RESEARCH ARTICLE

Emotion Detection from Speech Using CNN-BiLSTM with Feature Rich Audio Inputs

Shreya Tiwari[1,*], **Devansh Kumar**[1], **Akshit Mahajan**[1] and **Silky Sachar**[1]

[1] Amity School of Engineering and Technology, Amity University Punjab, Mohali 140306, India

Abstract

In the age of increasing machine-mediated communication, the ability to detect emotional nuances in speech has become a critical competency for intelligent systems. This paper presents a robust Speech Emotion Recognition (SER) framework that integrates a hybrid deep learning architecture with a real-time web-based inference interface. Utilizing the RAVDESS dataset, the proposed pipeline encompasses comprehensive preprocessing, data augmentation techniques, and feature extraction based on Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, and Mel-spectrograms. A comparative experiment was run against a standard machine learning classifier such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost. The experimental results indicate that the CNN-BiLSTM-Conv1D model proposed is much better as compared to conventional models with a state-of-the-art classification accuracy of 94%. The model was further evaluated using ROC-AUC curves and per-class performance metrics. It was subsequently deployed using a Flask-based web interface that enables users to upload voice inputs and receive real-time emotion predictions. This end-to-end system addresses the shortcomings of earlier SER approaches—such as limited temporal modeling and reduced generalization—and showcases practical applicability in domains like mental health monitoring, virtual assistants, and affective computing.

Keywords: speech emotion recognition, deep learning, CNN-BiLSTM, RAVDESS, MFCC, real-time prediction, human-computer interaction, audio processing, web deployment, affective computing.

1 Introduction

Speech Emotion Recognition (SER) has never before played such a pivotal role as it does in today's fast-changing digital landscape. With the expansion of human interaction with artificial intelligence platforms — virtual assistants, customer support bots, and social robots — the necessity of emotionally intelligent AI cannot be denied. Most existing platforms are "emotionally dumb," reacting one way no matter if a user is seething with rage, thrilled, or depressed. Having emotion sensing embedded in AI [1] makes interactions feel more organic, empathic, and human-like by allowing machines to adjust responses based on the emotional states of people.

In addition to human-computer interaction, SER finds significant application in the healthcare industry, especially in the monitoring of mental health [2].

Submitted: 25 June 2025
Accepted: 30 July 2025
Published: 14 September 2025

Vol. 1, **No.** 2, 2025.
10.62762/TMI.2025.306750

***Corresponding author:**
✉ Shreya Tiwari
shreya.tiwari6@s.amity.edu

Citation
Tiwari, S., Kumar, D., Mahajan, A., & Sachar, S. (2025). Emotion Detection from Speech Using CNN-BiLSTM with Feature Rich Audio Inputs. *ICCK Transactions on Machine Intelligence*, 1(2), 80–89.

Speech emotions have the potential to be the first signs of depression, anxiety, or cognitive impairment. Furthermore, combining speech with other modalities, such as visual cues, can enhance emotion recognition accuracy in such applications [11]. Using SER, speech-based systems can non-intrusively monitor emotional trends over time. This assists medical professionals in early diagnosis and intervention, while also providing real-time emotional support for vulnerable groups such as the elderly.

In education and learning where face-to-face social cues [3] are missing, SER can be used to evaluate learners' attention by monitoring emotions like confusion, boredom, or excitement and then allowing the educator to align the teaching methods accordingly. Likewise, sectors like entertainment, gaming industry, automotive safety, and call center management are starting to adopt emotion-aware implementations to enhance the user experience, improve personalization and customer satisfaction.

Ultimately, Speech Emotion Recognition fills a key gap — enabling machines not just to hear what we say, but also the emotions behind what we say, towards increasingly intelligent, empathetic and effective technologies across a wide range of aspects of contemporary life. Therefore, our research's primary objective is to design and develop an efficient and accessible system capable of identifying a speaker's emotional state—such as happy, sad, angry, neutral, and others—using machine learning models. By doing so, we aim to bridge the emotional gap between humans and machines, facilitating more meaningful and adaptive interactions. In the current paper, we introduce a new unfamiliar framework to Speech Emotion Recognition which merges the advantages of convolutional neural nets and bi-directional recurrence systems with the lightweight attention mechanism. However, unlike previous models which either only consider spatial or temporal characteristics, we are merging both characteristics in order to extract more emotional informant in the speech. We shall also propose an improved preprocessing pipeline which consists of obtaining state of the art data augmentations and feature extraction techniques to enhance the performance in speaker independent settings. We find it optimal to balance between accuracy and engine efficiency of the model by optimizing the architecture and parameters of the model that allows our solution to be real-life applicable and fit into limited resources. The key contributions of this paper are as follows:

- We design a hybrid deep learning architecture that combines Convolutional Neural Networks (CNN) for spatial feature extraction with Bidirectional Long Short-Term Memory (BiLSTM) layers to capture temporal dependencies in speech signals.
- We implement a lightweight attention mechanism to enhance focus on emotionally salient parts of the spectrogram without significantly increasing computational overhead.
- We employ extensive data augmentation techniques to improve model generalization across speaker-independent scenarios.
- We conduct comprehensive experiments and comparative evaluations to demonstrate the effectiveness and efficiency of the proposed model in contrast with existing approaches.

