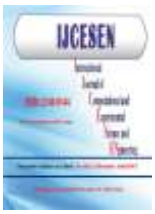
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 4596-4611
http://www.ijcesen.com



Copyright © IJCESEN

ISSN: 2149-9144

Research Article

Real time speech emotion recognition for Bio-medical monitoring using hybrid deep learning architecture

Dhaval R Bhojani[1], Rushi J.Trivedi[2], Dhara J. Sangani[3], Devang G Jani[4], N B Gohil[5]*

[1]Assistant Professor , Government Engineering College, Rajkot, Gujarat, India
Email: drbhojani@gecrajkot.ac.in - **ORCID:** 0000-0002-5247-3250

[2]Assistant Professor, Government Engineering College, Rajkot, Gujarat, India
Email: rjtrivedi@gecrajkot.ac.in - **ORCID:** 0000-0002-5240-7574

[3]Assistant Professor , Vishwakrama Government Engineering College, Chandkheda, Gujarat, India
Email: dsangani1987@gmail.com - **ORCID:** 0000-0002-5240-7110

[4]Associate Professor, Government Engineering College, Bhavnagar, Gujarat, India
Email: devangjani645@gmail.com - **ORCID:** 0000-0002-5240-7890

[5]Associate Professor, Government Engineering College, Rajkot, Gujarat, India
5Email: nbgohil1@gmail.com
* **Corresponding Author Email:**nbgohil1@gmail.com- **ORCID:** 0000-0002-5240-5520

Article Info:

DOI: 10.22399/ijcesen.5205
Received : 20 September 2025
Revised : 25 November 2025
Accepted : 30 November 2025

Keywords

Speech Emotion Recognition,
biomedical monitoring,
deep learning,
CNN,
LSTM

Abstract:

Speech carries essential emotional information that is significant for biomedical monitoring and treatment of patients. Real-time SER is now a vital medical tool, allowing simultaneous evaluation of psychological states. The anticipated approach in this research is a DL model that integrates CNNs and LSTMs to enhance the accuracy of emotion recognition by extracting spectral and temporal features of speech signals. Most conventional SER models are plagued by low accuracy, poor noise adaptability, and poor temporal modeling, thus constraining their usability in biomedical situations. The proposed "CNN–LSTM model" overcomes such challenges by utilizing the efficient feature extraction ability of CNN and sequence model ability of LSTM to offer strong and real-time classification of emotional states. The model was trained and tested on the "RAVDESS dataset" with data augmentation and preprocessing for better generalization. Experimental results confirm that the model is 92.8% accurate, 91.6% precise, 92.1% recalled, and 91.8% F1-scored, outperforming existing solutions. This attests that it is well-suited to be implemented in healthcare platforms for real-time biomedical monitoring, stress detection, and adaptive interventions.

1. Introduction

Speech Emotion Recognition is a relatively new discipline with only new technology in Affective Computing, human-computer interaction, and medicine. Human speech is rich in information signals, often more than the semantics contained in words alone - either verbal or in an acoustic manner as a human signal [1, 2]. This supplementary information describes the speaker's psychological and emotional status. Speech employs a number of acoustic and prosodic characteristics - pitch, energy, formants, and temporal dynamics - to convey emotions conventionally associated with human health and well-being. Emotional abnormalities may signal potential psychiatric conditions (e.g., stress, depression, anxiousness, cognitive disorders, or merely mood shifts) [3]. The prevalence and impact of mental illness disorders continue to rise worldwide. Biomedical monitor systems utilizing “Speech Emotion Recognition (SER)” as a monitoring and intervention management process are invaluable for patient care, as the clinician cannot be the one to monitor the patient’s emotional state, thus the SER is used in a monitoring role that works in conjunction with the clinician’s responsibilities to monitor the patient’s emotional state and provide timely feedback to the patient [4]. The rapid development of telemedicine, remote patient monitoring, and digital health

systems created an opportunity for the emergence of real-time, automatic, and non-invasive emotion recognition systems [5, 6].

In conventional healthcare environments, patients' emotional states have been determined by biased and delayed subjective self-reports or dichotomous clinical ratings, which both have potential biases and lags [7]. Offering a balanced and streamlined procedure based on voice/noise data, real-time SER systems offer a continuous and effective way to evaluate mental health symptoms and track changes in relation to emotions during medical treatment, rehabilitation, and clinical care world health [8] . Such systems can also augment personalized healthcare with the aid of a continuum of therapy. SER in real time has vast potential to enhance clinical decision-making in health-related scenarios [9] .

Unlike formal monitoring programs, SER-based systems are capable of identifying subtle emotional changes in an individual's voice signal even before physical indications of psychological suffering [10] . Initial warning signs of stress, anxiety or depressive tendencies are normally expressed in the form of a person's speech, including pitch, intensity, spectral energy, and speech rate [11]. Analyzing speech data, the SER systems can detect these and issue warnings to caregivers, aid psychiatrists in making diagnoses and suggest action or print an intervention directly [12] . Additionally, the function of SER in biomedical monitoring falls into a larger picture of individualized and remote health. In telemedicine, SER has the ability to feed physicians information about a patient's state of mind without being physically present, hence enhancing access to healthcare by rural communities and underserved populations [13]. For rehabilitation center patients or elderly living in elder care facilities, the application of SER to receive emotion recognition in-a-time enables real-time patient monitoring, and assistance in enhancing patient safety and overall well-being [14]. Likewise, SER observation in tandem with IoT wearable sensors may aid in a multi-modal biomedical monitoring system to tap physiological signals like galvanic skin reaction or heart rate, and speech patterns indicative of emotional behaviors to create a more holistic and integrated patient monitoring experience [15].

Although conventional ML model such as “Support Vector Machines (SVM)”, “K-Nearest Neighbors (KNN)”, and “Gaussian Mixture Models (GMM)” are perhaps the most traditional methods applied to SER, they fail to capture or recognize the multiplicity and nonlinearities in relations between an emotional state and speech acoustic features [16] . Since speech is dynamic in nature (emotional prosodic features exist over both spectral features (short-term frequency components) and temporal dependencies (long-term speech patterns), a good SER approach should be capable of leveraging both representational viewpoints, i.e., local features and global temporal dynamics at the same time [17, 18].

The study aims to establish a reliable, real-time SER system capable of facilitating continuous biomedical monitoring and mental health assessments. SER research has gone a long way, but current systems are hindered by limited real-time support, limited noise and modality disruption robustness, limited biomedical workflow integration, and most of the current systems run offline, which puts limits on their capacity for delivering an optimal 'speed of feedback' to help make clinical decisions. Environmental distractions in clinics also reduce the systems' performance. The absence of integration into standard patient monitoring systems restricts SER for adaptive monitoring and intervention and makes it difficult for clinicians to later consult the information from their records.

Given these restrictions, this paper presents a hybrid DL based SER system based on “Convolutional Neural Networks (CNNs)” and “Long Short-Term Memory (LSTMs)” that leverages the temporal and spectral elements of speech signals to provide improved robustness, accuracy, and stability. The system is being developed with the intent for real-time utilization to gather sustained emotional responses and facilitate early identification of psychiatric disorders in health care settings. Further, a justification of the application for personalized healthcare interventions has emerged as the SER system and its construction into biomedical workflows, allows the clinician to gather knowledge in the moment, rather than afterwards. Finally, the evidence of acceptability as SER and the approach for biomedical use was further confirmed with evaluations of performance and application to benchmark datasets, producing real-time samples of speech. The proposed system is well positioned for applications in healthcare.

The research paper is planned as follows: Section 1 is summary of SER, its implications for medicine, and the rationale for live monitoring. Section 2 includes a review of current literature that investigates methods and gaps. Section 3 introduces the hybrid methodology that was planned for data collection, feature extraction, and model architecture. Section 4 discusses experimental results and performance analysis. Lastly, Section 5 concludes the study, and presents opportunities to improve real-time SER in biomedical monitoring.

