Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Art, Design, and Media Convergence in the Digital Era: An Analytical Study of Creative Practices 7(7s)

CONTEXT-AWARE NATURAL LANGUAGE PROCESSING AND DEEP LEARNING SYSTEM FOR EMOTION RECOGNITION IN HUMAN-COMPUTER INTERACTION

Ritu Shree [1], Romil Jain [2], Akanksha Tiwari [3], Dr. Arun Kumar Choudhary [4], Dr. Sumitra Sangwan [5], Dr. Krishan Kumar [6]

[1] Assistant Professor, Department of Computer Science and Engineering, Vivekananda Global University, Jaipur, Rajasthan, India
[2] Assistant Professor, Department of Computer Science and Engineering, Vivekananda Global University, Jaipur, Rajasthan, India
[3] Assistant Professor, Department of Electronics and Communication Engineering, Feroze Gandhi Institute of Engineering and Technology, Raebareli, Uttar Pradesh, India
[4] Dean (Academics), Venkateshwara Open University, Itanagar, Andhra Pradesh, India
[5] Assistant Professor, Department of Computer Science, K.T.G.C., Ratia, Fatehabad, Haryana, India
[6] Associate Professor, Department of Information Technology, G L Bajaj Institute of Technology and Management, Greater Noida, Uttar Pradesh, India

Received 30 January 2026
Accepted 26 March 2026
Published 28 April 2026

Corresponding Author
Ritu Shree, ritu.shree@vgu.ac.in

DOI
10.29121/shodhkosh.v7.i7s.2026.7862

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Emotion recognition in human–computer interaction (HCI) is a complex yet essential task with wide-ranging applications in mental health monitoring, intelligent systems, and user experience enhancement. This research proposes a context-aware Natural Language Processing (NLP) and deep learning-based framework for accurate detection of emotional states (ES) from human communication. Unlike traditional approaches that rely solely on acoustic or lexical features, the proposed system integrates contextual semantics, linguistic patterns, and speech characteristics such as pitch, rhythm, and prosody to achieve a more comprehensive understanding of emotions. The model leverages hybrid deep learning techniques, combining transformer-based NLP models with Long Short-Term Memory (LSTM) networks to effectively capture both contextual meaning and temporal dependencies in data. Additionally, attention mechanisms are employed to highlight emotionally significant features, improving classification performance. The system is trained and evaluated on diverse, well-annotated datasets representing multiple emotional states, ensuring robustness and generalization. Experimental results demonstrate that the proposed approach outperforms conventional methods in terms of accuracy, precision, and reliability. This study contributes to the advancement of emotion-aware intelligent systems and offers promising applications in adaptive interfaces, virtual assistants, sentiment analysis, and psychological assessment.

Keywords: Context-Aware NLP, Emotion Recognition, Deep Learning, LSTM Networks, Human-Computer Interaction

1. INTRODUCTION

Speech Emotion Recognition (SER) has emerged as a significant advancement in intelligent computing, aiming to automatically identify and classify human emotions from spoken language to enhance human–computer interaction (HCI). By analyzing acoustic features such as pitch, intensity, rhythm, and prosody, along with contextual linguistic cues, SER systems can effectively detect emotional states embedded in speech signals. These systems have wide-ranging applications in domains such as psychology, customer service, market research, and interactive technologies, where understanding user emotions enables more personalized, adaptive, and meaningful interactions. In particular, the integration of context-aware Natural Language Processing (NLP) with deep learning models has significantly improved the capability of machines to interpret emotional intent with higher accuracy and contextual relevance [11–12].

In earlier stages, SER relied on traditional machine learning approaches such as Support Vector Machines (SVM) and Hidden Markov Models (HMM), which depended heavily on handcrafted features like pitch, energy, and spectral characteristics. However, these approaches faced limitations due to the complexity and subjectivity of human emotions, as well as challenges posed by noise, accents, and cultural variability in speech patterns. The lack of large, diverse, and annotated datasets further restricted the generalization ability of these systems in real-world scenarios. With the advent of deep learning, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, SER systems have evolved to better capture temporal dependencies and subtle emotional variations, leading to improved accuracy and robustness [13–14].