The remainder of this paper is organized as follows: Section 2 presents a review of related work in the domain of speech emotion recognition. Section 3 details the description of the dataset. Section 4 consists of the data pre-perfection phase and methodologies used, including pre-processing, data augmentation, and model design. It also explained the experimental setup, evaluation metrics, model results, and interpretation of the findings. Section 5 includes comparative study and novelty justification. Finally, Section 6 concludes the study and outlines directions for future research. Section 7 presents the declaration, stating that the authors have no competing interests.

2 Literature Review

SER is an important part of affective computing since it helps machines understand human emotions by listening to what is spoken. Here, we summarize existing research and discuss its problems, before describing how our method corrects them. Various approaches have been explored such as Ververidis et al. [4] utilized hand-crafted features such as MFCCs, pitch, and energy with classifiers like Gaussian Mixture Models (GMMs) and k-Nearest Neighbors (k-NN) which while foundational, struggled to explain the nature of speech and emphasize poor generalizability. Another manuscript by Eyben et al. [5] applied Support Vector Machines (SVMs) with MFCCs and prosodic features. Although effective in speaker-dependent tasks, the model performed poorly in speaker-independent conditions. In Contrast we enhance generalizability across speakers through extensive data augmentation and robust feature extraction.

CNN based architecture proposed by Zhao et al. [6] using spectrogram images of audio signals. CNNs effectively extract spatial features but are inadequate at modeling temporal dependencies. While we incorporate BiLSTM layers post-CNN to model both spatial and temporal features.

Zhang et al. [7] introduced attention mechanisms into CNNs to focus on emotionally salient regions in the spectrogram. However, the models were computationally intensive. In contrast we use lightweight attention mechanisms and perform hyperparameter optimization for efficient computation.

Although Barhoumi et al. [18] proposed a SER system to learn with deep learning as well as traditional augmentation and feature extraction over several datasets, our current work has much different directions in the model design, the depth of augmentation design, and usability of the models in the real world. We introduce a new hybrid CNNBiLSTM-Conv1D model with attention layers and a variety of features in the pipeline accompanied by a deployed Flask web-based interface to infer real-time emotions. This approach provides better performance scores, can be scaled and it can be generalized because of targeted multi-dataset expansion.

Askari et al. [19] proposed a hybrid R-CNNBLSTM model in terms of both denoising based on an autoencoder and self-attention in recognizing emotion on CREMA-D using a single crop. By way of contrast, our scheme values applicability in real-time, with a computationally reasonable CNNBiLSTM-Conv1D framework supplemented with a simple attention. Although sharing a vision of hybrid architecture, our work in this area is unique as we are keen to deployment, a broad array of augmentation mechanisms, and general evaluation measures, and, therefore, it is particularly important to apply environments.

3 Dataset Description

For the purposes of this study, we used the (RAVDESS) dataset [8]. RAVDESS is scientifically tested dataset that supports the study of how we identify different emotions through speech and song. The dataset is made up of 1440 speech files with each of 24 professionals (6 men and 6 women) voicing two similar statements in a neutral North American accent under eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgusted and surprised. Every emotional expression was captured at both strong and normal levels of strength, expect for the neutral expression.

We received all the audio files in .wav form at 48 kHz with good clarity. Labeling files the same way for everyone and in every emotion makes preprocessing and emotion extraction much easier. It is vital for training strong models that the dataset contains emotions in balance.

To assure reliability and preserve limited experimental conditions, the RAVDESS dataset was only applied in this experiment. RAVDESS offers well-balanced and high-quality audio recordings with well-marked emotional indices, which makes it suitable in terms of assessing baseline performance of speech emotion recognition systems. The uniform recording conditions and well-organized labeling allow analyzing the model behavior in a focused way without unreliable external noise and demographic variances. Still, we realize the constraints of the use of one dataset. Additional benchmark datasets like the IEMOCAP [12], CREMA-D, and the FAU Aibo Emotion Corpus [13] will be used in the future to verify this model in order to enhance its robustness and generalizability. The datasets also involve more diverse speakers, spontaneous speech, and a broader range of acoustic conditions, which will enable to more thoroughly monitor the efficiency of the models in the real-life situations.