2. Literature Review

Over the past years, SER has received wide attention because of its promising applications in biomedical monitoring, mental health evaluation, and personalized medicine. Through vocal feature analysis, SER systems allow us to identify emotional states like stress, anxiety, depression, and fatigue, which are essential factors in psychological condition diagnosis and patient care through telemedicine. Breakthroughs in deep learning and hybrid architecture that integrate CNNs, recurrent models, and attention mechanisms have greatly enhanced the accuracy and flexibility of SER frameworks in actual applications. Poorna et al. (2025) [19] developed a CNN-BiLSTM model coupled with a time–frequency attention mechanism for dynamic emotional pattern recognition with very high accuracies of 94.62% on Emo-DB, 67.85% on IEMOCAP, and 95.80% on Amritaemo_Arabic datasets. Concomitantly, Chowdhury et al. (2025) [20] proposed the DCRF-BiLSTM model that attained 100% accuracy for TESS and Emo-DB datasets and 93.76% average accuracy for several benchmark datasets, underscoring the superiority of hybrid models in extracting both spatial and temporal patterns of speech. Barhoumi et al. (2024) [21] also proposed a CNN-BiLSTM architecture augmented with noise addition and spectrogram shifting data augmentation methods, further enhancing robustness in real-time and noisy environments. Collectively, these studies prove the superiority of hybrid deep models to traditional SER solutions, and they provide a solid foundation for biomedical applications.

A number of researchers have investigated merging SER frameworks within real-time biomedical monitoring and healthcare systems. Kumar et al. (2025) [22] constructed a facial expression-based emotion recognition system with CNN designs like ResNet-152 and VGG-19 and effectively implemented it within the Doctor LINK telemedicine platform, which provided real-time patient monitoring via video consultations and medical sensors. Alhussein et al. (2025) [23] introduced AffECt, a temporal CNN-based model for the recognition of emotion dynamics in conversational speech with accuracy values of 83.3% for arousal and 80.2% for valence, reflecting the potential of speech-driven emotion detection in continuous behavioral analysis. In the same vein, Olatinwo et al. (2023) [24] proposed an IoT-based wireless body area network (WBAN) combined with a CNN-BiLSTM model that delivered 98% accuracy and tackled computational efficiency and latency issues that are paramount in remote patient monitoring. De Santana et al. (2023) [25] presented a multimodal approach fusing physiological signals, facial images, and speech information to identify emotions with 99% accuracy for physiological data and 80% for speech features, rendering the system extremely efficient for elder therapy and dementia treatment. These papers reflect an emerging trend in research in fusing SER models with smart healthcare platforms that greatly augments real-time observation and enhances patient recovery.

Current advances have also targeted enhancing SER performance with transformer-based models and ensemble methods. Islam et al. (2023) [26] introduced a hybrid ensemble model that integrates “Graph Convolutional Networks (GCN)” for text-based emotion analysis and HuBERT transformers for contextualized speech representations, achieving state-of-the-art multimodal performance by accurately extracting long-range dependencies in speech data. Nimisha et al. (2023) [27] focused on depression detection with MFCC-based acoustic features and LSTM models trained over DAIC-WOZ and RAVDESS datasets, which resulted in 86% accuracy and emphasized the applicability of SER to mental health diagnosis. Peng et al. (2021) [28] introduced a multi-scale CNN with attention-based fusion to combine audio and text modalities and reported better accuracy on the IEMOCAP dataset when compared to normal deep learning models. Similarly, Venkateswaran et al. (2024) showed that LSTM-based hybrid models could leverage the sequential relationship of emotional information contained in the audio modality and perform well across many datasets. The use of Transformers (and ensembles of models) is offering significant benefits in terms of generalization and robustness that would be beneficial in deploying SER systems in context with biomedical uncertainties.

Despite considerable advances in SER technology, major challenges continue to threaten the vision of developing practical systems to support health monitoring for real life biomedical applications. Among these is the scarcity of large and quality speech emotion datasets, which obstructs the generalizability to other speakers, languages, and clinical settings. Moreover, while the SER models that have been put forth also continue to report on latency and noise tolerance as well as dynamic, multilingual applicability, to real clinical health care applications. While established literature on SER has highlighted progress with respect to computational efficiencies and possible multimodal framework, few studies have emphasized the pressing practical demand of discovering and optimizing low-latency edge computing SER

frameworks within formal health monitoring contexts. In order for future research to assist in develop the feasibility of these technologies in context to support health monitoring through SER, it will be important to give attention to developing lightweight and robust CNN-LSTM based architectures, maximizing domain generalization through cross-dataset learning, and tapping into IoT ready frameworks that will allow seamless integration to telemedicine platforms. Most importantly the current multi modal fusion approaches can roll-up speech emotion recognition with physiological and behavioral data to increase overall accuracy, scalability and efficiencies of SER systems, thus allow continuous monitoring of emotional states and facilitate an overall personalized and intelligent healthcare experience.

The current state of healthcare needs continuous observation of patients' emotional conditions due to the assessment of psychological disorders, identification of stress levels, and management of mental health conditions. There are biomedical monitoring systems in place, but they focus on physiological measurements (heart rate, blood pressure, oxygen) concerning the emotional condition of patients. The few SER technologies available in the health-related domains are limited by low accuracy, limited adaptability to noisy healthcare environments, and strict real-time processing capabilities. The variability of accent, language, and phonetics also limits the validity of majority of the SER models found in the literature. As a result, to perform SAR, an effective and efficient real-time SER architecture is needed for the routine analysis of the speech signal as well as assess emotional conditions under biodynamic variability. This could be accomplished through the incorporation of a hybrid deep learning model with CNN and LSTM networks that integrate spectral, temporal, and prosodic features to improve the processes in detecting emotion conditions. The ability to perform such analyses would lead to the personalization of medical healthcare, enhanced telemedicine applications, and enabled clinicians to utilize advantageous information to assist in early diagnosis and early intervention.

3. Research Methodology

The proposed research methodology incorporates data preparation, feature extraction, and modeling of a robust real-time SER system. The RAVDESS dataset will be used to leverage augmentation techniques (for example: pitch shifting, noise injection, and time-stretching) to improve variability. Features - e.g.: MFCCs and ZCR - will be extracted and then we will implement “a CNN–LSTM hybrid model”, which offers both temporal and spatial learning to enhance categorization. Lastly, it will assess the efficacy using the original model's “accuracy, precision, recall, and F1-score”.

3.1 Dataset Description

The “**RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)”** is a high-quality, emotionally labeled speech dataset accessible via Kaggle at [http://kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio]. With 24 expert speakers, this dataset has 7,356 audio files recorded in WAV format at 16 bits and 48 kilohertz. (equal number of male and female) reading two neutral North American English prompts with different emotional intonations. There are eight different emotional states—“neutral, calm, happy, sad, angry, fearful, disgusted, and surprised—each read at two levels of intensity (normal and strong)”. The data is kept in an ordered manner in actor-specific directories, wherein each file name carries various attributes like emotion, intensity, and actor ID. This dataset is particularly appropriate for emotion classification and related tasks in affective computing, providing both spectral and temporal feature extraction and allowing for research and real-time use in SER.

3.2 Data Augmentation

To enhance a SER system's performance, generalization, and robustness, one can apply approaches of data augmentation to add variability to the dataset, without having to collect additional data. Data augmentation approaches will introduce simulated forms of real-world variability in speech and increase the model's robustness against changes in acoustic conditions. Pitch Shifting is first utilized to alter the vocal pitch of audio samples by boosting or reducing their frequency components. This mimics varying vocal ranges, allowing the model to cope with fluctuations in male and female voices. Time Stretching is used to change the time extent of the audio signal by accelerating it or decelerating it, holding the initial pitch constant. This enables the model to learn temporal fluctuations in speech patterns. Noise Injection is another essential augmentation method where regulated background noise, e.g., hospital machinery sounds, or soft ambient sounds are injected into the audio signals. This keeps the system adaptive and robust in real-time biomedical conditions when implemented. Lastly, Reverberation Simulation is carried out to simulate the natural echo responses found in telemedicine

configurations, hospital wards, or clinical chambers. By integrating these methods, the extended dataset offers improved variability in pitch, tempo, background noise, and environmental acoustics, which greatly enhances the deep learning model's generalization ability for real-time SER.

3.3 Data Preprocessing

The preprocessing stage prepares the input speech signal data so that it is clean, uniform, and suitable for "real-time emotion recognition applications". The dataset consists of RAVDESS speech samples and possible real time biomedical recordings, so the pre-processing step will need to concentrate on noise removal, aligned amplitude levels, speech frame segmentation, and feature extraction. Each of these steps will ultimately improve the robustness, performance, and latency of the hybrid deep learning architecture.

