A Novel Deep Hybrid Parallel Model of CNN-BiLSTM-Transformer for Speech Emotion Recognition

Original Scientific Paper

Sara Rekkal*
Tahri Mohammed University of Bechar, Faulty of Exact Sciences,
Dept of Mathematics and Computer Science, Bechar, Algeria
rekkal.sara@univ-bechar.dz

Kahina Rekkal
University of Naama, Department of Electrical Engineering,
Naama, Algeria
rekkal@cuniv-naama.dz

Adem Abdelhamid Boulhadid
Badji Mokhtar Annaba University, Computer Science Department,
Annaba, Algeria
adembhd@gmail.com

*Corresponding author

Abstract *– Human speech carries not only words but also rich paralinguistic signals that reflect emotional states. Despite significant progress in automatic speech emotion recognition (SER), many existing models still struggle to fully exploit vocal information. In this paper, we propose a novel deep hybrid model with three parallel structures: Convolutional Neural Networks (CNNs), Bi-directional Long Short-Term Memory Networks (BiLSTMs), and Transformers. Our model extracts local, temporal, and contextual information separately. This parallel architecture retains completeness and mitigates potential information loss in sequential architectures. Tested on five datasets for classifying discrete emotion types (neutral, happy, sad, angry, fearful, disgust, surprised, and calm), our model performs outstandingly, with a classification accuracy of 99.89% for RAVDESS, 100% for TESS, 99.82% for SAVEE, 95.64% for CREMA-D, and 100% for the EMO-DB dataset, consistently outperforming recent advanced methods.*

Keywords*: Emotion Recognition, Voice, Deep Learning, Signal Processing, CNN, Bidirectional LSTM, Transformer*

Received: December 21, 2025; Received in revised form: April 4, 2026; Accepted: April 7, 2026

1. INTRODUCTION

The human voice is a natural, salient, and prominent carrier of interpersonal communication, transmitting not only linguistic messages but also emotional cues that can be inferred even when they are implicit.

Emotion recognition based on vocal signals has seen growing applications in various fields. In medical contexts, vocal signal-based emotion analysis contributes to research on affective disorders and the diagnosis of mental illnesses such as depression and suicidal tendencies. Several studies demonstrate the use of speech emotion recognition (SER) in these areas. For example, a systematic review highlights SER applications in diagnosing depression, suicide, and other psychological disorders [1]. Other works include phoneme-level CNN models for clinical depression diagnosis from vocal signals [2] and ensemble CNN models in machine learning [3].

In human-computer interaction, SER enhances intelligent systems by making them more responsive and intuitive. Real-time HCI systems have been improved using one-dimensional CNNs with attention mechanisms [4], while CNNs have also been applied for real-time detection of speaker emotions [5].

In interactive systems and conversational agents, deep learning frameworks combining CNNs with Transformers or recurrent architectures allow advanced emotion recognition from speech [6]. These approaches also support behavioral analysis, improving communication systems, adaptive therapeutic tools, and dynamic behavior modeling [7].

Despite these advances, challenges remain: language, cultural, and environmental variability affect analysis reliability. Performance drops in noisy or reverberant environments [8], while interlinguistic and intercultural differences reduce generality [9]. Accent, gender, and other linguistic factors also influence recognition accuracy [10].

To address these limitations, this work proposes a new hybrid deep parallel architecture combining CNN, BiLSTM, and Transformer networks. This parallel design leverages the strengths of each model in feature learning, temporal dependency modeling, and global dependency modeling while mitigating information loss, as each model learns complementary aspects from the input. To the best of our knowledge, this represents a novel approach in the context of speech emotion recognition.

The rest of this paper is organized as follows: Section 2 presents the related work, Section 3 describes the methodology, Section 4 introduces the proposed architecture, Section 5 discusses the experimental results, and Section 6 concludes the paper.

2. RELATED WORKS

Recently, speech emotion recognition has gained attention due to advances in deep learning. Various approaches have been proposed to improve accuracy, robustness, and generalization. This section reviews representative works.

The research work described in [11] proposes a speech emotion recognition system using audio features including MFCC, chroma, Mel spectrogram, Tonnetz, and spectral contrast, which are used as inputs to a one-dimensional convolutional neural network (1D CNN), demonstrating improvement over existing solutions on the RAVDESS, EMO-DB, and IEMOCAP databases.

Paper [12] proposes a DCRF-BiLSTM approach that integrates BiLSTM learning of temporal features with Deep Conditional Random Fields (DeepCRF) sequence modeling, achieving high accuracy on the RAVDESS, TESS, SAVEE, EMO-DB, and CREMA-D databases.

In [13], the audio signals are converted to Mel spectrograms and processed by a CNN, yielding excellent results on the RAVDESS, SAVEE, TESS, and a combined version of these databases.

In the research study [14], a hybrid architecture combining CNNs and Transformers is proposed for SER, leveraging the strengths of both approaches to achieve performance surpassing current methods, although further improvements remain possible.

In paper [15], the importance of modeling temporal aspects is highlighted through a comparison of attention-based CNN-LSTM models with Mel-Spectrogram-ViT models.

For Arabic speech emotion recognition, it has been cited in [16] that data augmentation methods such as noise injection, time shift, and volume changes improve the performance of CNNs on the Saudi Dialect and BAVED databases.

In [17], separate one-dimensional CNN-LSTM and two-dimensional CNN-LSTM models are proposed, with the 2D CNN-LSTM achieving 95.89% on EMO-DB and 89.16% on IEMOCAP.

In [18], the authors present a systematic review of speech emotion recognition (SER) from a machine learning perspective, focusing on data processing, feature extraction, and classification, while discussing key challenges and model evaluation methods.

In conclusion, the use of CNNs, LSTMs, Transformers, and data augmentation methods has improved SER performance, allowing effective handling of complex audio inputs.

