Speech-Based Emotion Recognition and Classification Integrating a CNN and BiLSTM Network

Fatima Uroosa[1], **Asim Abbas**[2]*, **Muhammad Tayyab Zamir**[1], **Grigori Sidorov**[1]
[1]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),
Av. Juan de Dios Batiz, s/n, Mexico City 07320, Mexico
[2]School of Computer Science, University of Birmingham, B15 2TT, UK
{furoosa2024, mzamir2023, sidorov}@cic.ipn.mx, axa2233@student.bham.ac.uk

Abstract

Speech emotion recognition (SER) has gained significant interest in recent times, which utilizes speech signals to identify the emotional state of speakers. Accurate recognition of subtle emotional variations in speech, such as distinguishing closely related emotional states, remains a challenging problem due to the variability of speech signals and the acoustic similarity among emotion classes across different speakers and linguistic contexts. This paper proposes a hybrid deep learning model that integrates a Convolutional Neural Network (CNN) with a Bidirectional Long Short-Term Memory (BiLSTM) network to effectively identify both spectral and temporal features of speech. The log Mel-frequency spectral coefficients (MFSC) are used as input features to represent discriminative spectral representations, while the BiLSTM layer model represents long-range temporal dependencies in speech signals. The proposed framework is evaluated on the Toronto Emotional Speech Set (TESS), a publicly available dataset of acted emotional speech containing seven emotion classes. The experimental findings show that the hybrid CNN-BiLSTM achieved an overall classification accuracy of 96.36%, significantly outperforming baseline models including GRU (91.84%), BiLSTM (93.12%), and CNN–GRU (94.67%). These findings highlight the effectiveness of combining spectral and temporal modeling for improved speech emotion recognition performance. Furthermore, our CNN+BiLSTM approach offers a computationally efficient and data-efficient alternative to transformer-based models, while still effectively capturing both spatial and temporal emotional cues in speech, making it suitable for real-time and resource-constrained applications.

Keywords: Speech emotion recognition, Deep Learning, Convolutional Neural Networks, Bidirectional LSTM, Hybrid CNN–BiLSTM Model, Log Mel-Frequency Spectral Coefficients (MFSC)

1. Introduction

Emotions are complex psychological and physiological states of consciousness influenced by factors such as environmental challenges, attitudes, and personality. Emotions play a crucial role in human life and significantly affect our psychological and physical health, as well as decision-making (Lerner et al., 2015), (Gross, 1998). It is important to understand the nature of emotional states, as it helps to assess psychological conditions, such as mental health issues like depression, anxiety, and bipolar disorder, that affect millions of people across the world (World Health Organization, 2025).

Moreover, emotions are multidimensional and are often expressed at the same time in many different ways, including tonal variations of voice, facial expression, language, and other forms of nonverbal communication. Proper understanding of these expressions is critical in the development of human-computer interaction, mental health diagnosis, and security systems. This intermodal manifestation of emotions is a major challenge for computational recognition systems, which require interdisciplinary methods to examine and decipher complex emotional states from different data sources (Dao et al., 2022). Among the different modalities used for emotion recognition, speech has gained significant attention due to its natural availability, low acquisition cost, and rich emotional content. Speech-based emotion recognition is increasingly used across various fields, including healthcare (Xiefeng et al., 2019), Human-Computer Interaction (Alsabhan, 2023), mental health assessment and psychological monitoring (Jordan et al., 2025), and assistive clinical care (Vaida and Huang, 2025). Several recent studies have further highlighted the effectiveness of Speech Emotion Recognition (SER) in real-world applications (Botelho et al., 2024). In general, speech signals encode two types of complementary information: linguistic and paralinguistic. Linguistic information is spoken content, including language structure, accent, and dialect, whereas paralinguistic information conveys cues about the speaker's emotional state, attitude, context, and environmental surroundings. Emotional recognition can thus only be achieved through the usage of paralinguistic attributes, as emotional states are usually conveyed by how something is said rather than what is said. However, emotional expression also differs across languages and cultures and may be subjective; hence, the speech emotion recognition task is challenging (Chen et al., 2021).

Furthermore, the challenge arises from the presence of overlapping or similar emotional states, such as calm vs. neutral and excited vs. happy, which are hard to differentiate using a simple decision boundary. The initial methods were based on

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 29–37
16 May 2026.

manually created low-level acoustic attributes, such as pitch, energy, and formants, and classical machine learning (ML) classifiers, such as support vector machines (SVM), k-nearest neighbors (k-NN), random forests (RF), and decision trees, yielding moderate performance results (Abdelfattah et al., 2016). However, more recently, deep learning (DL) approaches that learn high-level paralinguistic representations directly from data have shown better performance. Developments in neural structures, such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and gated recurrent units (GRUs), have greatly enhanced the ability to model the dynamics of speech signals over time, resulting in more powerful speech emotion recognition systems (Jahangir et al., 2021).