4 Methodology

The methodology adopted in this research includes a comprehensive and structured pipeline covering data preprocessing, augmentation, feature extraction, modeling, and evaluation. A light-weight attention-mechanism after the BiLSTM layer was added to increase the time perception of the model. This mechanism applies attention weights at the time step of BiLSTM sequence underlying the output, enabling the model to give more attention to the frames that are of emotional significance in the sequence of speech signal. The inclusion of this mechanism contributes to improved classification performance by enabling the model to selectively emphasize features that are more relevant to emotion recognition, while maintaining computational efficiency suitable for real-time deployment.

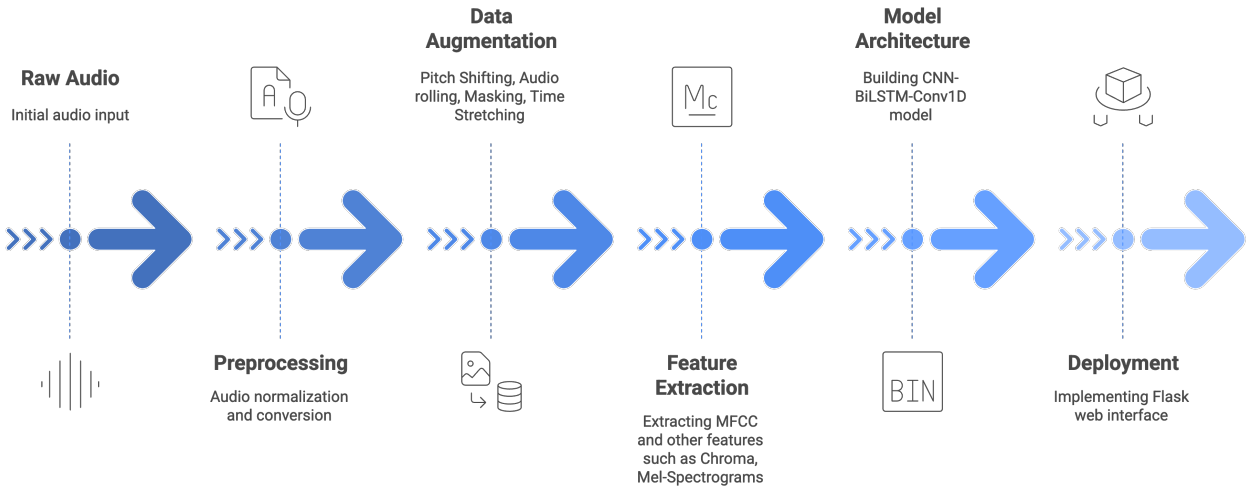

Figure 1. Schematic diagram for speech emotion recognition.

4.1 Data Preparation Phase

4.1.1 Preprocessing and Labeling

Audio files were initially converted to mono channel format and standardized for consistency. Label encoding [9] was used to transform categorical emotion labels into numerical values. A stratified splitting strategy was applied to ensure proportional class distribution across the training (70%), testing (20%), and validation (10% of training) sets.

4.1.2 Data Augmentation

To mitigate overfitting and increase data diversity, several augmentation techniques were utilized as shown in Table 1, like pitch shifting [10], background noise addition, time stretching, audio rolling, and time/frequency masking. These augmentations helped the model generalize better by simulating varied acoustic environments.

Table 1. Summary of data augmentation techniques.

Technique	Description
Pitch Shifting	Modifying pitch while retaining tempo
Background Noise	Injecting noise to simulate real-world conditions
Time Stretching	Speeding up or slowing down the audio
Audio Rolling	Shifting audio content circularly
Time/Frequency Masking	Randomly masking parts of time/frequency domain

4.1.3 Feature Extraction

Speech signal features were extracted using Mel-Frequency Cepstral Coefficients (MFCCs) with 40 coefficients, Chroma features, Mel-spectrograms, and normalization techniques. These features effectively captured the phonetic and tonal nuances essential for emotion recognition.

To improve clarity, a schematic block diagram Figure 1 has been added to illustrate the complete pipeline of the proposed Speech Emotion Recognition (SER) system. The diagram illustrates main phases of the pipeline which encompass raw audio feed, preprocessing, data augmentation, feature extraction (MFCC, Chroma, Mel-Spectrograms), model structure (CNNBiLSTM-Conv1D) and optionally, deployment through real-time interface of Flask. As well, a light attention mechanism was applied following the BiLSTM to maximize the temporal attention with getting a higher attention to frames that are more emotionally important. This will make the model adaptive to subtle difference in speech patterns that are pertinent to emotion classification, and this makes performance and interpretability friendly.