3. METHODOLOGY

This proposed framework starts with a data preparation phase. After the acquisition of audio data, an initial filtering process is applied to remove audio samples with low intensity, to ensure that training is performed using clear audio signals, followed by data preprocessing and data augmentation.

The preprocessing phase includes the following steps: loading the audio samples, converting them to a standard sampling rate, removing quiet parts, normalizing the amplitude of the signal, and finally resizing the signal samples to a constant size suitable for input to the model. Various data augmentation approaches are used to help diversify the data and make it more robust for the model. These approaches consist of adding Gaussian noise, time shifting, fixed gain transformation, and time stretching.

Once the audio signals are prepared, feature extraction is performed. In both original and augmented samples, three commonly used features of audio are extracted: Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, and the energy of the Mel spectrogram, which provide complementary information about the voice characteristics.

After feature extraction, the data are divided into training data and test data with emotional labels converted to numerical form. The features are then reshaped to meet the input requirements of the proposed CNN–BiLSTM–Transformer architecture.

Finally, the data are used to train the model to classify different emotional classes and obtain accurate emotion recognition results.

3.1. DATASETS

For the proposed study, five datasets for speech emotion recognition are used, including RAVDESS, TESS, SAVEE, EMO-DB, and CREMA-D. The number of audio samples for each emotion is shown in Table 1.

Table 1. Distribution of emotion classes across the various datasets

	RAVDESS [19]	SAVEE [20]	EMO-DB [21]	TESS [22]	CREMAD [23]
Happy	192	60	71	400	1271
Sad	192	60	62	400	1271
Angry	192	60	127	400	1271
Fear	192	60	69	400	1271
Disgust	192	60	46	400	1271
Surprise	192	60	0	400	0
Neutral	96	120	79	400	1087
Calm	192	0	0	0	0

3.2. PRE-PROCESSING, DATA AUGMENTATION, AND ACOUSTIC FEATURE EXTRACTION

Feature extraction is the process of using raw audio information to create useful representations for emotion recognition. It is important since audio data contain some redundancies as well as irrelevant features. Feature extraction aids in reducing data complexity by highlighting the important characteristics of audio that indicate emotional information.

During this phase, audio files with the WAV extension are processed through preprocessing and data augmentation, and then converted into feature vectors with the help of the Librosa library, which is a Python-based library developed for audio or music analysis.

3.2.1. Audio Preprocessing

Before extracting features from the signals, a standardized cleaning step is applied for each:

- Silence removal: Silence before and after the audio is removed using librosa.effects.trim () to ensure that only the relevant sections of the speech are analyzed.
- Resampling: All audio signals are resampled to a uniform sampling rate of 16 kHz to ensure consistency across datasets and compatibility with the subsequent processing pipeline.
- Normalization: All audio signals are amplitude-normalized to the range [−1, 1] to ensure numerical stability and facilitate efficient training of deep learning models.

3.2.2. Data Augmentation

To improve the robustness and generalization capability of the models, audio signal transformations are employed:

- Gaussian noise: Gaussian noise is added to the speech signals to simulate recording variability and improve model robustness. A noise scale of 0.0001 is used for the TESS, EMO-DB, and RAVDESS datasets, while a lower scale of 0.00001 is applied to the CREMA-D and SAVEE datasets. These values are empirically selected to preserve the natural dynamic range of the signals while introducing limited variability.
- Time shifting: Time shifting is applied by temporally translating the speech signals forward or backward to increase temporal diversity while preserving the original signal content.
- Fixed gain: The gain is applied by multiplying the signal by a fixed scalar to adjust the signal intensity. A gain of 0.95 is used for the TESS, EMO-DB, RAVDESS, and SAVEE datasets, while a gain of 0.80 is applied to the CREMA-D dataset. These values are empirically selected based on the observed intensity characteristics of each dataset.
- Time stretching: The audio signals are subjected to slight temporal stretching or compression using a ratio of 1.1 to introduce variability in speech rate while preserving the spectral characteristics of the signal.
- Padding: All speech signals are padded to a fixed length to ensure a uniform input size for the model. A length of 32,000 samples is used for the CREMA-D, TESS, and RAVDESS datasets, while 48,000 samples are applied for the SAVEE and EMO-DB datasets. These values correspond to the maximum signal lengths in each group, thereby preventing data truncation and preserving all temporal information.

3.2.3. Feature extraction

Feature extraction is a-key step in the proposed model. The Librosa library can extract three important categories of features:

- MFCCs (Mel-Frequency Cepstral Coefficients): These features capture the spectral envelope of the signal, representing the shape of the vocal tract and the dynamics of speech. To extract these features, a vector of forty MFCC coefficients is computed for each audio frame, and the average of these vectors is calculated over time.
- Chroma-STFT: This feature represents the energy distribution across the 12 chroma pitch classes, which helps the model identify harmonic and tonal patterns. The chroma features are averaged for each audio sample, resulting in a compact vector representation.
- Mel-Spectrogram: This technique helps measure the intensity of the signal in the time-frequency domain using the Mel scale, which gives a good approximation of the intensity of sound that can be detected by the human auditory system. This gives a vector of about 128 dimensions for each audio signal.

The feature vectors are concatenated horizontally into a 180-dimensional representation (40 MFCCs, 12 Chroma-STFT, 128 Mel-Spectrogram), capturing temporal and spectral characteristics. The feature extraction, augmentation, and preprocessing pipeline

provides consistent inputs, enabling effective speech emotion recognition. Fig. 1 illustrates the impact of the proposed data augmentation processes.