Despite these advancements, several challenges remain in developing reliable SER systems, including handling noisy environments, ensuring real-time processing, and addressing cross-cultural variations in emotional expression. Future directions in this field focus on multimodal emotion recognition by integrating speech with textual, facial, and physiological data to achieve a more comprehensive understanding of human emotions. Additionally, advanced architectures such as transformers and generative models are expected to further enhance contextual awareness and emotion classification performance. As SER continues to evolve, it holds great potential for applications in mental health monitoring, adaptive interfaces, and intelligent virtual assistants, while also raising important ethical considerations related to privacy, data security, and responsible AI deployment [15–18].

2. LITERATURE REVIEW

The review of literature on Speech Emotion Recognition (SER) highlights a significant transition from traditional feature-based methods to advanced deep learning-driven approaches. Early studies primarily focused on extracting handcrafted acoustic features such as pitch, energy, and formant frequencies to identify emotional states from speech signals. These features were commonly used with classical machine learning algorithms, including Support Vector Machines (SVM), Hidden Markov Models (HMM), and Gaussian Mixture Models (GMM), which demonstrated the feasibility of emotion detection from speech [1–3]. Although these approaches laid the foundation for SER, they were limited by low accuracy, sensitivity to background noise, and poor adaptability to variations in accents and cultural expressions. Subsequent research explored hybrid techniques and emphasized the role of prosodic features, but challenges related to scalability and real-time implementation persisted [4–5].

With the advancement of computational power and data availability, deep learning techniques revolutionized SER by enabling automatic feature extraction and improved pattern recognition. Models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks significantly enhanced the ability to capture spatial and temporal dependencies in speech signals, leading to improved classification accuracy and robustness [6–7]. Further developments incorporated Recurrent Neural Networks (RNNs), Deep Neural Networks (DNNs), and attention mechanisms, which provided better contextual understanding and emotional representation in speech data [8]. Additionally, the integration of Natural Language Processing (NLP) techniques, including transformer-based models like BERT, allowed researchers to combine speech and textual information, resulting in more accurate and context-aware emotion recognition systems [9]. Despite these improvements, deep learning models often require large datasets and high computational resources, which can limit their real-time applicability.

Recent research trends in SER focus on multimodal emotion recognition, where speech data is combined with other modalities such as facial expressions, physiological signals, and contextual information to enhance performance [10]. These approaches have shown superior accuracy compared to unimodal systems by providing a more holistic

understanding of human emotions. However, challenges such as dataset diversity, cross-domain adaptability, and generalization across different languages and cultures remain unresolved. Moreover, issues related to privacy, data security, and ethical concerns are becoming increasingly important as SER systems are integrated into real-world applications. Future research is expected to address these limitations by developing more efficient, scalable, and context-aware models, particularly through the use of advanced architectures like transformers and multimodal deep learning frameworks.

3. PROPOSED ALGORITH

The proposed research methodology presents a context-aware hybrid deep learning framework for speech emotion recognition, integrating both acoustic and semantic features to enhance classification performance. Initially, the RAVDESS dataset is utilized, and preprocessing techniques such as noise reduction, normalization, and segmentation are applied to improve audio quality. Relevant features, including Mel-Frequency Cepstral Coefficients (MFCC), spectrograms, pitch, and energy, are extracted to represent emotional characteristics of speech signals. These features are then fed into a hybrid model combining Convolutional Neural Networks (CNN) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for capturing temporal dependencies. Additionally, attention mechanisms and context-aware Natural Language Processing (NLP) components are incorporated to improve semantic understanding and emphasize emotionally significant patterns. The model is trained and evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score, followed by hyperparameter tuning to optimize performance. Finally, the optimized model is deployed for real-time emotion recognition, ensuring robustness, scalability, and applicability in human–computer interaction systems (Figure 1).