In contrast to prior works that focus on proposing a single deep architecture, this study presents a comprehensive comparative analysis of multiple deep learning models for speech emotion recognition using a unified experimental setup. LSTM, BiLSTM, GRU, CNN, and a hybrid CNN–BiLSTM model are evaluated on the Toronto Emotional Speech Set (TESS) dataset using MFCC features. The results indicate that sequence-based models such as LSTM, BiLSTM, and GRU consistently outperform CNN-only architectures, highlighting the importance of temporal dynamics in emotional speech. The hybrid CNN–BiLSTM model further improves performance by jointly modeling local spectral patterns and long-range temporal dependencies.

1.1. Research Questions

This study aims to investigate the effectiveness of a hybrid CNN–BiLSTM architecture for speech emotion recognition. The following research questions are addressed:

RQ1: How effectively can the proposed CNN–BiLSTM model learn and combine spectral and temporal features for accurate emotion classification.

RQ2: Does the proposed hybrid architecture outperform baseline models such as CNN, LSTM, and GRU in terms of accuracy and F1-score on the TESS dataset.

RQ3: How well does the model distinguish between acoustically similar emotions in speech signals.

RQ4: What is the impact of combining CNN-based feature extraction with BiLSTM-based temporal modeling on overall system performance.

The remaining paper is structured as follows. Section 2 reviews related work and presents the problem statement for speech emotion recognition. Section 3 describes the methodology, including dataset acquisition, preprocessing, feature extraction, and the proposed hybrid CNN–BiLSTM architecture. Section 4 outlines the experiment configuration, metrics of evaluation, and performance analysis. Section 5 presents the results, and Section 6 discusses the limitations of the study. Lastly, Section 7 brings the paper to an end.

2. Related Work

Due to the increasing significance of Speech emotion recognition (SER) in social sciences and biomedical applications over a period of time (Atmaja and Sasou, 2022), there has been a growing interest in the development of numerous DL-based methods in emotion classification (Monisha and Sultana, 2022). Some studies have analyzed the issues of SER and the different model structures to improve recognition performance by utilizing multiple attention mechanisms to combine both audio features (MFCCs) and text embeddings (Zhang, 2023). Their multimodal approach achieved better emotion recognition results than using audio or text alone.

(Kumbhar and Bhandari, 2019) trained an LSTM-based model to perform emotion recognition, using extracted Mel-Frequency Cepstral Coefficients (MFCCs) features of the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), a benchmark emotional speech dataset containing both speech and song recordings. Their approach achieved an accuracy of 84.81%; however, a high loss value of 67.21% was observed during implementation, highlighting the need to reduce false-positive predictions for improved system performance.

(Zehra et al., 2021) proposed an ensemble-based approach for cross-corpus multilingual emotion recognition using a majority voting strategy. Their study employed speech corpora in Urdu, Italian, English, and German and tested classifiers, including SVM, random forest, and decision trees, and classified the English (TESS) dataset at 70.14% under within-corpus evaluation. (Luna-Jiménez et al., 2021) explored transfer learning for SER by leveraging knowledge from supervised pre-trained models, specifically CNN-14 from the PANNs framework (Kong et al., 2020). Though the identification of the multilingual machine-generated text is done with the help of transformer-based machine learning and multilingual modeling methods, there are trends in efficient representation learning, which are applicable in cross-modal emotion recognition (Abiola et al., 2025).

By fine-tuning this model on the RAVDESS dataset using speech signals, the authors achieved an accuracy of 76.58%. More recently, research in speech processing has increasingly shifted toward self-supervised learning (SSL) approaches that pretrain speech representations directly from

raw audio signals rather than relying solely on supervised learning architectures. When fine-tuned on benchmark datasets, SSL-based models have demonstrated improved performance, particularly in low-resource scenarios. For example, (Gavali and Verma, 2025) reported an accuracy of 81.82% on the RAVDESS dataset by using wav2vec2-xlsr as a feature extractor combined with a multilayer perceptron (MLP), outperforming their earlier supervised CNN- and PANNs-based approaches, which achieved 76.58% accuracy. Even though this previous work dealt with text-based emotion intensity detection, the information on lightweight model design is encouraging to apply comparable efficient designs to speech-based emotion classification (Eyasu et al., 2025).