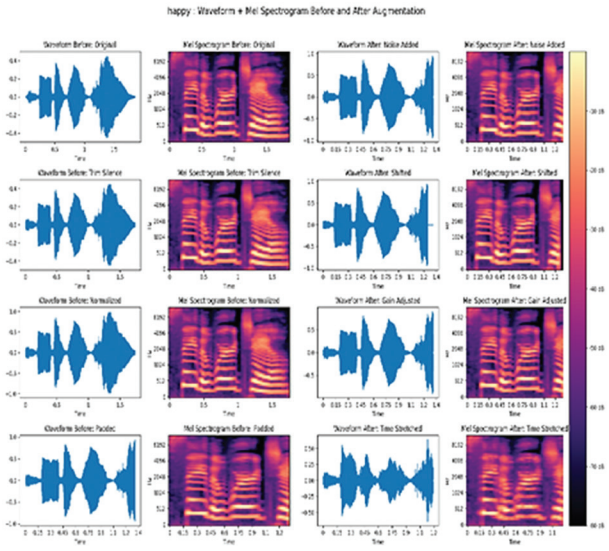


Fig. 1. Illustrative example: Waveform and Mel Spectrogram before and after augmentation for a "happy" audio file (TESS)

4. PROPOSED ARCHITECTURE

This proposed approach suggests a hybrid model architecture divided into three parallel parts: the Convolutional Neural Network (CNN) structure, the Bidirectional Long Short-Term Memory (BiLSTM) structure, and the Transformer model (Fig. 2). This architecture aims to effectively leverage the hierarchical features of the audio signal for the emotion recognition task.

4.1. CNN BRANCH

The Convolutional Branch includes four Conv1D layers for extracting local and spectro-temporal features from the audio signals.

- First layer: 256 filters, kernel size 5, ReLU activation, capturing basic local patterns.
- Second layer: 128 filters, kernel size 5, L1 and L2 regularization, Dropout 0.1, MaxPooling1D with pool size 8, reducing overfitting and downsampling the features.
- Third and fourth layers: 128 filters each, kernel size 5, L1 and L2 regularization, Dropout 0.1 before flattening, capturing more complex local patterns.
- Intermediate dense layer: Supports L1 and L2 regularization on weights, biases, and activations, enabling effective generalization while focusing on essential local acoustic patterns.

4.2. BILSTM BRANCH

The second branch consists of three bidirectional LSTM layers, which model the dynamics of the data along the time axis.

The first two BiLSTM layers have 256 units per direction, and return_sequences=True is used to pass the entire sequence to the next layers. Dropout with a rate of 0.1 is used after each. The third BiLSTM layer holds 128 units per direction and results in a compact representation over time. This is immediately followed by a Dense layer with 128 units, using ReLU activation, along with a Dropout of 0.1.

Such a structure allows for the identification of complex temporal dependencies and the representation of the distributed emotional context.

This branch ensures a robust representation of the audio data, ready for integration with the CNN and Transformer branches in the hybrid model.

4.3. TRANSFORMER BRANCH

Finally, the Transformer branch is designed to incorporate global dependencies using self-attention. It begins with Layer Normalization. This is followed by two Multi-Head Self-Attention layers with eight heads, where key_dim is set to 128. It is followed by a Global MaxPooling1D layer, which helps to obtain a fixed-size vector from the input sequence. This fixed-size vector is then projected using a Dense layer with a ReLU activation function and a dropout rate of 0.1. It is an effective branch for capturing global relationships that cannot be easily captured by either convolutional or BiLSTM layers.

4.4. BRANCH FUSION AND CLASSIFICATION

A multimodal feature is formed by concatenating the outputs of the three components: Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory Network (BiLSTM), and Transformer. This integrates the modalities as follows:

- Local spectral features learned by the CNN,
- Bidirectional temporal dependencies captured by the BiLSTM,
- Global and contextual relationships captured by the Transformer.

This diverse and complementary representation is then fed into the final Dense layer. Softmax activation produces a probability distribution over the emotional classes.

The model is compiled using the Adam optimizer with a learning rate of 0.0003, a combination known for its robustness in deep models, together with the sparse categorical cross-entropy loss function, optimal for multi-class classification when labels are integers. Learning was tracked using accuracy.

4.5. MODEL TRAINING

To facilitate in-depth analysis and allow for proper comparisons with existing studies, a standardized division of datasets was executed. Once audio features were derived, normalization and splitting for training and testing were conducted with 80% and 20%,

respectively. The model was trained for a total of 40 epochs. The optimizer used in this work is the Adam optimizer with a learning rate of 0.0003 and batch size of 32, which strikes a good balance between speed of convergence and model stability during training. The loss function used in this model is sparse categorical cross-entropy, which is ideal for a multi-class problem.

To counter overfitting and keep the model with the highest accuracy, EarlyStopping and ModelCheckpoint callbacks were used. The code for this project was written using TensorFlow and Keras, which ran on the Kaggle platform with two NVIDIA T4 GPUs available. The setup used in this project worked well with the computation involved in arriving at the proposed hybrid model for speech emotion recognition.