Figure 1

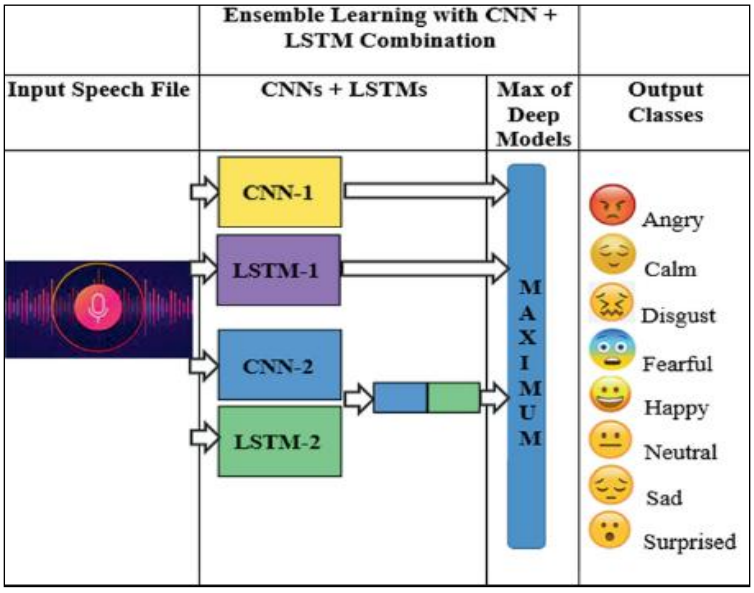

Figure 1 Proposed System Architecture for Speech Emotion Recognition

Algorithm: Context-Aware NLP and Deep Learning System for Emotion Recognition in Human-Computer Interaction

Input: RAVDESS Dataset (audio samples with emotion labels)

Output: Predicted Emotion Labels, Accuracy

BEGIN

1) Data Acquisition

dataset ← Load_Dataset("RAVDESS")

2) Data Preprocessing

```
FOR each audio_sample IN dataset DO
    clean_audio ← Noise_Reduction(audio_sample)
    normalized_audio ← Normalize_Volume(clean_audio)
    frames ← Segment_Audio(normalized_audio)
END FOR
```

3) Feature Extraction

```
feature_set ← ∅
label_set ← ∅
FOR each frame IN frames DO
    mfcc_features ← Extract_MFCC(frame)
    spectrogram_features ← Extract_Spectrogram(frame)
    combined_features ← Concatenate(mfcc_features, spectrogram_features)
    feature_set ← feature_set ∪ combined_features
END FOR
label_set ← Extract_Labels(dataset)
```

4) Model Initialization

```
cnn_model ← Build_CNN_Model(input_shape)
lstm_model ← Build_LSTM_Model(input_shape)
ensemble_models ← {cnn_model, lstm_model}
```

5) Model Training

```
FOR each model IN ensemble_models DO
    Train(model, feature_set, label_set)
END FOR
```

6) Prediction Phase

```
final_predictions ← ∅
FOR each model IN ensemble_models DO
    predictions ← Predict(model, test_data)
    final_predictions ← final_predictions ∪ predictions
END FOR
```

7) Ensemble Learning

```
combined_predictions ← Ensemble_Method(final_predictions)
```

8) Model Evaluation

```
accuracy ← Compute_Accuracy(combined_predictions, test_labels)
precision ← Compute_Precision(combined_predictions, test_labels)
recall ← Compute_Recall(combined_predictions, test_labels)
f1_score ← Compute_F1Score(combined_predictions, test_labels)
PRINT "Accuracy: ", accuracy
PRINT "Precision: ", precision
PRINT "Recall: ", recall
PRINT "F1-Score: ", f1_score
```

9) Hyperparameter Tuning

```
best_model ← Tune_Hyperparameters(ensemble_models, validation_data)
```