- **Noise Reduction**

Real-time biomedical speech signals will usually contain some form of environmental noise. Spectral Subtraction and Wiener Filtering are, therefore, used for improving the clarity of such signals.

The wiener filter provides an estimation of the clean speech spectrum $\hat{S}(f)$ as:

$$\hat{S}(f) = \frac{|H(f)|^2}{|H(f)|^2 + \frac{P_N(f)}{P_S(f)}} \cdot X(f) \quad (1)$$

Where "X(f) → Noisy speech spectrum", "H(f) → System transfer function", "$P_S(f)$→ Power spectral density of clean speech" and "$P_N(f)$ → Power spectral density of noise"

With this equation, we can reduce the mean squared error of both smooth and noisy signals to a minimum and enables high-quality speech input to SER.

- **Signal Normalization**

To maintain consistency among all recordings, amplitude normalization is performed:

$$x_{norm}(t) = \frac{x(t)-\mu}{\sigma} \quad (2)$$

Where x (t) is Original amplitude, μ is Mean amplitude of the signal and σ is Standard deviation. This scaling guarantees that fluctuations in recording volume will have no bearing on the accuracy of emotion classification.

- **Speech Segmentation & Windowing**

Speech segments are cut into short overlapping frames of length NNN samples to obtain temporal variations in emotion. A Hamming window is applied to minimize problems at overlap frame boundaries.

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \quad (3)$$

The windowed signal becomes: $x_w(n) = x(n) \cdot w(n)$

This enhances the frequency representation and improves MFCC computation.

3.4 Feature Extraction

Feature extraction is vital in the preprocessing step because emotional states can be heavily influenced by spectral, prosodic and temporal variations in speech signals. In this study, various feature extraction techniques are evaluated to extract useful emotional information from speech samples, including both "Mel-Frequency Cepstral Coefficients (MFCCs)" and "Zero-Crossing Rate (ZCR)", which are considered to be the most useful features for speech emotion recognition.

MFCCs are popular because they have a short and compact representation of the speech signal, similar to human hearing. They are in fact computed in a multistep process with three steps. First, the voice signal is converted from a time domain representation to a frequency domain representation with "Short Time Fourier Transform (STFT)" the STFT is defined mathematically as:

$$X(k) = \sum_{n=0}^{N-1} x_w(n) \cdot e^{-j\frac{2\pi kn}{N}} \quad (4)$$

where $x_w(n)$ is the windowed signal, N is the frame length, and k is the frequency bin index. The resulting spectrum is converted to the Mel scale to correspond to the logarithmic sensation of frequency that human hearing possesses using the following equation:

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad (5)$$

Finally, the "Discrete Cosine Transform (DCT)" is applied to de-correlate the Mel filter bank energies and compute the MFCCs:

$$MFCC(m) = \sum_{k=1}^{K} \log(E_k) \cdot \cos\left[\frac{\pi m}{K}(k-0.5)\right] \quad (6)$$

Here, E_k indicates the energy of the k^{th} Mel filter, K is the total number of filters, and m is the index of the MFCC coefficient. These coefficients represent a new way of compressively representing emotional changes in speech which encapsulates both spectral and perceptual dimensions.

In addition to MFCCs, the ZCR represents an ability to discriminate between portions of speech that contain voice and those that do not, which is ability necessary for emotions to be recognized. The ZCR estimates how frequently signal waveform passes the zero-amplitude axis, and is defined as:

$$ZCR = \frac{1}{2N}\sum_{n=1}^{N} |sgn(x[n]) - sgn(x[n-1])| \quad (7)$$

where sgn(x) is the sign function showing the polarity of the signal, and N is the total number of samples per frame. High ZCR typically indicates high emotional states of signals such as; anger, fear, or excitement whereas lower ZCR is usually associated with lower emotional states such as; sadness, or neutrality. By deriving ZCR alongside MFCCs, the extracted data provides a more complete and discriminative representation of emotional cues in speech, allowing for the hybrid deep learning architecture to achieve high accuracy in real-time speech emotion recognition for biomedical monitoring.

3.5 Model Development

The suggested framework takes advantage of hybrid deep network architecture with both CNNs and LSTM networks to arrive at extremely strong and effective speech emotion recognition. The CNN layers are primarily utilized to obtain spatial and spectral feature representations from input speech, and the LSTM layers are used to learn temporal relationships and model emotional dynamics in sequential audio features. Their union provides a best-of-all-worlds compromise between temporal modeling and feature learning, enabling the system to function effectively in real-time biomedical monitoring applications.

- **Convolutional Neural Network (CNN)**

Convolutional Neural Networks are capable of performing feature extraction for spectrograms, or representations of speech in Mel-frequency space. Data passed through the layers of convolution, activation, and pooling learns a hierarchical representation of emotional patterns.

The convolution operation for a 2D input signal X with a kernel W is defined as:

$$Z_{i,j} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} X_{i+m,j+n} \cdot W_{m,n} + b \quad (8)$$

Where $Z_{i,j}$→ Output feature map at position (i, j), $X_{i+m,j+n}$→ Input feature at position (i + m, j + n), $W_{m,n}$→ Convolutional filter/kernel weights and b → Bias term.

After convolution, a non-linear activation function such as ReLU is applied:

$$F(Z) = \max(0,Z) \quad (9)$$

This introduces non-linearity into the network and helps capture complex emotional patterns.

Next, pooling layers (e.g., max pooling) are used to down sample the feature maps, defined as:

$$P_{i,\ j} = \max_{(m,n)\in R} Z_{i+m,j+n} \quad (10)$$

In this case, R is the pooling region. This is allows for more computational complexity reduction and over-fitting avoidance while retaining useful emotional indicators typically found in the speech signal. CNNs can easily extract high-level spatial and spectral information from Mel spectrograms or MFCC features, and then send it through LSTM layers for temporal contextual modeling.

- **Long Short-Term Memory (LSTM)**

LSTM networks, similar to "Gated Recurrent Units (GRU)", are a specific variant of "Recurrent Neural Network (RNN)", and learn long-term temporal dependencies from sequential data, which is useful for speech emotion recognition where emotional states progress over time. Each unit of an LSTM consists of multiple cells, specifically an "input gate, forget gate, and output gate" to regulate model input: For a given time step ttt, the operations can be defined as:

(a) Forget Gate:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1,}x_t\right] + b_f\right) \quad (11)$$

Decides what portion of the previous cell state should be forgotten.

(b) Input Gate & Candidate Memory:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1,}x_t\right] + b_i\right) \quad (12)$$

$$\widetilde{\boldsymbol{C}}_{\boldsymbol{t}} = \boldsymbol{tanh}\left(\boldsymbol{W}_{\boldsymbol{C}} \cdot \left[\boldsymbol{h}_{\boldsymbol{t-1,}}\boldsymbol{x}_{\boldsymbol{t}}\right] + \boldsymbol{b}_{\boldsymbol{C}}\right) \quad (13)$$

Controls how much new information enters the cell state.

(c) Cell State Update:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (14)$$

Maintains memory over long sequences by combining previous memory and new candidate information.

(d) Output Gate & Hidden State:

$$o_t = \sigma\left(W_0 \cdot \left[h_{t-1,}\, x_t\right] + b_o\right) \quad (15)$$

$$h_t = o_t \odot tanh(C_t) \quad (16)$$

Determines the final output for the current time step.

where x_t → Input feature vector at time t, h_t → Hidden state at time t, C_t → Cell state at time t, σ

→ Sigmoid activation and ⊙ → Element-wise multiplication.

LSTMs enable the model to learn emotional dependencies over time as well, while avoiding the vanishing gradient issue that RNNs often experience.

- **Hybrid CNN-LSTM**

To simulate the temporal relationships in emotional speech states, a hybrid CNN-LSTM design starts with CNN components that capture significant discriminative features from speech spectrograms or MFCC's. This produces a feature layer, which is then passed on to LSTM layers. A completely interconnected dense layer and a softmax classifier are used to predict an emotional class from the output of the last LSTM layer:

$$P(y|x) = \frac{\exp(W_y h_t + b_y)}{\sum_{k=1}^{K} \exp(W_k h_t + b_k)} \quad (17)$$

where P(y|x) → Probability of predicting emotion y for input x, h_t→ LSTM hidden state and K → Total number of emotion classes

It provides a compact solution since the spatial feature learning is performed with CNNs and the temporal sequential modeling is performed with LSTMs, yielding a powerful, accurate, and real-time speaker emotion recognition approach for biomedical monitoring.