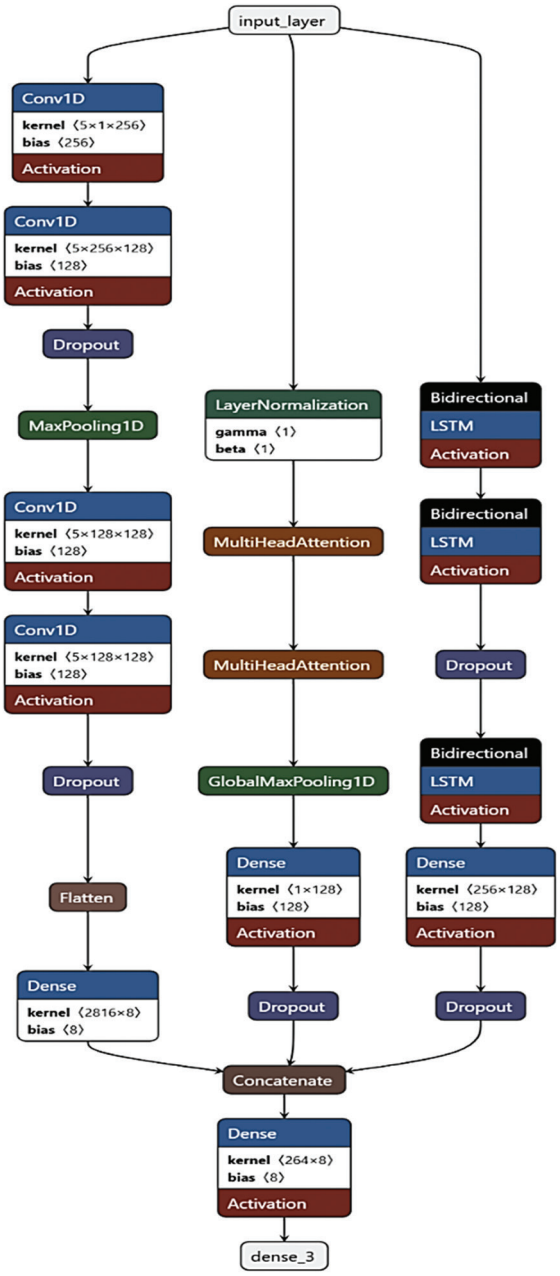


Fig. 2. Illustration of the proposed framework, consisting of three parallel branches: CNN, BiLSTM, and Transformer

5. EXPERIMENTS AND RESULTS

The proposed model was developed and tested on five different datasets. The results were analyzed using standard evaluation metrics, that is, accuracy, loss, and confusion matrix. This hybrid speech emotion recognition model combines three neural network architectures. The CNN captures local spectral features in speech, the BiLSTM analyzes temporal emotional transitions, and the Transformer with attention determines the most relevant voice segments. The outputs of all three models are combined to provide a robust and comprehensive emotion classification. Our model achieves outstanding results on all datasets considered, as shown in Table 2.

Table 2. Results on different datasets

Datasets	SAVEE	CREMA-D	TESS	EMO-DB	RAVDESS
Results	99.82%	95.64%	100%	100%	99.89%

These results reveal the robustness and excellent generalization skills of the architecture, achieving near-perfect accuracy for the majority of the datasets and maintaining a high level of accuracy (95.64%) for the largest, CREMA-D dataset.

For a better evaluation, precision, recall, and F1-score values are also presented for each dataset. This evaluation helps provide a clearer comparison of the results and highlights the model's robustness for the task of speech emotion recognition. Tables 3, 4, and 5 below summarize the evaluation results for all datasets.

The primary objective of the proposed model is to achieve high accuracy for recognizing speech emotion. Therefore, only the configuration achieving the highest level of accuracy is considered. The accuracy of the validation data is used as a major indicator of how well the proposed model will generalize, where the highest level of accuracy denotes the stage achieving the highest reliability of the predicted results.

Table 3. Class-wise emotion recognition performance on the CREMA-D and EMO-DB datasets

Emotion / Metric	EMO-DB			CREMA-D		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Neutral	1.00	1.00	1.00	0.92	0.96	0.94
Happy	1.00	1.00	1.00	0.96	0.97	0.96
Sad	1.00	1.00	1.00	0.94	0.93	0.93
Angry	1.00	1.00	1.00	0.98	0.97	0.98
Fear	1.00	1.00	1.00	0.94	0.95	0.94
Disgust	1.00	1.00	1.00	0.95	0.92	0.93
Accuracy	—	—	1.00	—	—	0.95
Macro Avg	1.00	1.00	1.00	0.95	0.95	0.95
Weighted Avg	1.00	1.00	1.00	0.95	0.95	0.95

Table 4. Class-wise emotion recognition performance on the RAVDESS and TESS datasets

Emotion / Metric	RAVDESS			TESS		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Neutral	1.00	1.00	1.00	1.00	1.00	1.00
Calm	1.00	0.99	1.00	—	—	—
Happy	1.00	1.00	1.00	1.00	0.99	1.00
Sad	0.97	1.00	0.98	1.00	1.00	1.00
Angry	1.00	0.98	0.99	1.00	1.00	1.00
Fear	0.99	0.99	0.99	1.00	1.00	1.00
Disgust	0.99	1.00	1.00	0.99	1.00	1.00
Surprise	1 .00	0.99	1 .00	1.00	1 .00	1.00
Accuracy	—	—	0.99	—	—	1.00
Macro Avg	0.99	0.99	0.99	1.00	1.00	1.00
Weighted Avg	0.99	0.99	0.99	1.00	1.00	1.00

Table 5. Class-wise emotion recognition performance on the SAVEE dataset

Emotion / Metric	SAVEE		
	Precision	Precision	Precision
Neutral	0.98	0.99	0.99
Calm	—	—	—
Happy	1.00	1.00	1.00
Sad	1.00	0.97	0.98
Angry	0.98	1.00	0.99
Fear	1.00	1.00	1.00
Disgust	1.00	1.00	1.00
Surprise	1.00	1.00	1.00
Accuracy	—	—	0.99
Macro Avg	1.00	0.99	0.99
Weighted Avg	0.99	0.99	0.99

The confusion matrix provides information on the number of correct and incorrect predictions with respect to each class. Figs. 3, 4, 5, 6, and 7 below show the confusion matrix for all emotion classes. Among these datasets, the class with the highest correct classifications in RAVDESS is the "Calm" class, in TESS it is "Surprise", in SAVEE it is "Neutral", in EMO-DB it is "Anger", and in CREMA-D it is "Happiness".

Training and validation accuracy and loss provide information on the learning performance of the model during training.

Figs. 8, 9, 10, 11, and 12 below show the training and validation accuracy and loss curves for all considered datasets.