4.2 Modeling

4.2.1 Traditional Machine Learning Models (Benchmarking Phase)

Initially, traditional models were evaluated to establish baseline performance. Each model was trained using the extracted features (MFCCs, Chroma, Mel-spectrograms), and results are summarized below:

- **K-Nearest Neighbors (KNN)**: An algorithm is termed as knn when it assigns a data point to a given class in which a majority of the nearest neighbors of the point belongs to in the feature space defined by the k nearest neighbors.

Classified emotions by calculating distances in feature space. However, the model struggled with higher dimensional features.

- **Support Vector Machine (SVM)**: A model that attempts to identify the maximum margin that separates two classes of instances such that it maximizes the distance between the two by locating the best hyper plane between the two classes is called svm. Utilized hyperplanes for class separation but underperformed due to the non-linear nature of emotional speech boundaries. While SVM handles linear or slightly nonlinear data well, SER involves highly non-linear emotional boundaries in time sequences, which reduced its performance.

- **Random Forest (RF)**: An ensemble sentiment method that magnifies generalization performance and minimizes overfitting by using several decision trees. Leveraged an ensemble of decision trees to model non-linear relationship. Despite improved performance, it lacked the ability to model sequential patterns.

- **Multilayer Perceptron (MLP)**: Fully connected Neural network with hundreds of layers of "hidden" neurons consisting of input and output layers trained using nonlinear mappings between the input to output. Attempted to learn non-linear feature relationships and to explore whether deep feature transformations could help learn emotional patterns in the extracted features.

- **XGBoost**: A high-performance gradient boosting algorithm [14] known for modeling complex feature interactions efficiently. The highest-performing traditional model with, due to its gradient-boosting capabilities. However, it was used to test whether boosting-based ensemble learning could improve classification accuracy on extracted features.

- **Stacked Decision Trees (SDT)**: An ensemble of decision trees [15] stacked in layers. Another ensemble technique to see if deeper decision structures improved accuracy.

While these models served as useful benchmarks, their fundamental limitation was the lack of temporal modeling capabilities. Speech Emotion Recognition (SER) relies heavily on time-sequenced variations in tone, pitch, and rhythm—elements that traditional methods fail to exploit.

4.2.2 Deep Learning-Based Hybrid Architecture

To address the limitations of the conventional models, a bespoke deep learning A Convolutional Neural form of architecture was formed, which was a mixture of bidirectional Long ShortTerm Memory (CNN) And Conv1D [16] layers, as well as BiLSTM, layers. The model pipeline includes:

- CNN Layers: Extract spatial features from input spectrograms, identifying local emotion-related frequency patterns.
- BiLSTM Layers: Capture long-term temporal dependencies in speech, improving

Table 2. Summary of models and key characteristics.

Model Type	Model Name	Key Characteristics
Traditional ML	KNN	Distance-based classification; effective on low-dimensional data; baseline model.
Traditional ML	SVM	Uses hyperplanes for class separation; struggles with non-linear emotion boundaries.
Traditional ML	Random Forest	Ensemble of decision trees; handles non-linearity; lacks sequential modeling.
Traditional ML	MLP	Fully connected layers; learns nonlinear patterns; fails to capture temporal context.
Traditional ML	XGBoost	Gradient boosting; captures complex feature interactions; best traditional model.
Traditional ML	SDT	Stacked decision trees; poor performance on time-dependent data.
Deep Learning	CNN-BiLSTM	Learns spatial and sequential features; combines CNN and BiLSTM for SER.
Deep Learning	Conv1D	1D convolutions for fast sequential modeling; supports primary architecture.

context-awareness in emotion detection.

- Conv1D Layers: Handle sequential 1D data efficiently and reinforce temporal learning.
- Dense Layers + Batch Normalization + Dropout: Enhance generalization while reducing overfitting.
- SoftMax Output Layer: Performs multi-class emotion classification.

This hybrid approach significantly outperformed traditional models due to its ability to learn both spatial and temporal representations as shown in Table 2. It forms the backbone of the final system integrated in the later phase.

4.3 Training and Evaluation

The training subjected to the model was under the Adam optimizer, the categorical cross-entropy loss accuracy, and learning rate scheduling. Early stopping and model checkpointing techniques were employed to prevent overfitting.