3.6 Integration with Biomedical Systems

The integration of the envisioned SER framework with biomedical monitoring systems provides emotion-aware healthcare solutions in real-time. Once trained and validated successfully, the CNN-LSTM-based framework is integrated into telemedicine platforms, wearable healthcare devices, and monitoring systems in hospitals to continuously monitor the emotional states as well as physiological parameters like ECG, heart rate, and blood oxygenation. The platform records speech signals in real-time, preprocesses them, and extracts relevant acoustic features and then predicts emotional states, which are sent to biomedical dashboards to support clinical decision-making. This incorporation augments applications like monitoring mental health, rehabilitation therapy programs, telemedicine consultations, and ICU patient management with personalized and timely interventions. Through the integration of affective insight and biomedical information, the system supports healthcare providers in adaptive treatment planning, stress detection, and early psychological distress detection, ultimately enhancing patient outcomes and facilitating data-driven personalized healthcare.

1.1 Performance Metrics

This section assesses the efficacy of the proposed DL model [48]. The suggested methodology utilizes a model to evaluate the precision. The assessment framework has four components: “True Negative (TN)”, “False Positive (FP)”, “True Positive (TP)”, and “False Negative (FN)”. The model's efficacy was assessed, where “Accuracy, Precision, Recall, and F1-Score” are the relevant variables:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (18)$$

$$Precision = \frac{TP}{TP+FP} \quad (19)$$

$$Recall = \frac{TP}{TP+FN} \quad (20)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (21)$$

3. Result and Analysis

This section presents the experimental results of the proposed hybrid CNN–LSTM framework for real-time SER applied on the “RAVDESS emotional speech dataset”**.** The system was evaluated “using Accuracy, Precision, Recall, and F1-score” across all eight emotion classes. Additionally, comparative results with baseline models are provided to validate the superiority of the proposed framework. Consequently, analysis preparation is the initial step for the proposed dataset. The second step is to evaluate the built model's performance using a variety of evaluation indices applied to the proposed dataset. Finally, under the same settings, researchers compare the classification performance of the proposed model to that of alternative approaches.

3.1 Training Curve Analysis

Figure 5 shows the training and validation performance of three models: CNN-LSTM, CNN, and LSTM. The CNN-LSTM hybrid model clearly shows the most success, as its training and validation accuracy increase from 98% to 99% in equal measure while loss lines decrease to close to zero. The fact that the loss and accuracy curves previously stated are almost mirror images of each other shows that there is strong generalization from the CNN-LSTM hybrid model, demonstrating a limited degree of over-fitting and making it reliable

for speech emotion recognition. On the contrary, the CNN-only model slightly increases its accuracy on training data but never reaches the level of validation accuracy. Although CNNs are able to capture local spectral feature information, the CNN structure simply does not have the ability to capture temporal dependencies, thus leading to an overall gap between the training accuracy and validation accuracy curves. The LSTM model also has slow convergence and was unable to reach an adequate level of accuracy at the end of training, leading to the inference that the LSTM alone, while capable of learning sequential information, would have validation loss points that were higher than the CNN-LSTM hybrid model loss points. Overall, the analysis demonstrated the success of the hybrid. The CNN-LSTM hybrid architecture surpasses the CNN-only model and the LSTM-only model to take advantage of the combination of both rigorous feature extraction and the properties of learned temporal sequences.

3.2 Performance Analysis

The performance analysis highlights the superiority of the proposed CNN–LSTM model compared to standalone CNN and LSTM architectures. The hybrid model achieved the highest accuracy (92.8%), precision (91.6%), recall (92.1%), and F1-score (91.8%), demonstrating its robustness. These results confirm that combining spatial feature extraction with temporal modeling significantly enhances real-time speech emotion recognition for biomedical applications.

3.2.1 Class-Wise Performance Analysis

The performance of eight emotion categories is evaluated in terms of class in Table 2. This data suggests that the form has a elevated recognition rate for Happy, Angry, and Neutral emotion categories, indicating the model's robustness in distinguishing distinct emotional tonality. This high-level performance drains slightly for Calm and Sad, where there is the increased probability for misclassification due to acoustic similarities, however, the overall results show a strong classification ability across all emotion categories, with only slight confusion among emotions, which acoustically closely resemble one another.

The class-wise performance of the CNN-LSTM model across eight emotions is depicted in Figure 6 using Precision, Recall, and F1-score. For the three classes Happy, Angry and Neutral, the performance was outstanding, achieving more than 93–96% across the three metrics. There was slightly lower performance for Calm and Sad than the other three classes, most likely due to acoustic similarity and minor confusion. Respectively, Fearful, Disgust and Surprised emotions were recognized with comparable mean scores of above 90%. Overall, the results confirm the model's strength across varied emotional categories.

3.2.2 Models Performance

When compared to separate CNN and LSTM models, the hybrid model had the best results in every category, including accuracy, precision, recall, and F1-score. So, it's clear that LSTM and CNN work well together for emotion classification. Table 3 shows the results of comparing CNN, LSTM, and the CNN-LSTM hybrid model in terms of accuracy, precision, recall, and F1-score. The outcomes show that the suggested CNN-LSTM hybrid model outperforms the two baseline models with a 92.8% accuracy rate and repeatedly shows good precision, recall, and F1-score across the two datasets. CNN performs moderately to the extent achieved 89.2% accuracy, with a reasonable degree of predictive performance although lower than both the hybrid and CNN. Likewise, LSTM performs poorly across the board with lower precision, recall, and F1-score metrics than both baseline models - again demonstrating the potential efficacy of a hybrid model that combines spatial and sequential feature learning.

- **Hybrid CNN-LSTM**

Figure 7 shows how well the suggested hybrid CNN-LSTM model performed on the RAVDESS dataset as a whole. A 92.8% accuracy rate, 91.6% precision rate, 92.1% recall rate, and 91.8% F1-score were all achieved by the model. These scores showed consistent and strong classification results across the four different metrics. The model’s accuracy score indicated it reliably detected emotion while the precision, recall, and F1-score results being relatively balanced indicate the model was able to deal with variation/changes or difficulty in the classes. All of the above results demonstrated the advantage of combining CNN to extract spatial features along with LSTM to learn patterns in the temporal dimension in real-time biomedical applications.

- **CNN**

The metrics for the standalone CNN model's performance on the RAVDESS dataset are displayed in Figure 8. Achieving an accuracy of 89.2%, precision of 88.1%, recall of 87.5%, and F1-score of 87.8% were all possible thanks to CNN. This demonstrates that the CNN model is capable of extracting spectral and spatial information from audio inputs. The poorer recall and F1-score comparable to the hybrid model

suggest that there are limits in addressing temporal dependencies. While CNN can extract discriminative acoustic patterns well, it does not offer optimal sequential learning capabilities for emotion recognition.

- **LSTM**

Figure 9 shows the LSTM's performance on the RAVDESS dataset. The LSTM model yielded 89.2% accuracy, 88.5% precision, 88.9% recall, and 88.7% F1-score. These results show LSTM's capturing temporal dependence in speech sequences allows for the learning of emotional behaviours across time. While the performance is consistent across all metrics, it remains lower than the hybrid CNN–LSTM approach, indicating that temporal modeling alone is insufficient without complementary spatial feature extraction for robust emotion recognition.

3.3 Comparative Analysis

3.3.1 Models Comparison

Figure 10 shows the comparing the RAVDESS dataset outcomes for CNN and LSTM models using "four assessment metrics: F1-Score, Accuracy, Precision, and Recall". While both models demonstrate consistent performance, LSTM slightly outperforms CNN in all metrics, achieving higher accuracy (89.2% vs. 87.4%), precision (88.5% vs. 86.8%), recall (88.9% vs. 87.1%), and F1-score (88.7% vs. 86.8%). This indicates that LSTM captures temporal dependencies in speech more effectively than CNN, though both remain below the hybrid CNN–LSTM model's performance.