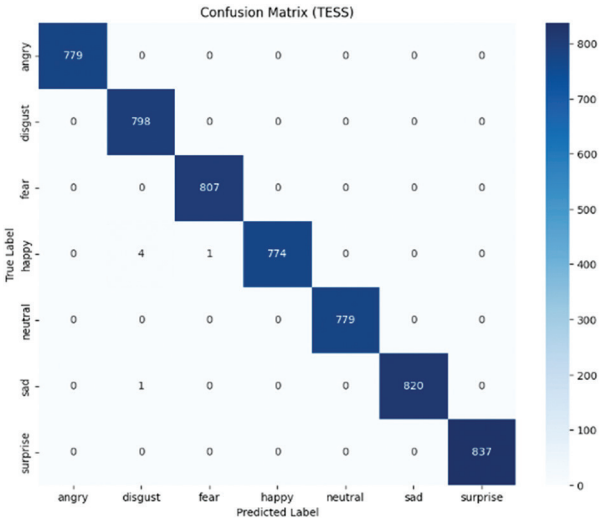


Fig. 3. Confusion matrices for the TESS dataset

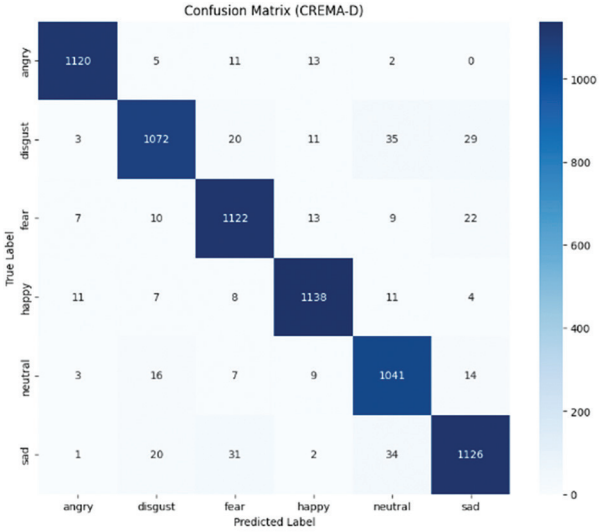


Fig. 4. Confusion matrices for the CREMA-D dataset

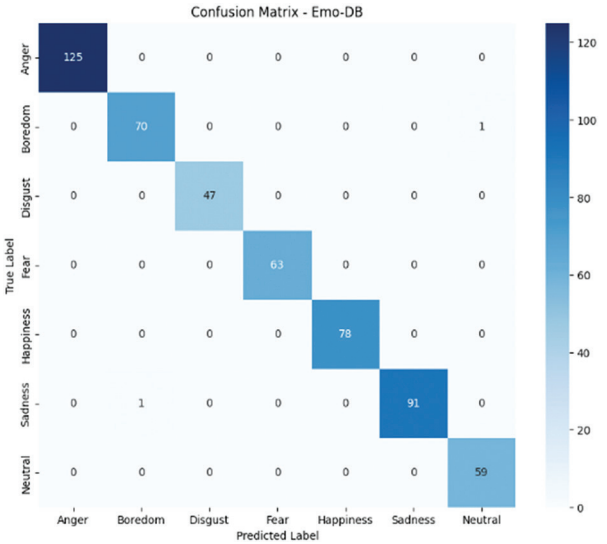


Fig. 5. Confusion matrices for the EMO-DB dataset

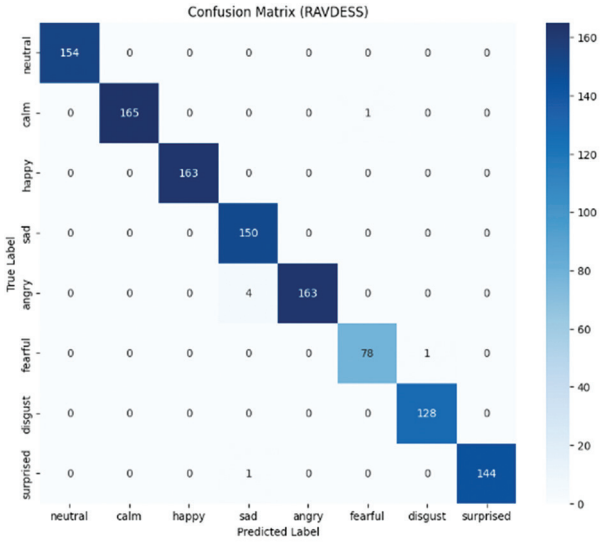


Fig. 6. Confusion matrices for the RAVDESS dataset

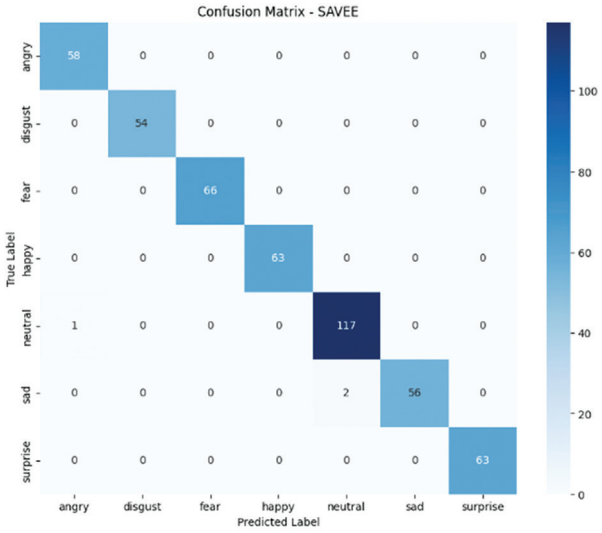


Fig. 7. Confusion matrices for the SAVEE dataset

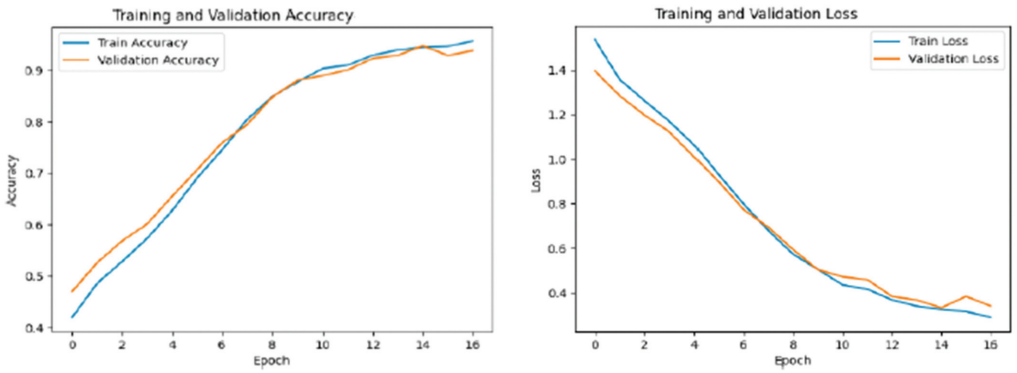


Fig. 8. Training vs. validation accuracy and loss for the CREMA-D dataset

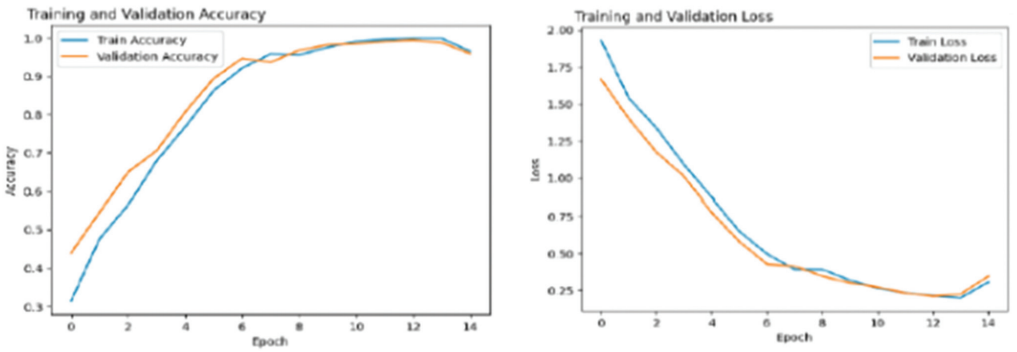