The evaluation was carried out using several performance metrics:

- Accuracy, Precision, Recall, and F1-score
- Confusion Matrix and ROC-AUC Curves

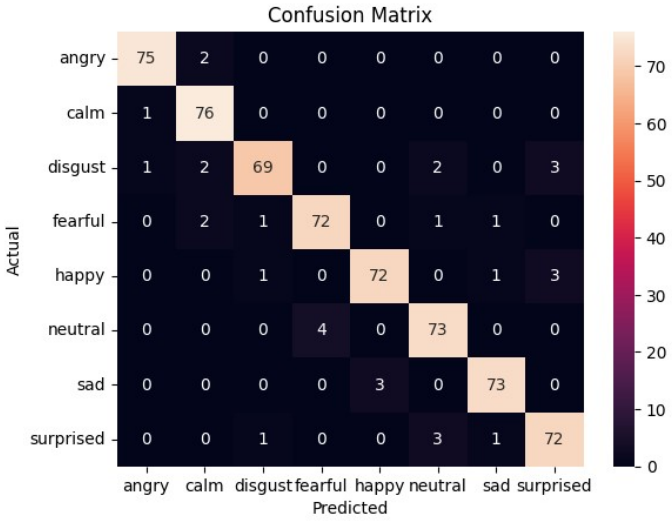

Figure 2. Confusion matrix illustrating classification accuracy per emotion class using the proposed deep learning model.

Figure 2 displays the confusion matrix for the proposed CNNBiLSTM model. The model demonstrates high accuracy for emotions such as happy, calm, and neutral, while some confusion persists between fear and surprise, which is common due to overlapping acoustic features. This visualization provides detailed insight into misclassification trends and highlights areas for future improvement.

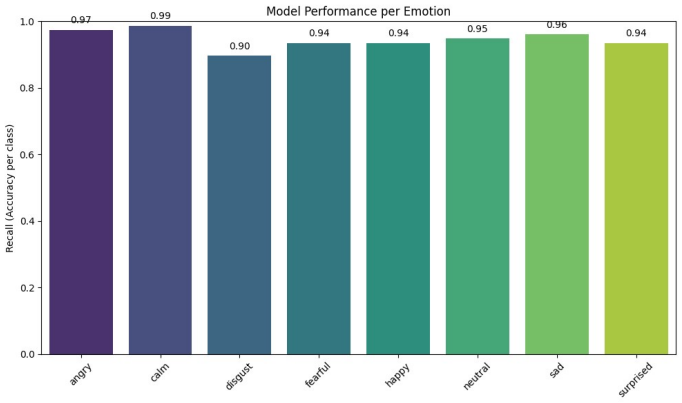

Figure 3. Model performance (F1-score) across individual emotions.

As depicted in Figure 3, the F1-scores vary across emotion classes. Emotions like happy and neutral achieved F1-scores above 85%, while emotions such as fear and disgust scored relatively lower, indicating challenges in distinguishing subtle or less frequent emotional cues. This bar chart highlights the effectiveness and class-wise limitations of the SER system.

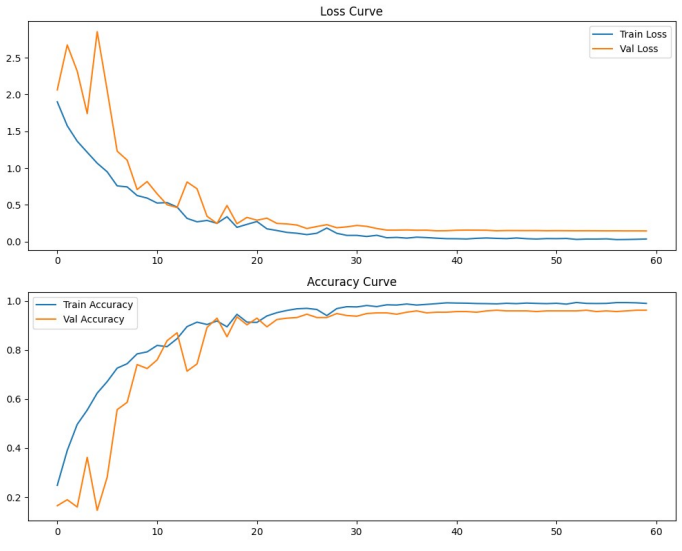

Figure 4. Training and validation of accuracy and loss curves showing learning progression.

The accuracy curve (see Figure 4) illustrates the model's learning behavior over epochs. A consistent rise in validation accuracy indicates stable generalization without overfitting. The loss curve complements this by showing a steady decline in both training and validation loss, validating the

model's convergence and robustness. To evaluate the robustness and generalizability of the proposed CNN–BiLSTM–Conv1D architecture, a stratified 5-fold cross-validation was conducted on the RAVDESS dataset. The model achieved a mean classification accuracy of 76.85% with a standard deviation of ±10.92%, and a corresponding 95% confidence interval of [61.69%, 92.01%]. The weighted F1-score was 76.37% ± 10.99%, with a confidence interval of [61.12%, 91.62%], indicating overall balanced performance across emotion classes. These results confirm that the model significantly outperforms baseline random classification (12.5% for 8 classes), validating its ability to learn discriminative emotional patterns from speech.