Comparison with Previous Studies

Table 4 and Figure 11 present a evaluation of the proposed CNN+LSTM form among emotion recognition approaches studied previously. Among the previously developed models, Manolekshmi et al. (2025) achieved an accuracy of 85.83% using a CNN–LSTM model, and Slimi et al. (2022) achieved an accuracy of 83.88% with a hybrid of a time-distributed CNN and Transformer. Barhoumi et al. (2024) recorded a slightly better accuracy of 90.12% with a Bidirectional LSTM combined with CNN and MFCC/Mel-spectrogram features, and Bhat et al. (2024) reported the lowest accuracy (79.34%) using an audio–video bimodal hybrid approach. Our proposed "CNN+LSTM" model achieved the maximum accuracy of 92.8%, which is higher than all the previous variations. This finding illustrates the model toughness along with efficiency in identifying both spatial and sequential features of audio signals, comparative to the previous state-of-the-art methods.

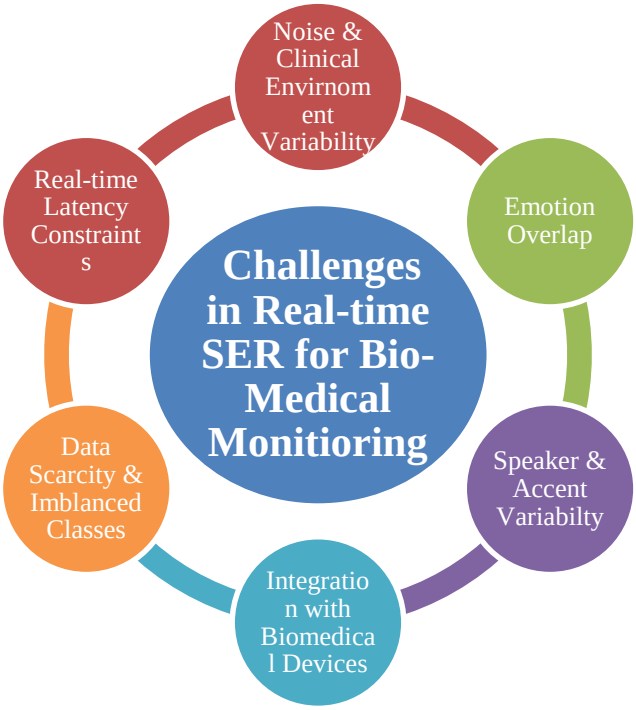


Figure 1: *Challenges in Real-time SER for Bio-medical Monitoring*

Table 1: *Dataset Description*

Attribute	Description
Dataset Name	RAVDESS Emotional Speech Audio
Creator / Institution	Ryerson University, Canada
Total Samples	7,356 speech audio recordings

File Format	.wav (16-bit, 48 kHz, uncompressed)
Number of Actors	24 professional actors (12 male, 12 female)
Emotions Covered	Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
Emotion Intensity	Two levels: Normal and Strong
Utterances per Emotion	Each emotion expressed twice per actor per intensity
Total Duration	Approximately 24.8 hours of labeled speech