Recent studies on deep learning SER have further analyzed hybrid convolutional-recurrent architectures to identify both spectral and temporal speech audio patterns. (Liu et al., 2024) developed a CNNLSTM algorithm that includes attention and multi-feature fusion, and demonstrated excellent results with several benchmark corpora, such as SAVEE and CASIA. (Ahmed et al., 2023) Demonstrate an ensemble 1D-CNN-LSTM-GRU architecture with data augmentation, which enhanced the generalization of five standard datasets. (Bhanbhro et al., 2025) compared CNN-LSTM and attention-enhanced CNN-LSTM architectures, reporting competitive results on the RAVDESS dataset. Other studies examined real-time CNN +BiLSTM architecture augmented with vocal features, evaluated on benchmark datasets such as TESS, RAVDESS, and EmoDB (Berlin Database of Emotional Speech), a widely used benchmark dataset for speech emotion recognition. Furthermore, an attention-based CNN-BiLSTM architecture was proposed to capture both local and global contextual speech signals, achieving improved performance on several datasets.

Previous studies have identified several limitations in speech emotion recognition. (Jahangir et al., 2021) highlights challenges such as limited dataset size, reliance on acted emotional speech, and difficulties in modeling temporal dependencies. Similarly, (Singh and Goel, 2022) emphasizes issues related to feature selection, model generalization, and the difficulty of distinguishing acoustically similar emotions. These limitations indicate the need for more robust architectures. In this work, the proposed CNN–BiLSTM model addresses these challenges by effectively combining spectral feature extraction with temporal sequence modeling, leading to improved emotion recognition performance.

2.1. Problem Description

Speech Emotion Recognition (SER) is to identify the speaker's emotional state based on the speech signals using paralinguistic information like pitch, energy, spectral patterns, etc. Despite the recent significant progress, SER is still a difficult task because the variability of speech signals is very important, and the acoustic similarity between emotion classes, such as happy and surprise or neutral and calm, is quite significant (Singh and Goel, 2022) (Jahangir et al., 2021). Emotional dynamics are shared over both the local spectral properties and long-range time dynamics that cannot be easily modeled with conventional machine learning methods of handcrafted features and shallow classifiers (Atmaja and Sasou, 2022).

More recent deep learning models have enhanced the SER performance by learning a hierarchical representation of time-frequency directly based on time-frequency features (Jahangir et al., 2021). Convolutional Neural Networks (CNNs) are effective to use in order to capture local spectral patterns, whereas recurrent networks like Long Short-Term Memory (LSTM) networks are designed to capture the temporal dependencies in speech. However, the model's capacity to completely capture both spectral and temporal information is constrained when either approach is used separately. This encourages the adoption of a hybrid CNN–BiLSTM architecture for robust unimodal speech emotion identification that simultaneously incorporates long-range temporal correlations and local spectral characteristics.

3. Methodology

This section explains the overall workflow of the proposed hybrid CNN-BiLSTM models for speech emotion classification, as shown in Figure 1.

3.1. Dataset

The experiments in this research study are performed on the TESS (Pichora-Fuller and Dupuis, 2020), a publicly available dataset for speech emotion recognition (SER). TESS contains emotional speech samples from two native English speakers recorded in a controlled environment. The dataset consists of seven emotional categories: angry, disgust, fear, happy, sad, surprise, and neutral. Each emotion class contains an equal number of audio samples, making the dataset balanced and suitable for supervised learning. The controlled recording environment and clear articulation of emotions facilitate reliable evaluation of SER models. In this study, TESS is used to train and evaluate the proposed hybrid CNN–BiLSTM framework.

3.2. Data Preprocessing

All audio samples are preprocessed to ensure consistency. Preprocessing steps include normaliza-