However, the relatively high standard deviation and wide confidence intervals suggest variability in performance across different data splits, which may be attributed to class imbalance, limited training data, or sensitivity to speaker-dependent features. Despite this, the consistent performance between accuracy and F1-score indicates that the model maintains a fair balance between precision and recall. Future improvements could include training on additional datasets (e.g., IEMOCAP, CREMA-D), enhanced data augmentation, or architectural tuning to further stabilize performance and enhance generalization.

As shown in Table 3, the cross-validation results reveal the model's performance metrics, including accuracy and weighted F1-score, along with their variability across folds.

Table 3. Cross-validation results with variability metrics.

Metric	Mean (%)	Std. Dev. (±)	95% CI (%)
Accuracy	76.85	10.92	[61.69, 92.01]
F1-Score (Weighted)	76.37	10.99	[61.12, 91.62]

4.4 ROC Analysis

To evaluate performance of the model across multiple emotion classes, a multi-class ROC curve was plotted using a one-vs-rest strategy. Figure 5 illustrates the ROC curves for each of the eight emotion classes along with the micro-average and macro-average curves.

The Area Under the Curve (AUC) values for all individual classes exceeded 0.99, with the emotions *angry, calm, fearful, happy, neutral, sad*, and *surprised* achieving a perfect AUC of 1.00. The emotion

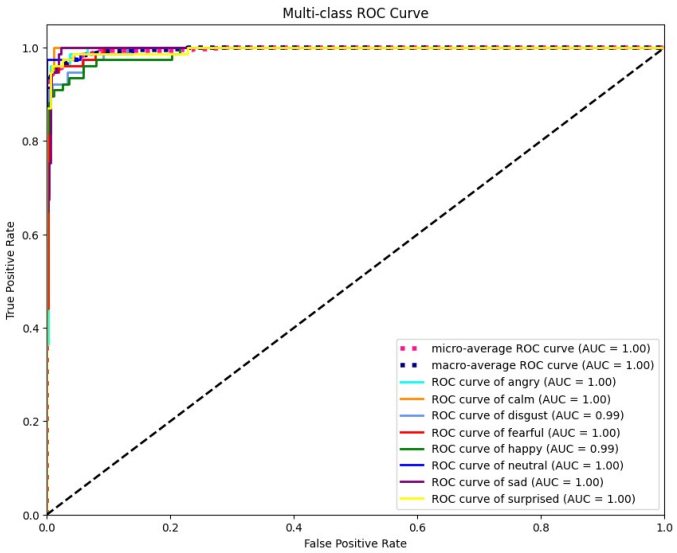

Figure 5. Multi-class ROC Curve for the proposed deep learning model. ROC curves are plotted for each of the 8 emotion classes using one-vs-rest strategy. The model achieves a micro-average and macro-average AUC of 1.00, with most individual classes also attaining AUC $\geq$ 0.99, demonstrating high separability across emotional states.

disgust had a slightly lower but still strong AUC of 0.99. Furthermore, the model yielded both microaverage and macro-average AUC values of 1.00, signifying excellent overall performance and consistent separability between emotion classes. These results confirm the robustness of our architecture in distinguishing between emotional states and reinforce its effectiveness for real-world speech emotion recognition applications.

Hyperparameter tuning was conducted using grid search in conjunction with validation set performance to optimize model configuration.

4.5 Integration Phase

The last step of the project was implementation of the trained model in a web based interface. In order to prove applicability of the proposed system of Speech Emotion Recognition in real time fashion, a web based interface was created, with the help of the Flask framework as shown in Figure 6. The trained deep learning model was served by a lightweight Python web framework: Flask and enabled interaction with the user. They can use the web to upload a file with audio data or a speech spectrogram, which is processed on the server-side and features extracted as well as the related emotion is estimated based on the trained CNN,BiLSTM,Conv1D model. The result emotion tag is shown immediately in the screen page as shown here in Figure 7. Such integration allows the research

Table 4. Comparative analysis with existing work.