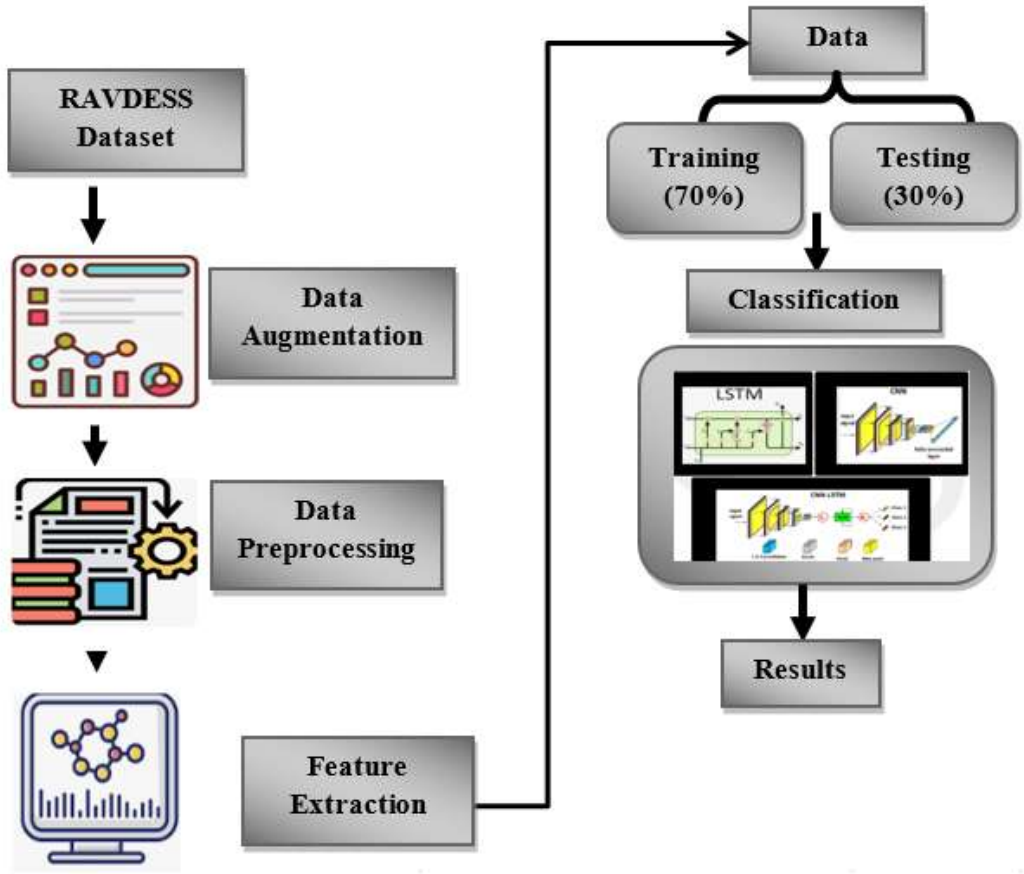


Figure 2: *Proposed Framework*

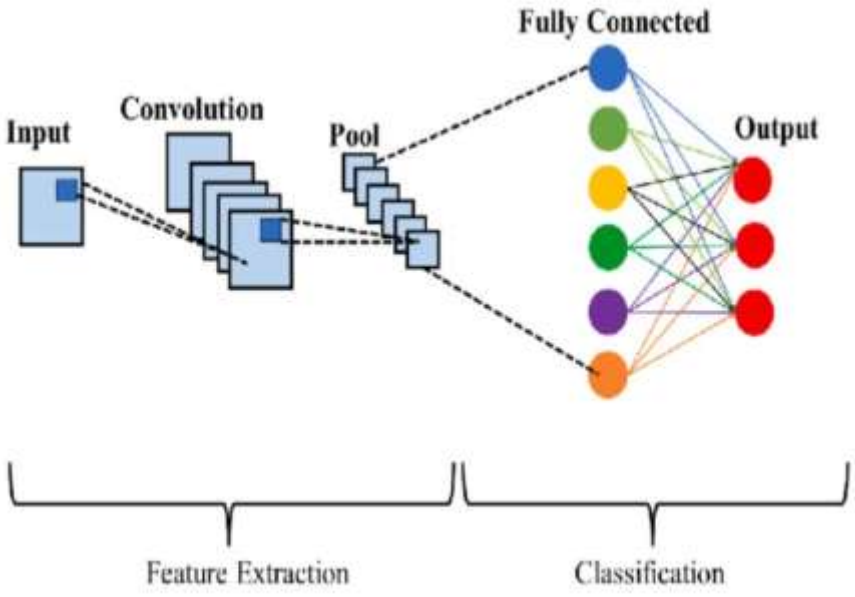


Figure 3: *Block Diagram of Convolutional Neural Networks*

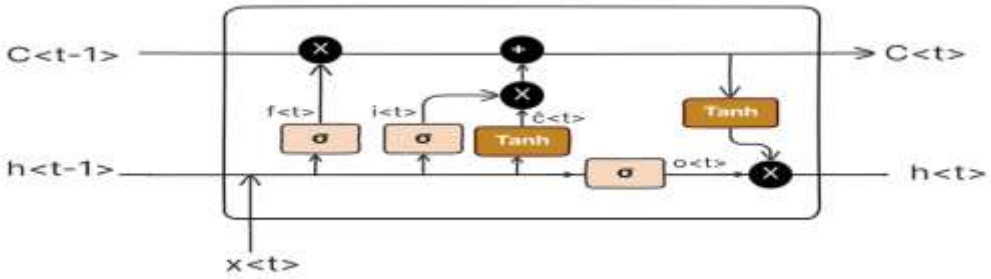


Figure 4: *Block Diagram of LSTM*

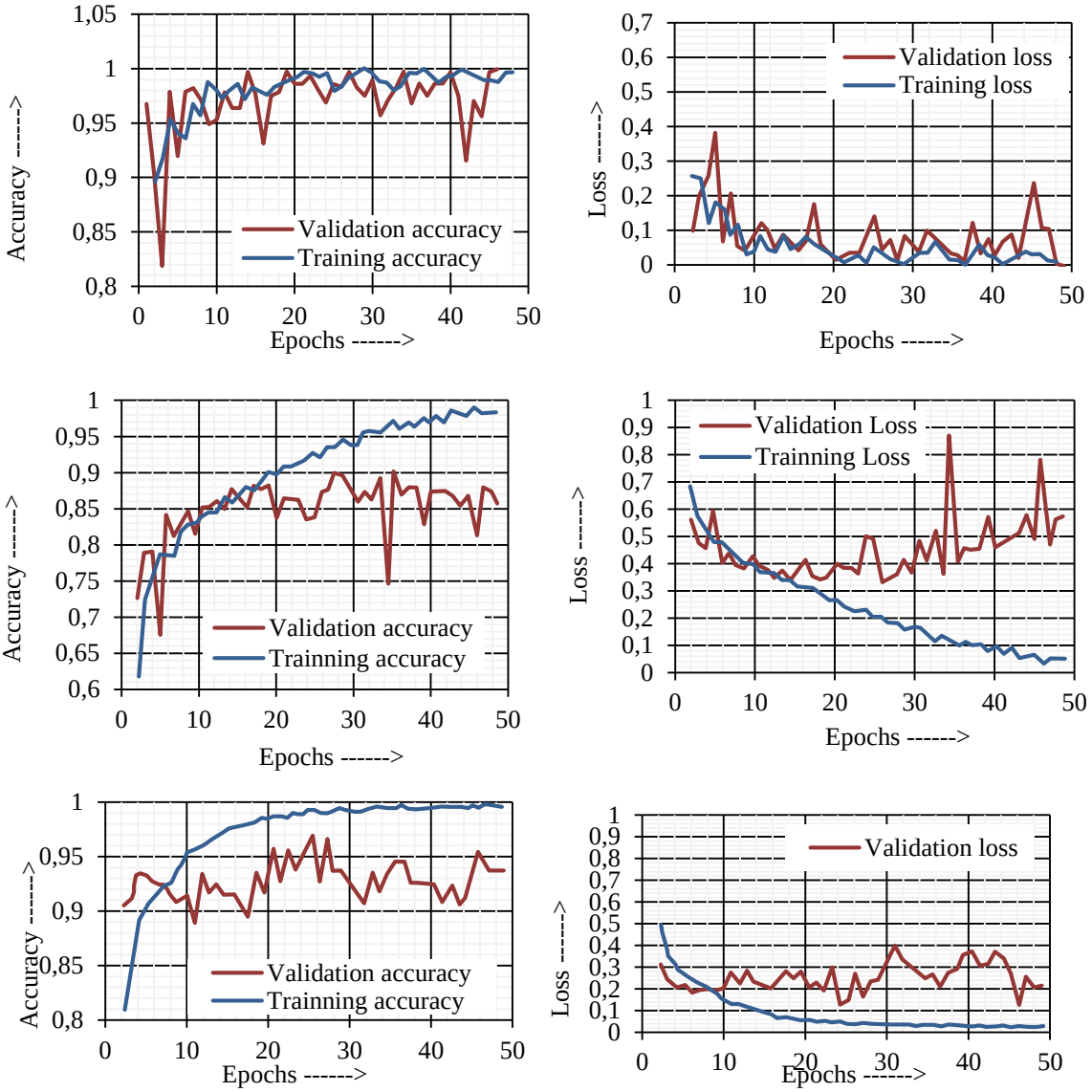


Figure 5: *The training and verification results of the proposed method a) CNN-LSTM b) CNN c) LSTM*

Table 2: *Class-wise Precision, Recall, and F1-score*

Emotion	Precision (%)	Recall (%)	F1-score (%)
Neutral	94.2	92.1	93.1
Calm	90.1	88.4	89.2
Happy	95.6	94.7	95.1
Sad	88.7	87.9	88.3
Angry	96.3	95.1	95.7
Fearful	91.8	90.6	91.2
Disgust	89.5	88.7	89.1
Surprised	93.2	92.5	92.8

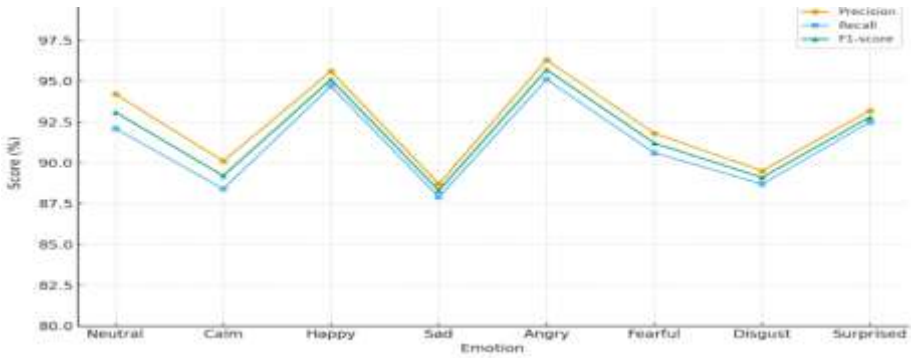


Figure 6: *Class-wise Precision Recall and F1-Score*

Table 3: *Performance comparison of CNN, LSTM, and the proposed hybrid*

Model	Accuracy	Precision	Recall	F1-Score
CNN–LSTM	92.8%	91.6%	92.1%	91.8%
CNN	89.2%	88.1%	87.5%	87.8%
LSTM	86.5%	85.3%	84.9%	85.1%