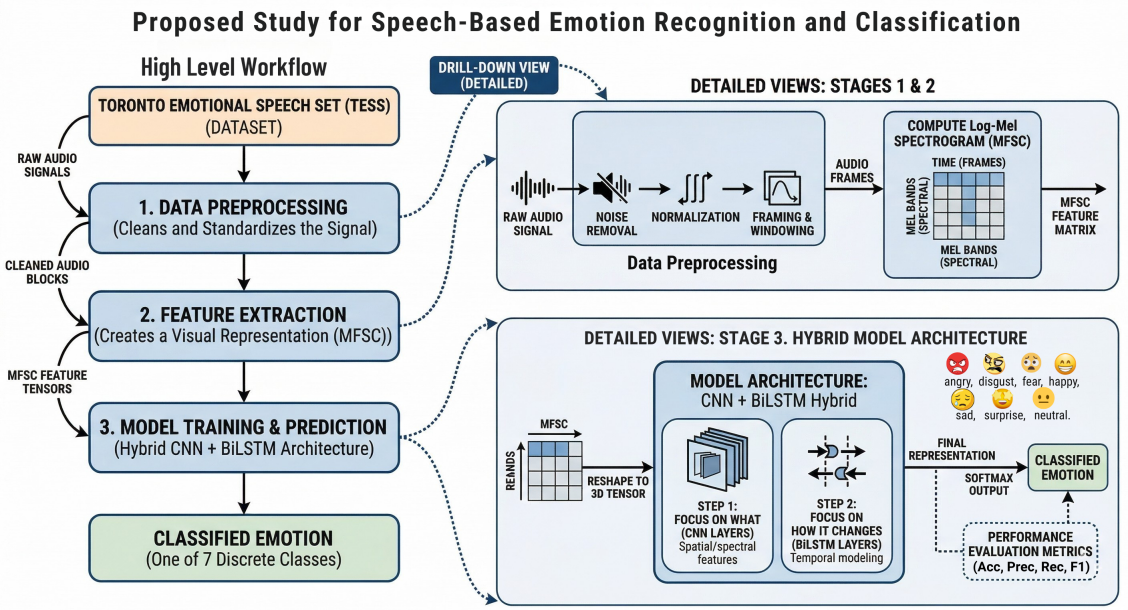
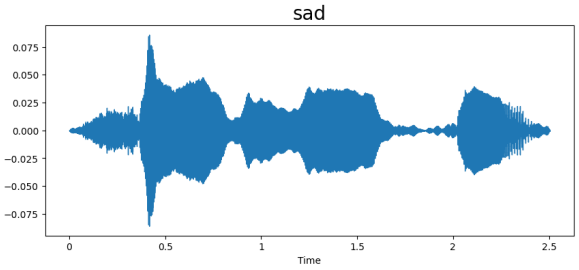


Figure 1: Overall workflow of the proposed CNN–BiLSTM-based speech emotion recognition framework.

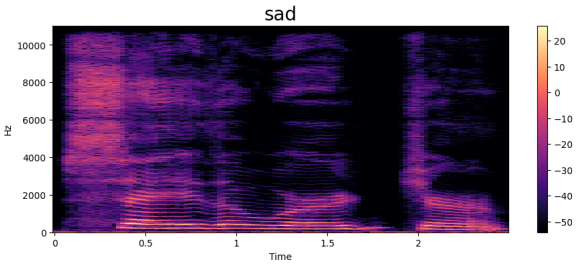
tion of audio amplitude, resampling to a standard sampling rate, and trimming of silence at the start and end of recordings. These steps reduce variability and improve model performance.

3.3. Feature Extraction

Log Mel-filterbank spectral coefficients (MFSCs) are extracted from each speech signal and used to train the proposed hybrid CNN–BiLSTM model. MFSC features are chosen because they generally outperform MFCCs in capturing spectral information for emotion recognition. Since speech recordings vary in duration, the extracted feature sequences result in variable-length representations. To ensure a consistent input format for the DL models, zero-padding is applied to all MFSC sequences, producing uniform feature dimensions across samples. A sample speech signal along with its corresponding spectrogram and chromagram are shown in Figures 2a and 2b.



(a) Waveplot for Sad Emotion



(b) Spectrogram for Sad Emotion

Figure 2: Waveform and spectrogram representations for the sad emotion sample.

3.4. Proposed Hybrid Deep CNN-BiLSTM Model

To jointly learn spectral and temporal representations, a hybrid CNN-BiLSTM architecture is proposed for speech emotion recognition. Log-Mel spectrogram (MFSC) features extracted from speech signals are first fed into a CNN component, which captures high-level local spectral patterns through convolution and pooling operations.

The output feature maps generated by the CNN are reshaped into a sequential representation and passed to the BiLSTM layer to model temporal dependencies across speech frames. By incorporating bidirectional temporal context, the BiLSTM effectively captures both short-term and long-term emotional characteristics of speech.

Finally, the BiLSTM outputs are forwarded to fully connected layers followed by a softmax classifier to predict emotion categories such as *neutral*, *happy*, *sad*, and *angry*. The entire CNN-BiLSTM model is trained end-to-end using categorical cross-entropy loss. The CNN-BiLSTM model is proposed to include a 1D convolutional layer (32 filters) and max-pooling and dropout, a Bi-Directional LSTM (128-dimensional output), and two fully connected lay-

Table 1: Architecture and training hyperparameters of the proposed CNN–BiLSTM model

Component	Configuration
Input Features	Log-Mel Spectrogram (MFSC)
Convolution Layer	Conv1D, 32 filters
Activation Function	ReLU
Pooling Layer	MaxPooling1D
Recurrent Layer	Bidirectional LSTM (64 units per direction)
Dense Layer	64 units (ReLU)
Output Layer	7 units (Softmax)
Dropout Rate	0.5
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Evaluation Metric	Accuracy
Total Parameters	58,503

ers (64 hidden units) and a 7-class softmax output layer. The model is trained on the Adam optimizer and categorical cross-entropy loss. The number of trainable parameters is 58,503, as shown in Table 1." The relatively small number of trainable parameters highlights the computational efficiency of the proposed architecture. Compared to deeper or attention-based models, this lightweight design makes the model suitable for real-time applications and deployment in resource-constrained environments.