Fig. 9. Training vs. validation accuracy and loss for the RAVDESS dataset

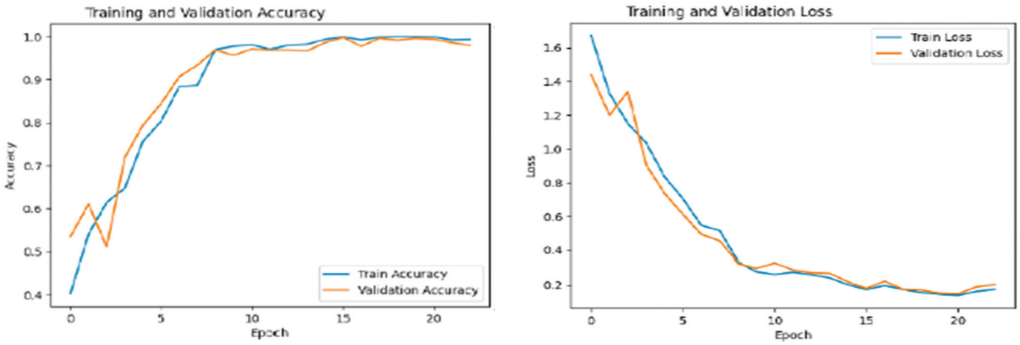


Fig. 10. Training vs. validation accuracy and loss for the SAVEE datasets

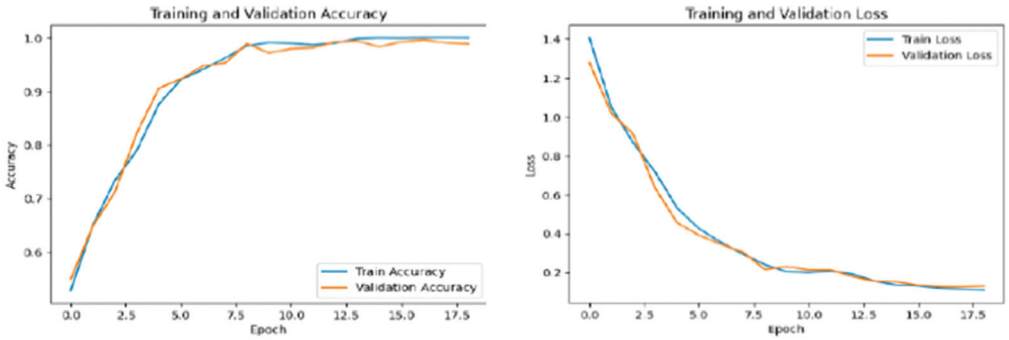


Fig. 11. Training vs. validation accuracy and loss for the EMO-DB datasets

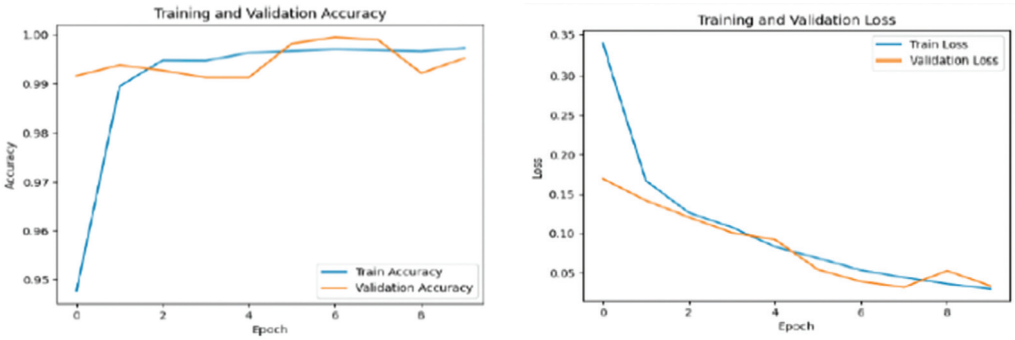


Fig. 12. Training vs. validation accuracy and loss for the TESS datasets

Among these datasets, the model achieves stable convergence and consistent performance across training and validation phases, with clear generalization observed for RAVDESS, TESS, SAVEE, EMO-DB, and CREMA-D datasets.