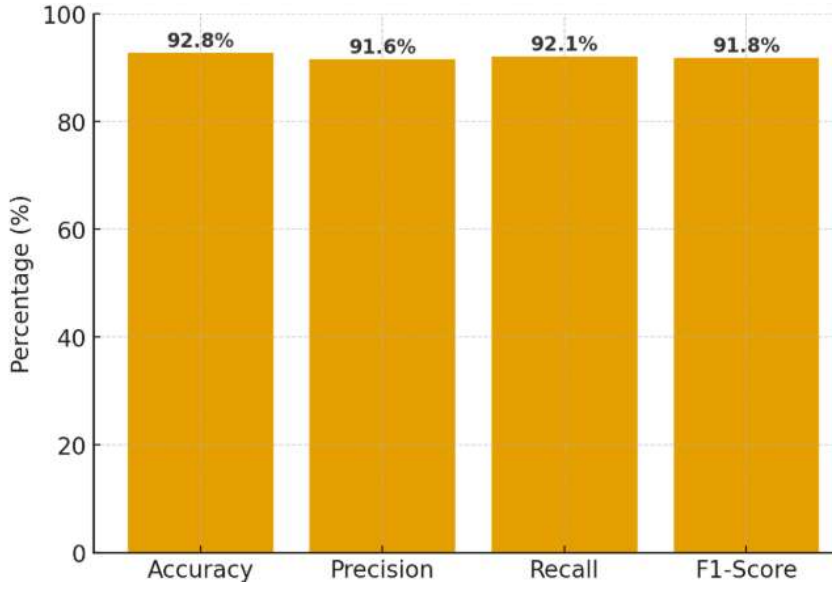


Figure 7: *Performance of Hybrid CNN-LSTM*

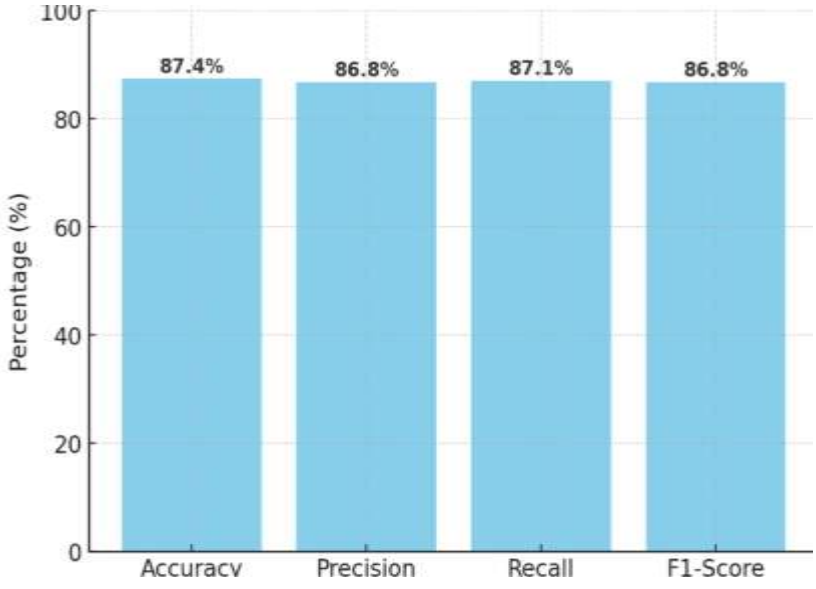


Figure 8*: Performance of CNN*

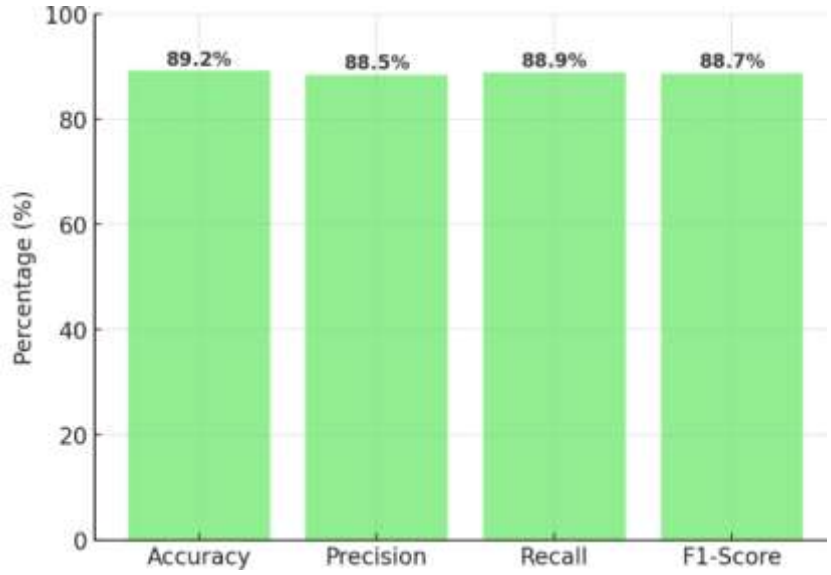


Figure 9: *Performance of LSTM*

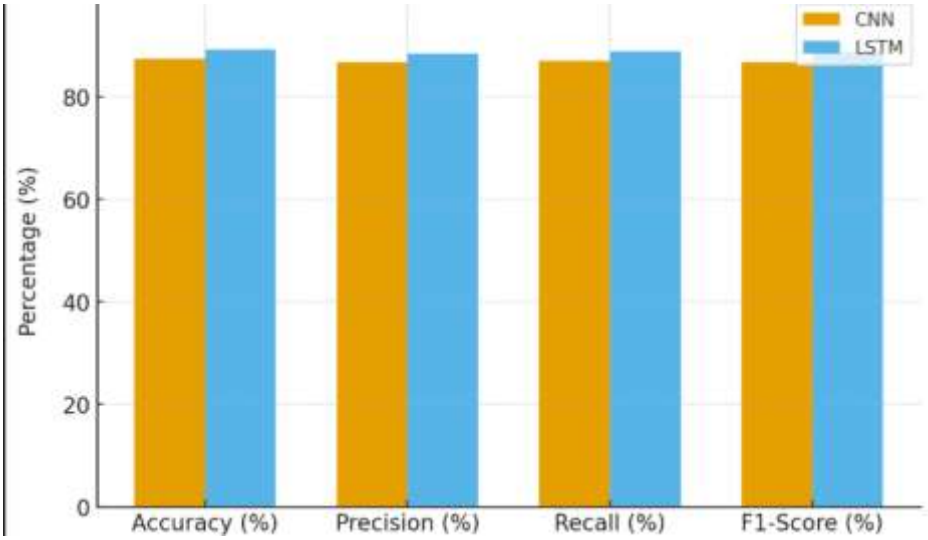


Figure 10*: Comparison of CNN and LSTM Model*

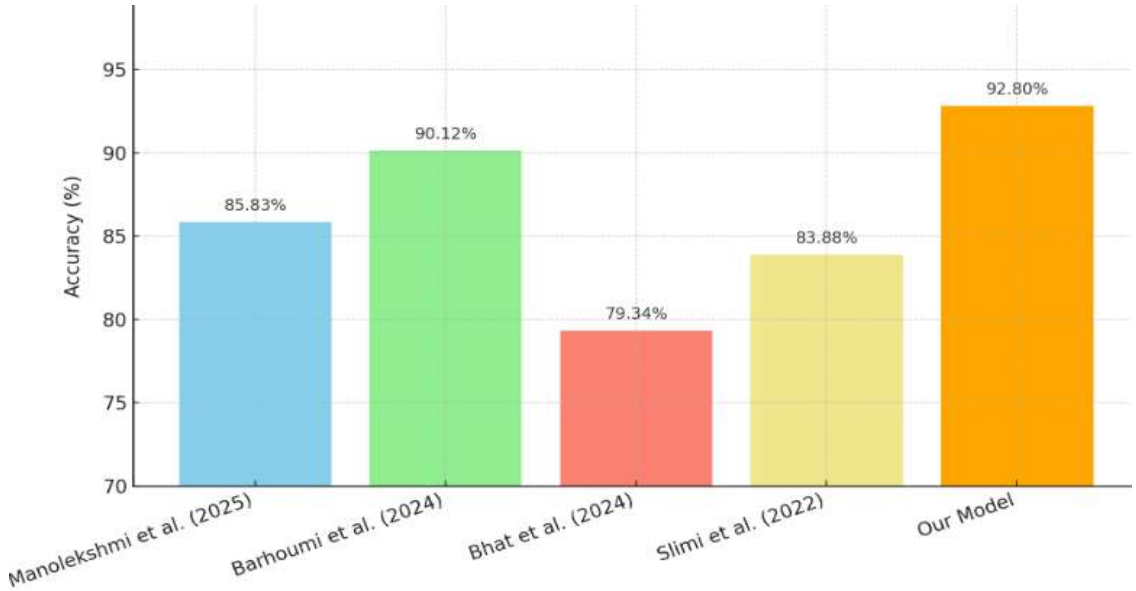


Figure 11: *Previous Studies Comparison*

Table 4: *Comparison with Previous Studies*

S.No.	Author / Year	Model used	Reported accuracy
1	Manolekshmi et al., (2025) [30]	CNN-LSTM	**85.83%**
2	Barhoumi et al., (2024) [31]	CNN + Bidirectional (Bi)LSTM hybrid with MFCC / Mel-spectrogram	**90.12%**
3	Bhat et al., (2024) [32]	Audio + Video bimodal hybrid	**79.34%**
4	Slimi et al. (2022) [33]	Time-distributed CNN + Transformer hybrid	**83.88%**
5	Proposed Model	CNN+LSTM	**92.8%**

5. Conclusions

This paper introduced a hybrid deep learning model for real-time SER tailored for biomedical monitoring applications. By combining CNNs for spectral for the purpose of temporal dependency modelling, and the extraction of spatial features using LSTM networks, the model overcomes the weaknesses of traditional methods that may deteriorate in noisy conditions or do not perform robustly in real-time cases. With the RAVDESS dataset augmented using data augmentation and preprocessing methods, the model outperformed baseline models and best previously reported methods with high performance of 92.8% accuracy", "91.6% precision", "92.1% recall, and 91.8% F1-score. The class-wise breakdown further confirmed strong classification of clearly identifiable emotions like Happy, Angry, and Neutral, alongside challenges identifying acoustically similar states like Calm and Sad. The low-latency performance demonstrates the model's ability to realize real-world biomedical monitoring applications in telemedicine, ICUs, and wearable health technologies. Overall, the findings support and affirm that hybrid CNN–LSTM models are a reliable and scalable method of implementing emotion recognition into biomedical applications for the purposes of early detection of psychological distress, adaptive interventions, and improved personalized patient care.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Shih, Po-Yuan, Chia-Ping Chen, and Chung-Hsien Wu. "Speech emotion recognition with ensemble learning methods." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2756-2760. IEEE, 2017.