4. Experimental Setup

All experiments were conducted in a GPU-enabled Google Colab environment. The proposed models were trained and evaluated on an NVIDIA Tesla T4 GPU with 16 GB of GPU memory and 12.7 GB of system RAM. The implementation was carried out in Python using the TensorFlow-Keras framework with CUDA version 12.1 support. This setup enabled efficient training and evaluation of the proposed DL architectures for SER.

The experiments were carried out on two publicly available benchmark datasets: TESS and SAVEE. The TESS dataset consists of 2,800 speech utterances, while the SAVEE dataset contains 480 utterances. The combined dataset was split into training, validation, and testing sets following a ratio consistent with prior work, resulting in 2,296 samples for training, 656 for validation, and 328 for testing.

4.1. Evaluation Metrics and Performance Analysis

The performance of the proposed hybrid CNN-BiLSTM model was evaluated using standard multiclass classification metrics, including accuracy, precision, recall, and F1-score, along with a confusion matrix. These metrics provide a comprehensive assessment of the model's ability to classify different emotion categories.

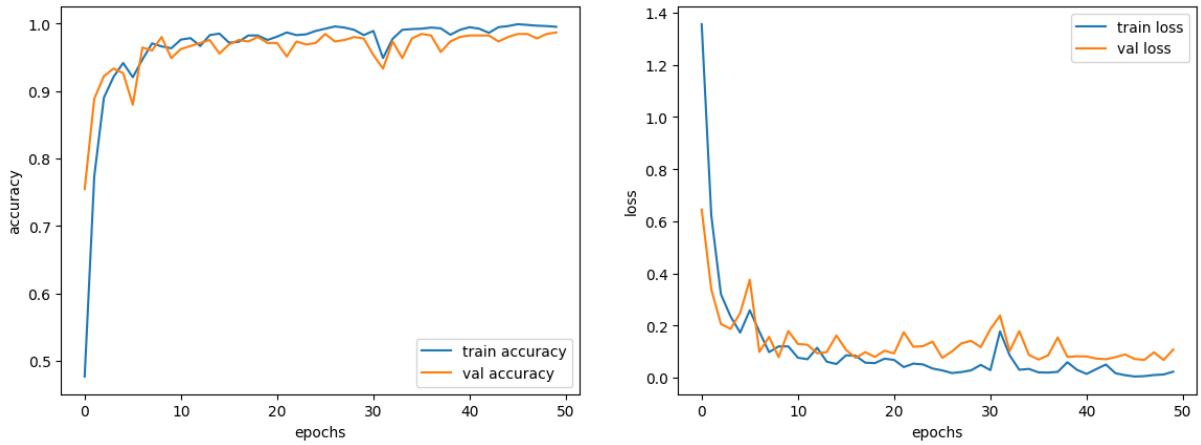
The confusion matrix summarizes the classification results by comparing the predicted labels with the ground truth labels, providing insights into both correct and incorrect predictions across emotion categories, as shown in 4.

5. Results

The hybrid CNN-BiLSTM model was trained on the TESS dataset and evaluated on a distinct test set to determine its capability in speech emotion recognition. The analysis focused on conventional classification metrics, including accuracy, precision, recall, and F1 Score. The class-wise performance of the proposed model on the TESS dataset is presented in Table 2. The results indicate that the Recognition performance is high across the majority of emotion categories, which shows the effectiveness of utilizing convolutional feature extraction and bidirectional temporal modeling. Figure 3a shows the accuracy trends, while Figure 3b illustrates the loss convergence of the proposed model. The training and validation curves show stable convergence, with the validation accuracy following the training accuracy over the epochs. The small gap between the curves indicates that the model does not suffer from significant overfitting. Furthermore, the loss in the validation process gradually decreases and stabilizes after convergence, which shows that the spectral and temporal representations learned are effectively being generalized.