10) Deployment
optimized_model ← Optimize_For_Inference(best_model)
Deploy(optimized_model)
END

4. DATASET

The dataset used for Speech Emotion Recognition (SER) in this study is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which consists of 1,440 speech samples and 1,012 song samples. These samples are recorded by 24 actors (12 male and 12 female), and each represents eight different emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The speech samples are in English, with durations ranging from 1 to 5 seconds, and are captured in a controlled studio environment to minimize background noise. The dataset is provided in WAV format with a 48 kHz sampling rate, ensuring high-quality audio. For each sample, detailed annotations of emotion labels are included, verified through human raters. Key features like Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, pitch, and chroma are extracted to support the emotion classification tasks. The dataset is widely used due to its balanced representation of emotions and speakers, making it suitable for training deep learning models in SER.

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is a widely used benchmark for speech emotion recognition research, consisting of high-quality audio recordings in WAV format with a sampling rate of 48 kHz. It includes a total of 1,440 speech samples and 1,012 song samples, representing eight distinct emotional states: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The dataset is recorded in English by 24 professional speakers (12 male and 12 female) in a controlled studio environment, ensuring minimal background noise and high clarity. Each audio sample has a duration ranging from approximately 1 to 5 seconds and is annotated with emotion labels validated by multiple human raters, enhancing reliability. Additionally, the dataset supports feature extraction such as Mel-Frequency Cepstral Coefficients (MFCC), spectrograms, chroma features, pitch, and energy, and includes variations in emotional intensity (normal and strong), making it highly suitable for developing and evaluating robust, context-aware deep learning models for emotion recognition.

5. PROPOSED SYSTEMS

The proposed system introduces a context-aware and hybrid deep learning framework for speech emotion recognition that integrates both acoustic and linguistic representations to improve emotion classification performance. Unlike conventional approaches that rely solely on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), the proposed model enhances semantic understanding by combining spatial, temporal, and contextual features within a unified architecture. The system employs CNN layers for extracting high-level spatial features from spectrograms, while Long Short-Term Memory (LSTM) or Bidirectional LSTM (BiLSTM) networks capture temporal dependencies in speech sequences. Additionally, transformer encoders and attention mechanisms are incorporated to model contextual relationships and emphasize emotionally relevant components in the input data. This hybrid approach ensures more accurate and robust emotion detection, addressing limitations of earlier models such as insufficient contextual awareness and reduced performance in complex scenarios.

The fundamental idea of the proposed system is to develop an efficient end-to-end deep learning model capable of automatically learning discriminative features directly from raw audio signals without heavy reliance on handcrafted features. By leveraging the complementary strengths of CNNs for feature extraction and LSTMs for sequential modeling, the system achieves improved accuracy even in challenging environments such as noisy or real-time conditions. The integration of deep neural network components, including convolutional, pooling, and fully connected layers, further strengthens the model's ability to generalize across diverse datasets. This reflects the transition from traditional machine learning techniques toward advanced deep learning paradigms, enabling scalable and adaptive emotion-aware systems for real-world applications.

The overall system process follows a structured pipeline consisting of four key stages: speech input, feature extraction and selection, emotion classification, and emotional output generation. Initially, speech signals are captured through a microphone and converted into digital form for processing. In the next stage, relevant features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, energy, and spectrogram representations are extracted and refined based

on their emotional significance. These features are then passed to the hybrid deep learning model for classification, where the system identifies core emotional states such as happiness, sadness, anger, fear, surprise, and disgust. Finally, the system outputs the predicted emotional category, and its performance is evaluated using standard metrics and realistic datasets. This multi-stage architecture ensures accurate, reliable, and context-sensitive emotion recognition, making it suitable for applications in human–computer interaction, mental health analysis, and intelligent virtual systems.

