.wav	Happy	Happy	21.06	Correct
03-01-01-01-01-05.wav	Neutral	Neutral	22.30	Correct
03-01-04-01-01-06.wav	Sad	Sad	22.79	Correct

TABLE VI. SAMPLE-LEVEL PREDICTIONS FROM THE MATLAB GUI ON THE RAVDESS TEST SET.

As shown in Table VI, the GUI correctly identified the ground-truth emotion for all six representative test recordings, with confidence scores ranging from 19.97% to 23.23%. These qualitative results, obtained independently of the aggregate accuracy reported in Table IV, further confirm that the proposed hybrid CNN–SVM framework performs reliably in interactive, real-time operation.

E. Comparative Analysis

To evaluate the effectiveness of the proposed framework, its performance was compared with recently published Speech Emotion Recognition methods reported in the literature.

Method	Dataset	Accuracy (%)

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 08, August 2026
Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026

CNN	RAVDESS	88.60
CNN + LSTM	RAVDESS	91.20
ResNet18	RAVDESS	92.50
CNN + SVM	RAVDESS	92.10
Proposed Hybrid CNN–SVM	**RAVDESS**	**93.04**

TABLE V. COMPARISON OF THE PROPOSED FRAMEWORK WITH EXISTING SER METHODS.

The comparison presented in Table V indicates that the proposed hybrid ResNet18–SVM framework achieves competitive performance compared with recently published SER approaches. Although several deep neural networks demonstrate strong classification accuracy, many require considerably larger computational resources and end-to-end training. In contrast, the proposed framework separates deep feature extraction from classification, allowing ResNet18 to generate discriminative feature representations while the RBF-SVM constructs robust decision boundaries. This hybrid strategy provides a favorable balance between recognition performance and computational efficiency, making the framework suitable for practical MATLAB-based deployment.

The comparison demonstrates that the proposed framework achieves competitive recognition accuracy while maintaining a relatively simple architecture. By combining transfer learning with an RBF-kernel SVM, the framework balances classification performance and computational efficiency, making the framework suitable for practical speech emotion recognition applications.

F. Discussion

The proposed hybrid CNN–SVM framework demonstrates that combining transfer learning with conventional machine learning is an effective strategy for Speech Emotion Recognition. The Mel–Delta–Delta² spectrogram representation preserves complementary spectral and temporal information, enabling the ResNet18 network to learn robust emotional representations. The SVM classifier further improves classification by effectively separating high-dimensional deep feature vectors.

In addition to its classification performance, the developed MATLAB-based GUI provides an interactive platform for loading speech recordings, visualizing spectrograms, predicting emotions, and displaying confidence scores. This makes the proposed framework suitable for educational demonstrations, research validation, and potential real-time affective computing applications.

V. CONCLUSION AND FUTURE WORK

This paper presented a hybrid Speech Emotion Recognition (SER) framework that combines deep transfer learning with machine learning for accurate emotion classification from speech signals. The proposed approach transforms speech recordings into three-channel Mel–Delta–Delta² spectrogram representations to preserve both spectral and temporal information. A pre-trained ResNet18 network was employed to extract discriminative deep features, while a multi-class Support Vector Machine (SVM) with an RBF kernel performed the final emotion classification.

The proposed framework was evaluated on the RAVDESS emotional speech dataset using six emotion categories. Experimental results demonstrated an overall classification accuracy of 93.04%, with precision, recall, and F1-score values of 0.94, 0.93, and 0.93, respectively. The confusion matrix analysis indicated that the proposed framework correctly classified most speech samples, with only a small number of misclassifications occurring between acoustically similar emotions. Furthermore, the developed MATLAB-based graphical user interface (GUI) enables interactive speech processing, spectrogram visualization, confidence estimation, emotion ranking, and real-time emotion prediction, illustrating the practical applicability of the proposed framework.

Although the proposed framework achieved competitive performance, several limitations remain. The evaluation was conducted on a single benchmark dataset, and the system was trained using clean speech recordings under controlled conditions. Performance may vary in real-world environments containing background noise, channel distortions, or speakers with diverse accents and speaking styles. Additionally, the current framework focuses exclusively on speech signals and does not incorporate complementary modalities such as facial expressions or physiological signals.

Future work will focus on improving the robustness and generalization capability of the proposed system by evaluating it on multiple benchmark datasets, including cross-corpus emotion recognition experiments. Future research may also investigate lightweight transformer-based architectures, self-supervised speech representations, and multimodal emotion recognition by integrating speech with facial and visual information. In addition, optimizing the proposed framework for real-time deployment on embedded and edge-computing platforms will further enhance its applicability in intelligent healthcare systems, human–computer interaction, virtual assistants, and affective computing applications.

A limitation of the present work is that evaluation was performed only on the RAVDESS benchmark dataset. Future studies will investigate cross-corpus evaluation under more diverse acoustic environments.

REFERENCES

[1] S.R.Livingstone and F.A.Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," PLOS ONE, vol.13, no.5, Article number e0196391, 2018.

[2] A.Krizhevsky, I.Sutskever, and G.E.Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, 2012.

[3] K.He, X.Zhang, S.Ren, and J.Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. [4] C.Szegedy et al., "Going Deeper with Convolutions," Proc.IEEE CVPR, 2015.

Received: 01-07-2026 Accepted: 06-08-2026 Published: 13-08-2026

[5] K.Simonyan and A.Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," ICLR, 2015.

[6] C.

Cortes and V.Vapnik, "Support-Vector Networks," Machine Learning, vol.20, no.3, pp.273–297, 1995.

[7] J.C.Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Microsoft Research, 1998.

[8] B.Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," Proc.ISMIR, 2000.

[9] D.P.W.Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," Columbia University, 2005.

[10] A.Satt, S.Rozenberg, and R.Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," Proc.Interspeech, 2017.

[11] N.-H.Ho, H.-J.Yang, S.-H.Kim, and G.Lee, "Multi-modal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," IEEE Access, vol.8, pp.61672–61686, 2020.

[12] A.Amjad, L.Khan, and H.-T.Chang, "Effect on Speech Emotion Classification of a Feature Selection Approach Using a Convolutional Neural Network," PeerJ Computer Science, 2021.

[13 S.

Latif, R.Rana, S.Khalifa, R.Jurdak, J.Qadir, and B.W.Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," IEEE Transactions on Affective Computing, vol.14, no.2, pp.1634–1654, April 2023, doi: 10.1109/TAFFC.2021.3114365

[14] S.Hochreiter and J.Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.

[15] Y.

LeCun, Y.Bengio, and G.Hinton, "Deep Learning," Nature, vol.521, pp.436–444, 2015.

[16] I.Goodfellow, Y.Bengio, and A.Courville, Deep Learning.MIT Press, 2016.

[17] D.

P.Kingma and J.Ba, "Adam: A Method for Stochastic Optimization," ICLR, 2015.

[18] J.

Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[19] A.Graves, Supervised Sequence Labelling with Recurrent Neural Networks.Springer, 2012.