To further test the architecture, performance was compared with baseline models, including GRU, BiLSTM, and CNN-GRU. Table 3 summarizes the overall classification performance of each of the models. The CNN–BiLSTM model with the highest accuracy and F1-score across assessment methods indicates the usefulness of combining convolutional feature extraction with bidirectional temporal modeling in speech emotion recognition. The proposed hybrid CNN-BiLSTM model was tested to determine its effectiveness in SER. The high performance of CNN–BiLSTM compared to CNN–GRU and standalone recurrent architectures indicates the significance of the bidirectional temporal modeling in SER. While convolutional layers effectively capture local spectral patterns, the BiLSTM layer facilitates the incorporation of context information of previous and subsequent frames. This complementary combination helps to increase the capacity of the model to identify the subtle emotional patterns that develop over time.

The model attained an overall classification accuracy of **96.36%** and an F1-score of **0.93**, on the TESS dataset, indicating improved performance compared to the baseline architectures. The class-wise analysis suggests that the model performs well, specifically on emotions such as neutral, surprise, and sadness, and is characterized by high precision and recall. Happy and disgust were ob-



(a) Training and validation accuracy

(b) Training and validation loss

Figure 3: Learning curves of the proposed CNN–BiLSTM model showing (a) training and validation accuracy and (b) training and validation loss.

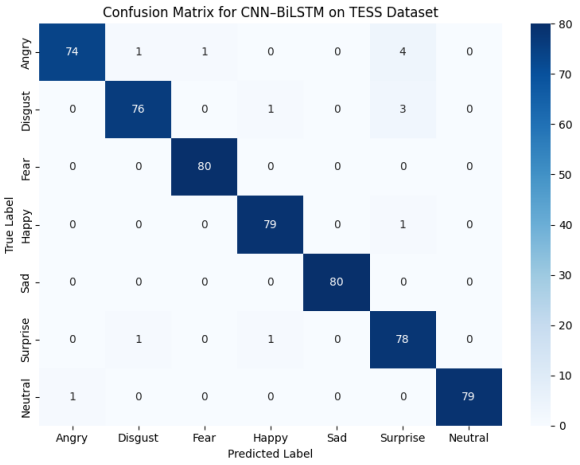


Figure 4: Confusion matrix of the proposed CNN–BiLSTM model on the TESS dataset.

served slightly lower, and this can be attributed to the acoustic similarity of these two emotion classes with other types of emotions, which has been a commonly observed challenge in speech emotion recognition.

In Figure 4, the confusion matrix, it can be observed from the confusion matrix that the number of samples along the diagonal is high, which implies that most of the emotion samples are accurately identified. Misclassifications are minimal and mainly occur with acoustically related emotions, e.g., between anger and surprise or between surprise and disgust. These observations are correlated with the precision and recall rates of the classes as mentioned in Table 2 and further confirm the strength of the proposed CNN–BiLSTM architecture.

Our experimental evaluation confirms that the hybrid CNN-BiLSTM framework is effective in capturing local spectral patterns and long-term temporal dependencies of speech signals. The model can differentiate between various categories of emotions by using convolutional feature extraction with bidirectional temporal modeling, even when subtle acoustic variations are present. In general, the findings indicate that the suggested method can be considered a valid solution for unimodal speech emotion recognition in a controlled experimental environment.

Emotion	Precision	Recall	F1-score	Acc. (%)
Surprise	0.98	0.98	0.98	96.42
Neutral	1.00	0.94	0.97	99.02
Sad	0.91	0.95	0.93	98.05
Disgust	0.97	0.73	0.84	95.77
Fear	0.87	0.77	0.82	96.10
Angry	0.91	0.73	0.81	93.83
Happy	0.68	0.90	0.78	95.12

Table 2: Class-wise performance metrics of the proposed CNN–BiLSTM model on TESS.

Model	Precision	F1-score	Acc. (%)
GRU	0.88	0.87	91.84
BiLSTM	0.90	0.89	93.12
CNN–GRU	0.91	0.90	94.67
CNN–BiLSTM	**0.93**	**0.93**	**96.36**

Table 3: Performance comparison of different models on the TESS dataset.

The proposed model has good performance in most of the emotion categories, as indicated in 2. The maximum F1-score is achieved with surprise (0.98), then neutral, and sad, which points to the fact that these feelings have their own acoustic peculiarities. Comparatively worse precision is, on the other hand, found with happy, which implies that there is a possible overlap with other acoustically

Table 4: Comparison of Proposed Model with the existing state-of-the-art

Author	Dataset-Language	Methodology used	Accuracy (%)
(Wang et al., 2020)	IEMOCAP-English	Dual-Sequence LSTM	73.3
(Kumaran et al., 2021)	RAVDESS-English	Deep C-RNN	75
(Zhang et al., 2021)	CASIA-Chinese	Parallel Convolution Bi-LSTM	79.67
(Senthilkumar et al., 2022)	EMO-DB-English	Bi-directional LSTM and DBN	85.57
(Muppidi and Radfar, 2021)	EMO-DB-English	QCNN	88.7
(Jahangir et al., 2021)	Benchmark-English	Deep Neural Network	80–90
(Prasanna et al., 2022)	Standard-English	CNN–BiLSTM	∼92
Proposed Model	**TESS-English**	**CNN–BiLSTM**	**96.36**

similar classes. The difference among classes indicates the natural complexity of separating the emotions having slight variations in prosodies, which is a frequently described problem in speech emotion recognition.