5.1. COMPARISON WITH THE STATE OF THE ART

The proposed model demonstrates better results compared to the existing works on datasets: RAVDESS, TESS, SAVEE, EMO-DB, and CREMA-D (see Table 6). This comparison study is carried out on the basis of the average classification accuracy of various models. Table 6 is based on [12] and further extended.

Table 6. Comparison of accuracy values of the proposed approach with state-of-the-art methods.

References/ Datasets	RAVDESS (%)	TESS (%)	SAVEE (%)	EMO-DB (%)	CREMA-D (%)
U. Asiya et al. [24]	68.00	-	-	-	-
D. Issa et al. [11]	71.61	-	-	86.10	-
S. Mekruksavanich et al. [25]	75.83	55.71	65.83	-	65.77
S. Jothiman et al. [26]	92.60	99.60	84.90	-	89.90
M. R. Ahmed et al. [27]	95.62	99.46	93.22	95.42	90.47
L. T. C. Ottoni et al. [28]	97.01	100.00	90.62	-	83.28
E. A. Alkhamali et al. [29]	96.30	-	96.50	99.86	-
R. M. Ben-Sauod et al. [30]	97.16	100.00	97.45	97.98	-
S. Y. Chowdhury et al. [12] (Config. 1)	97.83	100.00	97.02	100.00	95.10
S. Y. Chowdhury et al. [12] (Config. 2)	97.99	99.99	96.10	99.26	94.41
Tuncer et al. [31]	87.43	-	84.79	90.09	-
Saleem et al. [32]	-	-	-	94.2	-
Our proposed model	99.89	100.00	99.82	100.00	95.64

6. CONCLUSION

In this paper, a parallel hybrid model combining CNN, BiLSTM, and Transformer architectures was proposed for speech emotion recognition. The model enables the independent extraction of local, temporal, and contextual features while preserving the integrity of the learned representations.

The model was evaluated on five datasets: RAVDESS, TESS, SAVEE, CREMA-D, and EMO-DB, achieving outstanding accuracy rates of 99.89%, 100%, 99.82%, 95.64%, and 100%, respectively.

These results outperform previous studies and demonstrate the robustness and effectiveness of the proposed approach across emotional speech datasets.

Despite these promising results, validation in real-world environments remains a direction for future research, where the model will be evaluated under practical conditions.

DECLARATION OF THE USE OF AI TOOLS

The authors used ChatGPT and Grammarly for language improvement and error checking. All AI-assisted content was reviewed and edited by the authors, who take full responsibility for the integrity and accuracy of this work.

No AI tool is listed as an author, and AI tools are not cited as primary sources.

REFERENCES

[1] E. Jordan, R. Terrisse, V. Lucarini, M. Alrahabi, M. Krebs, J. Desclés, C. Lemey, "Speech Emotion Recognition in Mental Health: Systematic Review of Voice-Based Applications", JMIR Mental Health, Vol. 12, 2025, p. e74260, doi: 10.2196/74260.

[2] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, A. Othmani, "AudVowelConsNet: A Phoneme-Level Based Deep CNN Architecture for Clinical Depression Diagnosis", Machine Learning with Applications, Vol. 2, 2020, p. 100005. doi: 10.1016/j.mlwa.2020.100005.

[3] A. Vázquez-Romero, A. Gallardo-Antolín, "Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks", Entropy, Vol. 22, No. 6, 2020, p. 688. doi: 10.3390/e22060688.

[4] W. Alsabhan, "Human–Computer Interaction with a Real-Time Speech Emotion Recognition Using Ensembling Techniques, 1D Convolution Neural Network and Attention", Sensors, Vol. 23, No. 3, 2023, p. 1386. doi: 10.3390/s23031386.

[5] A. A. Alnuaim, M. Zakariah, A. Alhadlaq, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, "Human-Computer Interaction with Detection of Speaker Emotions Using Convolutional Neural Networks", Computational Intelligence and Neuroscience, Vol. 2022, 2022, p. 7463091. doi: 10.1155/2022/7463091.

[6] R. Ullah, M. Asif, W. A. Shah, F. Anjam, I. Ullah, T. Khurshaid, L. Wuttisittikulkij, S. Shah, S. M. Ali, M. Alibakhshikenari, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer", Sensors, Vol. 23, No. 13, 2023, p. 6212. doi: 10.3390/s23136212.

[7] H. Li, B. Baucom, P. Georgiou, "Linking Emotions to Behaviors Through Deep Transfer Learning", PeerJ Computer Science, Vol. 6, 2020, p. e246. doi: 10.7717/peerj-cs.246.

[8] F. Zhu-Zhou, R. Gil-Pita, J. García-Gómez, M. Rosa-Zurera, "Robust Multi-Scenario Speech-Based Emotion Recognition System", Sensors, Vol. 22, No. 6, 2022, p. 2343. doi: 10.3390/s22062343.

[9] S. Kshirsagar, T. H. Falk, "Cross-Language Speech Emotion Recognition Using Bag-of-Word Representations, Domain Adaptation, and Data Augmentation", Sensors, Vol. 22, No. 17, 2022, p. 6445. doi: 10.3390/s22176445.

[10] Z. Dair, R. Donovan, R. O'Reilly, "Linguistic and Gender Variation in Speech Emotion Recognition Using Spectral Features", arXiv.2112.09596, 2021. doi: 10.48550/arXiv.2112.09596

[11] M. Dias Issa, F. Demirci, A. Yazici, "Speech Emotion Recognition with Deep Convolutional Neural Networks", Biomedical Signal Processing and Control, Vol. 59, 2020, p. 101894. doi: 10.1016/j.bspc.2020.101894.