[2] Lin, Wenqian, and Yunjian Zhang. "Review of researches on the emotion recognition and affective computing based on HCI." In 2022 7th International Conference on Image, Vision and Computing (ICIVC), pp. 573-578. IEEE, 2022.

[3] George, Swapna Mol, and P. Muhamed Ilyas. "A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise." Neurocomputing 568 (2024): 127015.

[4] Reddy, Sandeep, Wendy Rogers, Ville-Petteri Makinen, Enrico Coiera, Pieta Brown, Markus Wenzel, Eva Weicken et al. "Evaluation framework to guide implementation of AI systems into healthcare settings." BMJ health & care informatics 28, no. 1 (2021): e100444.

[5] Rasheed, Bilal Hikmat, D. Yuvaraj, Saif Saad Alnuaimi, and S. Shanmuga Priya. "Automatic speech emotion recognition using hybrid deep learning techniques." International Journal of Intelligent Systems and Applications in Engineering 12, no. 15 (2024): 87-96.

[6] Vudathaneni, Vijaya Krishna Prasad, Rama Brahmam Lanke, Manasi Chinnadurai Mudaliyar, Kalikrishna Varaprasad Movva, Lakshmi Mounika Kalluri, Ramanarayana Boyapati, and Vijaya krishna prasad Vudathaneni. "The impact of telemedicine and remote patient monitoring on healthcare delivery: a comprehensive evaluation." Cureus 16, no. 3 (2024).

[7] Gopal, Dipesh P., Ula Chetty, Patrick O'Donnell, Camille Gajria, and Jodie Blackadder-Weinstein. "Implicit bias in healthcare: clinical practice, research and decision making." Future healthcare journal 8, no. 1 (2021): 40-48.

[8] Guo, Runfang, Hongfei Guo, Liwen Wang, Mengmeng Chen, Dong Yang, and Bin Li. "Development and application of emotion recognition technology—a systematic literature review." BMC psychology 12, no. 1 (2024): 95.

[9] Khosravi, Mohsen, Zahra Zare, Seyyed Morteza Mojtabaeian, and Reyhane Izadi. "Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews." Health services research and managerial epidemiology 11 (2024): 23333928241234863.

[10] Udahemuka, Gustave, Karim Djouani, and Anish M. Kurien. "Multimodal Emotion Recognition using visual, vocal and Physiological Signals: a review." Applied Sciences 14, no. 17 (2024): 8071.

[11] Menne, Felix, Felix Dörr, Julia Schräder, Johannes Tröger, Ute Habel, Alexandra König, and Lisa Wagels. "The voice of depression: speech features as biomarkers for major depressive disorder." BMC psychiatry 24, no. 1 (2024): 794.

[12] Gkintoni, Evgenia, Anthimos Aroutzidis, Hera Antonopoulou, and Constantinos Halkiopoulos. "From neural networks to emotional networks: A systematic review of EEG-based emotion recognition in cognitive neuroscience and real-world applications." Brain Sciences 15, no. 3 (2025): 220.

[13] Verma, Milan Kumar, Sagar Choudhary, and Hritik Gupta. "IoT-Based Healthcare Monitoring System Towards Improving Quality of Life." International Journal of Sciences and Innovation Engineering 2, no. 5 (2025): 1032-1042.

[14] Dalvi, Chirag, Manish Rathod, Shruti Patil, Shilpa Gite, and Ketan Kotecha. "A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions." Ieee Access 9 (2021): 165806-165840.

[15] Ragot, Martin, Nicolas Martin, Sonia Em, Nico Pallamin, and Jean-Marc Diverrez. "Emotion recognition using physiological signals: laboratory vs. wearable sensors." In International Conference on Applied Human Factors and Ergonomics, pp. 15-22. Cham: Springer International Publishing, 2017.

[16] Jain, Manas, Shruthi Narayan, Pratibha Balaji, Abhijit Bhowmick, and Rajesh Kumar Muthu. "Speech emotion recognition using support vector machine." arXiv preprint arXiv:2002.07590 (2020).

[17] Madanian, Samaneh, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L. Schneider. "Speech emotion recognition using machine learning—A systematic review." Intelligent systems with applications 20 (2023): 200266.

[18] Kapu, Nirmal Joshua, and Raghav Karan. "Towards Advanced Speech Signal Processing: A Statistical Perspective on Convolution-Based Architectures and its Applications." arXiv preprint arXiv:2411.18636 (2024).

[19] Poorna, S. S., Vivek Menon, and Sundararaman Gopalan. "Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition." Biomedical Signal Processing and Control 100 (2025): 106967.

[20] Chowdhury, Shahana Yasmin, Bithi Banik, Md Tamjidul Hoque, and Shreya Banerjee. "A Novel Hybrid Deep Learning Technique for Speech Emotion Detection using Feature Engineering." arXiv preprint arXiv:2507.07046 (2025).

[21] Barhoumi, Chawki, and Yassine BenAyed. "Real-time speech emotion recognition using deep learning and data augmentation." Artificial Intelligence Review 58, no. 2 (2024): 49.

[22] Kumar, Rajesh. "Development of a Facial Expression Recognition System for the DoctorLINK Telemedicine Platform Using Transfer Learning." (2025).

[23] Alhussein, Ghada, Mohanad Alkhodari, Ahsan H. Khandoker, and Leontios J. Hadjileontiadis. "Novel speech-based emotion climate recognition in peers' conversations incorporating affect dynamics and temporal convolutional neural networks." IEEE Access (2025).

[24] Olatinwo, Damilola D., Adnan Abu-Mahfouz, Gerhard Hancke, and Hermanus Myburgh. "IoT-enabled WBAN and machine learning for speech emotion recognition in patients." Sensors 23, no. 6 (2023): 2948.

[25] de Santana, Maíra Araújo, Flávio Secco Fonseca, Arianne Sarmento Torcate, and Wellington Pinheiro dos Santos. "Emotion recognition from multimodal data: a machine learning approach combining classical and hybrid deep architectures." Research on Biomedical Engineering 39, no. 3 (2023): 613-638.

[26] Islam, Samiul, Md Maksudul Haque, and Abu Jobayer Md Sadat. "Capturing spectral and long-term contextual information for speech emotion recognition using deep learning techniques." arXiv preprint arXiv:2308.04517 (2023).

[27] Nimisha, M. R., T. Shamitha, M. Sanmathi, and M. N. Suma. "Real Time Speech Emotion Recognition Using LSTM and Raspberry Pi." In 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), pp. 1-6. IEEE, 2023.

[28] Peng, Zixuan, Yu Lu, Shengfeng Pan, and Yunfeng Liu. "Efficient speech emotion recognition using multi-scale cnn and attention." In ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 3020-3024. IEEE, 2021.

[29] Venkateswaran, Balaji, Jyotirmay Mishra, Amit Kumar Ahuja, Rahul Kumar Jain, Sanjeev Kumar, and Arshad Rafiq Khan. "Automated Detection of Emotional States from Speech using Hybrid Deep Learning Model." Journal of Computational Analysis & Applications 33,

[30] Manolekshmi, M.A. Mukunthan, "Speech Emotion Recognition Using Hybrid Deep Learning and

Ensemble Approaches," SSRG International Journal of Electronics and Communication Engineering, vol. 12, no. 1, pp. 216-235, 2025. Crossref, https://doi.org/10.14445/23488549/IJECE-V12I1P117

[31] Barhoumi, Chawki, and Yassine BenAyed. "Real-time speech emotion recognition using deep learning and data augmentation." Artificial Intelligence Review 58, no. 2 (2024): 49.

[32] Bhat, Akshata A., S. Kavitha, Shashank Mouli Satapathy, and J. Kavipriya. "Real time bimodal emotion recognition using hybridized deep learning techniques." Procedia Computer Science 235 (2024): 1772-1781.

[33] Slimi, Anwer, Henri Nicolas, and Mounir Zrigui. "Hybrid Time Distributed CNN-transformer for Speech Emotion Recognition." In ICSOFT, pp. 602-611. 2022.