6. Comparative Analysis with Existing Studies

To present a comprehensive evaluation of the proposed CNN–BiLSTM model on the TESS dataset, we compare our results with existing state-of-the-art speech emotion recognition methods utilize different datasets, as shown in Table 4. The proposed approach achieves an accuracy of 96.36%, outperforming earlier approaches such as Dual LSTM (73.3%) and Deep C-RNN (75%), which are limited in their ability to capture complex spectral–temporal features.

Compared to CNN–BiLSTM-based and hybrid architectures, including those reported by (Zhang et al., 2021), (Senthilkumar et al., 2022), and (Muppidi and Radfar, 2021), which achieve accuracies of 79–88.7%, the proposed model reported improved performance. Additionally, prior studies such as (Jahangir et al., 2021) and (Prasanna et al., 2022)report accuracy of 80–92%, further supporting the competitiveness of the proposed approach.

Overall, the results indicate that the proposed CNN–BiLSTM model effectively captures both spectral and temporal features using TESS dataset only, achieving high accuracy while maintaining a relatively simple and efficient architecture.

7. Limitations of the Proposed Study

Despite achieving strong performance, the proposed CNN-BiLSTM model has certain limitations. First, the experiments are conducted on controlled benchmark datasets, TESS, which contain acted emotions recorded in relatively clean acoustic environments. As a result, the model's generalization to real-world, noisy, and spontaneous speech scenarios remains to be further investigated, as highlighted in prior studies (Jahangir et al., 2021) (Singh and Goel, 2022).

Second, the class imbalance and limited number of samples for certain emotions may affect recognition performance, particularly for emotions with overlapping acoustic characteristics (e.g., happy vs. surprise or neutral vs. calm), leading to misclassification. Such overlap underscores that it is not easy to differentiate emotions using acoustic characteristics alone, especially when dealing with emotions exhibiting subtle prosodic variations. Furthermore, the suggested framework only pays attention to the unimodal acoustic characteristics and excludes linguistic or multimodal data of a text or facial expression. It has been proven in the past that the performance of emotion recognition can be enhanced further when a variety of modalities is used, particularly in a complex and naturalistic context. The consideration of these limitations through the assessment of the model in the case of spontaneous speech datasets and further development of the framework into multimodal learning is a significant future direction of research.

8. Conclusion

This paper presented a hybrid deep CNN-BiLSTM model for SER, effectively integrating convolutional feature extraction with bidirectional temporal modeling to capture both spectral and temporal characteristics capture both spectral and long-range temporal characteristics of speech signals. Using log-Mel spectrogram and MFSC features, the proposed model was trained and evaluated on the TESS dataset only, achieving consistent performance across all emotion classes.

Experimental results show that the CNN-BiLSTM model outperforms baseline models, including GRU, BiLSTM, and CNN-GRU, in terms of accuracy, precision, and F1-score. The class-wise analysis further indicates that the model demonstrates robustness in distinguishing subtle emotional variations, particularly for emotions with distinct acoustic patterns.

Overall, the study findings prove that utilizing CNN-based spectral feature learning with BiLSTM temporal modeling provides a powerful framework for accurate and reliable SER. Future work will focus on extending this approach to real-world spontaneous speech datasets, addressing class imbalance, and exploring multimodal frameworks that integrate textual and visual cues to further enhance emotion recognition performance.

9. Ethical Consideration

This study used an open-source dataset publicly available on Borealis websites [1] that do not include personally identifiable information. The data were used intended research purposes.

ChatGPT was used exclusively to refine the English language and presentation of the manuscript. The conceptualization, design, experimental analysis and execution of the study were entirely performed by the authors.

References

Omar Abdelfattah, George Gal, Gordon W. Roberts, Ishiang Shih, and Yi-Chi Shih. 2016. A top-down design methodology encompassing component variations due to wide-range operation in frequency synthesizer plls. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(6):2050–2061.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *In Proceedings of the CIC-NLP Workshop on GenAI Detection*.

Md Rayhan Ahmed, Salekul Islam, AKM Muzahidul Islam, and Swakkhar Shatabda. 2023. An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition. *Expert Systems with Applications*, 218:119633.