[12] S. Y. Chowdhury, B. Banik, M. T. Hoque, S. Banerjee, "A Novel Hybrid Deep Learning Technique for Speech Emotion Detection Using Feature Engineering", Proceedings of the Fourth International Conference on Hybrid Human-Artificial Intelligence Workshops, Pisa, Italy, 9-13 June 2025. doi: 10.48550/arXiv.2507.07046.

[13] V. Sareen, K. R. Seeja, "Speech Emotion Recognition Using Mel Spectrogram and Convolutional Neural Networks", Procedia Computer Science, Vol. 258, 2025, pp. 3693–3702. doi: 10.1016/j.procs.2025.04.624.

[14] A. Slimi, H. Nicolas, M. Zrigui, "Hybrid Time Distributed CNN-Transformer for Speech Emotion Recognition", Proceedings of the 17th International Conference on Software Technologies, Lisbon, Portugal, 11-13 July 2022, pp. 602-611. doi: 10.5220/0011314900003266.

[15] C. S. A. Kumar, A. D. Maharana, S. M. Krishnan, S. S. S. Hanuma, G. J. Lal, V. Ravi, "Speech Emotion

Recognition Using CNN-LSTM and Vision Transformer", Innovations in Bio-Inspired Computing and Applications, Lecture Notes in Networks and Systems, Vol. 649, Springer, 2023, pp. 86-97. doi: 10.1007/978-3-031-27499-2_8.

[16] W. Bouchelligua, R. Al-Dayil, A. Algaith, "Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks", Applied Sciences, Vol. 15, No. 4, 2025, p. 2114. doi: 10.3390/app15042114.

[17] J. Zhao, X. Mao, L. Chen, "Speech Emotion Recognition Using Deep 1D & 2D CNN-LSTM Networks", Biomedical Signal Processing and Control, Vol. 47, 2019, pp. 312-323. doi: 10.1016/j.bspc.2018.08.035.

[18] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, S. L. Schneider, "Speech Emotion Recognition Using Machine Learning—A Systematic Review", Intelligent Systems with Applications, Vol. 20, 2023, p. 200266. doi: 10.1016/j.iswa.2023.200266.

[19] S. R. Livingstone, 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, 2018, p. e0196391. doi: 10.1371/journal.pone.0196391.

[20] P. Jackson, S. Haq, "Surrey Audio-Visual Expressed Emotion (SAVEE) Database", University of Surrey, Guildford, UK, 2014.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, "A Database of German Emotional Speech", Proceedings of Interspeech 2005, pp. 1517-1520. doi: 10.21437/Interspeech.2005-446.

[22] M. K. Pichora-Fuller, K. Dupuis, "Toronto Emotional Speech Set (TESS)", doi: 10.5683/SP2/E8H2MF.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset", IEEE Transactions on Affective Computing, Vol. 5, No. 4, 2014, pp. 377-390. doi: 10.1109/TAFFC.2014.2336244.

[24] U. A. Asiya, V. K. Kiran, "Speech Emotion Recognition - A Deep Learning Approach", Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), Palladam, India, 11-13 November 2021, pp. 867-871. doi: 10.1109/I-SMAC52330.2021.9640995.

[25] S. Mekruksavanich, A. Jitpattanakul, N. Hnoohom, "Negative Emotion Recognition Using Deep Learning for Thai Language", Proceedings of the Joint International Conference on Digital Arts, Media and Technology and ECTI Northern Section Conference, Pattaya, Thailand, 11-14 March 2020, pp. 71-74. doi: 10.1109/ECTIDAMTNCON48261.2020.9090768.

[26] S. Jothimani, K. Premalatha, "MFF-SAug: Multi-Feature Fusion with Spectrogram Augmentation for Speech Emotion Recognition Using Convolution Neural Networks", Chaos, Solitons & Fractals, Vol. 162, 2022, p. 112512. doi: 10.1016/j.chaos.2022.112512.

[27] M. R. Ahmed, S. Islam, A. K. M. Muzahidul Islam, S. Shatabda, "An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition", Expert Systems with Applications, Vol. 218, 2023, p. 119633. doi: 10.1016/j.eswa.2023.119633.

[28] L. T. C. Ottoni, A. L. C. Ottoni, J. d. J. F. Cerqueira, "A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta-Learning", Electronics, Vol. 12, 2023, p. 4859. doi: 10.3390/electronics12234859.

[29] E. A. Alkhamali, A. Allinjawi, R. B. Ashari, "Combining Transformer, Convolutional Neural Network, and Long Short-Term Memory Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition in Distance Education Classrooms", Applied Sciences, Vol. 14, No. 12, 2024, p. 5050. doi: 10.3390/app14125050.

[30] R. M. Ben-Sauod, R. S. Alshwehdi, W. I. Eltarhouni, "Enhancing Speech Emotion Recognition Through a Cross-Dataset Analysis: Exploring Improved Models", Proceedings of the IEEE 4th International Maghreb Meeting on Sciences and Techniques of Automatic Control and Computer Engineering, Tripoli, Libya, 19-21 May 2024, pp. 706-711. doi: 10.1109/MI-STA61267.2024.10599702.

[31] T. Tuncer, S. Dogan, U. R. Acharya, "Automated Accurate Speech Emotion Recognition System Using Twine Shuffle Pattern and Iterative Neighborhood Component Analysis Techniques", Knowledge-Based Systems, Vol. 211, 2021, p. 106547. doi: 10.1016/j.knosys.2020.106547.

[32] N. Saleem, J. Gao, R. Irfan, A. Almadhor, H. T. Rauf, Y. Zhang, S. Kadry, "DeepCNN: Spectro-Temporal Feature Representation for Speech Emotion Recognition", CAAI Transactions on Intelligence Technology, Vol. 8, No. 2, 2023, pp. 401-417. doi: 10.1049/cit2.12233.