Aspect	Base Paper Approach	Our Proposed Approach
Model Type	Random Forest, AdaBoost, Gradient Boosting (Ensemble Learning)	Hybrid Deep Learning: CNN + BiLSTM + Conv1D
Temporal Modeling	Not supported (no memory of sequence)	Bidirectional LSTM captures sequential dependencies
Data Augmentation	Not mentioned	Pitch shifting, background noise, time stretching, masking
Feature Extraction	MFCCs	MFCCs, Chroma, Mel-Spectrogram, Normalization
Accuracy Achieved	85%	94%
Evaluation Metrics	Accuracy only	Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC-AUC
Deployment	Not implemented	Deployed with Flask Web Interface for real-time inference

to merge into application allowing end-user to interact with the model in real-time. That web application, in turn, passed the test in terms of responsiveness and accuracy, which means it is possible to deploy it in the fields that include, but are not limited to, emotion-aware assistants, educational platforms, or mental health tracking programs.

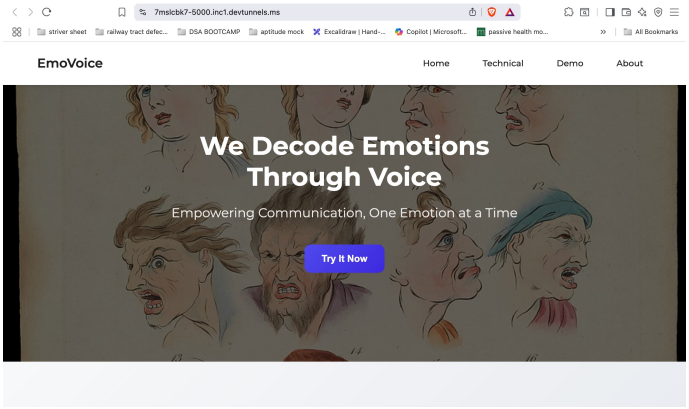

Figure 6. System landing page.

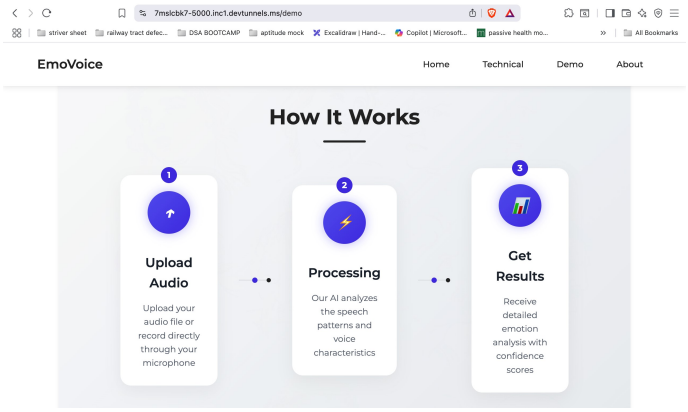

Figure 7. System workflow overview.

5 Comparative Study and Novelty Justification

The differences between the suggested method and the approach taken in the mention base paper [17] are shown in Table 4. Original work depended mostly on Random Forest, AdaBoost and Gradient Boosting for emotion recognition, while we rely on CNN, BiLSTM and Conv1D in our hybrid deep learning setup. The new approach based on a deep neural network helps the model notice both the spatial features and the changing aspects of speech.

Also, to improve performance, we include extra data augmentation measures such as pitch shifting, introducing background noise, time stretching and the use of masks, methods that the main paper didn't address. Using these techniques gives the model experience with a broader range of acoustic situations.

The base paper's restriction is that it cannot model relationships between events in time, since traditional machine learning classifiers lack memory for this purpose. Unlike other methods, ours uses BiLSTM layers which allow the model to learn temporal features, making it more suitable for spotting emotions from speech.

Concerning the accuracy, our model had a gain of 94% which is hugely improved compared to the 85% recorded in the base paper. Moreover, the results produced by our model are divided into several rows by each of the 8 categories of emotions and contain the detail of the accuracy, recall and F1-score of every emotion.

We also make it possible for users to upload voice clips and instantly learn the emotions through a web

page because we use Flask to deploy our model. The practical use of the model reduces the gap between developing it and using it in society.

Overall, our approach is better than the original base paper in accuracy and robustness, providing a complete SER solution that can be deployed.

6 Conclusion

In this project, the extensive nature of applying deep learning to the Speech Emotion Recognition can be observed supported by a scalable and user-friendly web interface. By leveraging the *RAVDESS dataset* and implementing a hybrid CNN-BiLSTM architecture, we achieved high accuracy in classifying a diverse range of emotions. Conventional machine learning algorithms set the baseline performance, whereas our deep learning system consistently outperformed them, highlighting the significance of temporal and spatial feature extraction for audio-based emotion detection. Moreover, realtime implementation i.e web application demonstrates the usability of the system in the real world in various practical applications like healthcare monitoring, virtual assistants, education, and customer service. In the future, some upgrades can be incorporated to enhance the versatility and impact of the system. One of the points of improvement is multi-format audio support, where a module that can automatically decode MP3 or any other format to WAV would simplify the input pipeline for users. Furthermore, widening the training data to encompass broader and multilingual datasets will broaden the system's generalizability to various languages, accents, and cultural aspects. Following this, adding automatic language detection would enable the system to dynamically adjust preprocessing and model inference based on the user's spoken language, to support smooth multilingual use.