Waleed Alsabhan. 2023. Human–computer interaction with real-time speech emotion recognition using ensemble techniques. *Sensors*, 23(3):1386.

Bagus Tris Atmaja and Akira Sasou. 2022. Sentiment analysis and emotion recognition from speech using universal speech representations. *Sensors*, 22(17):6369.

[1] Borealis (TESS) Dataset

Jamsher Bhanbhro, Asif Aziz Memon, Bharat Lal, Shahnawaz Talpur, and Madeha Memon. 2025. Speech emotion recognition: Comparative analysis of cnn-lstm and attention-enhanced cnn-lstm models. *Signals*, 6(2):22.

Catarina Botelho, John Mendonça, Anna Pompili, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2024. Macro-descriptors for alzheimer's disease detection using large language models. In *In Proceedings of Interspeech*.

Jing Chen, Chenhui Wang, Kejun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. 2021. Heu emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications*, 33(14):8669–8685.

Thi Le Trinh Dao, Xuan T. Le, and Eric Castelli. 2022. Emotional speech recognition using deep neural networks. *Sensors*, 22(4):1414.

M. Eyasu, W. S. Abebaw, N. Hafeez, F. Uroosa, T. A. Bizuneh, G. Sidorov, and A. Gelbukh. 2025. Tewodros at semeval-2025 task 11: Multilingual emotion intensity detection using small language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1485–1494.

Mrunal Prakash Gavali and Abhishek Verma. 2025. Ensemble of large self-supervised transformers for improving speech emotion recognition. *International Journal of Data Mining, Modelling and Management*, 17(2):217–244.

James J. Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299.

Rashid Jahangir, Ying Wah Teh, Faiqa Hanif, and Ghulam Mujtaba. 2021. Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications*, 80(16):23745–23812.

Eric Jordan, Raphaël Terrisse, Valeria Lucarini, Motasem Alrahabi, Marie-Odile Krebs, Julien Desclés, and Christophe Lemey. 2025. Speech emotion recognition in mental health: Systematic review of voice-based applications. *JMIR Mental Health*, 12(1):e74260.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik. 2021. Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep c-rnn. *International Journal of Speech Technology*, 24(2):303–314.

Harshawardhan S. Kumbhar and Sheetal U. Bhandari. 2019. Speech emotion recognition using mfcc features and lstm network. In *In Proceedings of ICCUBEA*, pages 1–3.

Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and decision making. *Annual Review of Psychology*, 66:799–823.

Yanlin Liu, Aibin Chen, Guoxiong Zhou, Jizheng Yi, Jin Xiang, and Yaru Wang. 2024. Combined cnn-lstm with attention for speech emotion recognition based on feature-level fusion. *Multimedia Tools and Applications*, 83(21):59839–59859.

Cristina Luna-Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M. Montero, and Fernando Fernández-Martínez. 2021. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22):7665.

Syeda Tamanna Alam Monisha and Sadia Sultana. 2022. A review of the advancement in speech emotion recognition for indo-aryan and dravidian languages. *Advances in Human-Computer Interaction*, 2022:9602429.

A. Muppidi and M. Radfar. 2021. Speech emotion recognition using quaternion convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6309–6313.

M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. Toronto emotional speech set (tess).

Y. Lakshmi Prasanna, Y. Tarakaram, Y. Mounika, Suja Palaniswamy, and Susmitha Vekkot. 2022. Comparative deep network analysis of speech emotion recognition models using data augmentation. In *Proceedings of the International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, pages 185–190.

N. Senthilkumar, S. Karpakam, M. G. Devi, R. Balakumaresan, and P. Dhilipkumar. 2022. Speech emotion recognition based on bidirectional lstm architecture and deep belief networks. *Materials Today: Proceedings*, 57:2180–2184.

Youddha Beer Singh and Shivani Goel. 2022. A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492:245–263.

Maria Vaida and Ziyuan Huang. 2025. Multimodal graph neural networks in healthcare: A review of fusion strategies across biomedical domains. *Frontiers in Artificial Intelligence*, 8:1716706.

Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. Speech emotion recognition with dual-sequence lstm architecture. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6474–6478.

World Health Organization. 2025. Over a billion people living with mental health conditions – services require urgent scale-up. Accessed: 2026-03-25.

Cheng Xiefeng, Yue Wang, Shicheng Dai, Pengjun Zhao, and Qifa Liu. 2019. Heart sound signals can be used for emotion recognition. *Scientific Reports*, 9(1):6486.

Wisha Zehra, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu. 2021. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, 7(4):1845–1854.

H. Zhang, H. Huang, and H. Han. 2021. A novel heterogeneous parallel convolution bi-lstm for speech emotion recognition. *Applied Sciences*, 11(21):9897.