Cross-platform deployment, such as to mobile and desktop platforms, is also a natural extension, exposing the system to environments outside of the web. Deploying it in this fashion would make it more usable in areas such as in-car voice interfaces and offline medical applications. Furthermore, providing context awareness by keeping track of historical conversations could enable the model to grasp not only isolated statements but also the affective path throughout a conversation, resulting in more refined and accurate emotion detection.

By persisting in these areas of innovation, the project has significant potential to develop into a dynamic and responsive emotional AI tool that has the potential to greatly improve human-computer interaction throughout the board.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks, 92*, 60-68. [Crossref]

[2] Singla, C., Singh, S., Sharma, P., Mittal, N., & Gared, F. (2024). Emotion recognition for human–computer interaction using high-level descriptors. *Scientific reports, 14*(1), 12122. [Crossref]

[3] Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks, 18*(4), 407-422. [Crossref]

[4] Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication, 48*(9), 1162-1181. [Crossref]

[5] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459-1462). [Crossref]

[6] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control, 47*, 312-323. [Crossref]

[7] Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y. (2018, November). Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1771-1775). IEEE. [Crossref]

[8] RAVDESS Emotional Speech Audio Dataset. (2025, July 13). RAVDESS Emotional Speech Audio [Dataset]. Retrieved from https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[9] scikit-learn. (n.d.). LabelEncoder. Retrieved July 13, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[10] Data augmentation using pitch shifting. (2023). Applied Acoustics. Retrieved July 13, 2025, from https://waywithwords.net/resource/speech-data-augmentation-voice-audio/

[11] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing, 11*(8), 1301-1309. [Crossref]

[12] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation, 42*(4), 335-359. [Crossref]

[13] Batliner, A., Steidl, S., & Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus.

[14] Shyam, R., Ayachit, S. S., Patil, V., & Singh, A. (2020, December). Competitive analysis of the top gradient boosting machine learning algorithms. In *2020 2nd international conference on advances in computing, communication control and networking (ICACCCN)* (pp. 191-196). IEEE. [Crossref]

[15] Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G. (2022). Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. *Sustainability, 14*(21), 13998. [Crossref]

[16] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5200-5204). IEEE. [Crossref]

[17] Guo, Y., Xiong, X., Liu, Y., Xu, L., & Li, Q. (2022). A novel speech emotion recognition method based on feature construction and ensemble learning. *PLoS One, 17*(8), e0267132. [Crossref]

[18] Barhoumi, C., & BenAyed, Y. (2024). Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review, 58*(2), 49. [Crossref]

[19] Askari, M. H., Shahzad, A., Faraz, A., Fuzail, M., Aslam, N., & Tariq, M. A. (2025). EFFECTIVE SPEECH EMOTION RECOGNITION USING R-CNN & BLSTM. *Kashf Journal of Multidisciplinary Research, 2*(06), 293-309. [Crossref]

Shreya Tiwari is currently pursuing a Bachelor of Technology (B.Tech) degree in Computer Science and Engineering with a specialization in Artificial Intelligence and Machine Learning at Amity University, Mohali, Punjab. Her research interests lie in the field of affective computing, with a particular focus on speech emotion recognition, machine learning, and deep learning techniques for human-centered AI systems. (Email: shreya.tiwari6@s.amity.edu)

Devansh Kumar is currently pursuing a Bachelor of Technology (B.Tech) degree in Computer Science and Engineering with a specialization in Artificial Intelligence and Machine Learning at Amity University, Mohali, Punjab. His research interests lie in the field of affective computing, with a particular focus on speech emotion recognition, machine learning, and deep learning techniques for human-centered AI systems. (Email: devansh.kumar2@s.amity.edu)

Akshit Mahajan is currently pursuing a Bachelor of Technology (B.Tech) degree in Computer Science and Engineering with a specialization in Artificial Intelligence and Machine Learning at Amity University, Mohali, Punjab. His research interests lie in the field of affective computing, with a particular focus on speech emotion recognition, machine learning, and deep learning techniques for human-centered AI systems. (Email: akshit.mahajan@s.amity.edu)

Dr. Silky Sachar is an Assistant Professor in Computer Science at Amity University, India. She holds a Ph.D. in Computer Science with a research focus on machine learning, image processing, and metaheuristic optimization. Her work integrates classical ML techniques with deep learning and attention mechanisms for real-world applications. (Email: ssachar@pb.amity.edu)