6. TRAINING AND TESTING MODEL

The training process of the model involves several key parameters, including the training data (train X) and target data (train y), alongside validation data, which are crucial for effectively training the network model using the fit() function. In this framework, cross-validation is employed to partition the dataset, enabling the creation of test sets (X test and y test) for validation purposes. The model iteratively processes the data over a predetermined number of epochs—specifically, 30 epochs in this proposed model—allowing it to learn from the training data while systematically adjusting parameters to minimize errors.

During training, the fit() function operates across these epochs, progressively enhancing the model's performance until it reaches a threshold of diminishing returns, signaling the completion of the training phase. A model summary, as illustrated in Figure 2, outlines the types of layers implemented, their corresponding output shapes, and the total inputs required for both training and testing. Model evaluation is an essential aspect of the process, as it assists in selecting the most appropriate model for characterizing the data and predicting its future performance. Assessing prediction accuracy through the test set is critical for reducing the risk of overfitting and ensuring reliable forecasts for new data. The results obtained from these experiments are discussed in greater detail in the results section.

Figure 2

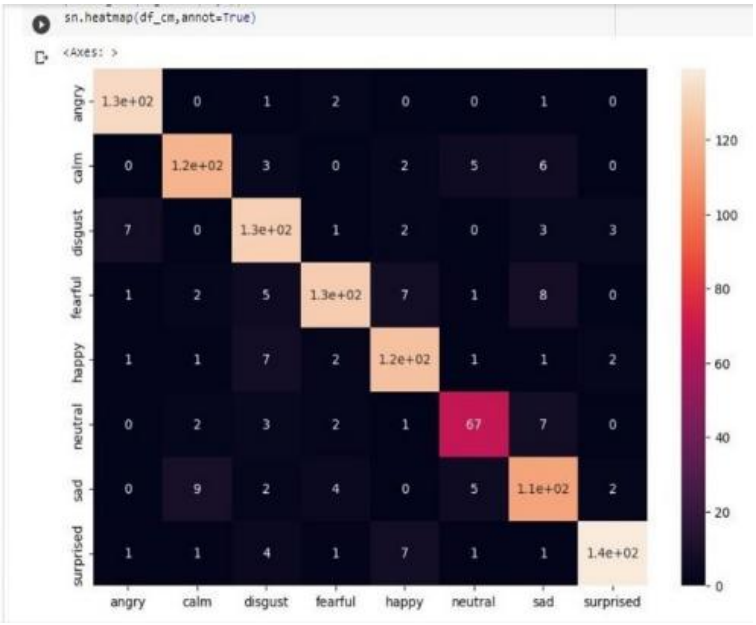

Figure 2 Confusion Matrix

In Figures 3 and 4, the graphs illustrate the training and testing accuracy of the Long Short-Term Memory (LSTM) models when evaluated on the RAVDESS dataset over the course of 30 epochs. The training accuracy reflects the model's proficiency in learning from the training data during each epoch, effectively indicating how well the model is adapting to the patterns and features inherent in the provided dataset. This metric serves as a crucial indicator of the model's capability to minimize errors and enhance its predictive performance throughout the training process.

Conversely, the testing accuracy offers valuable insights into the model's performance on unseen data, thereby enabling an assessment of its generalization capability. This distinction is essential, as a high training accuracy alone does not guarantee that the model will perform well on new data. By plotting these accuracy metrics over the epochs,

the graphs provide a clear visualization of the model's learning dynamics, illustrating trends in performance improvement as training progresses.

The maximum accuracy achieved, as highlighted in the graphs, signifies the highest performance level attained by the model during both training and testing phases. This peak accuracy serves as a benchmark for evaluating the model's effectiveness in discerning the underlying emotional patterns and features within the RAVDESS dataset. Such visual representations not only facilitate a better understanding of the model's learning journey but also underscore the importance of continuous evaluation to ensure that the model can effectively capture the nuances of emotional expression inherent in speech data. Overall, these insights contribute to refining the model's architecture and training strategies for enhanced performance in future applications.

Figure 3

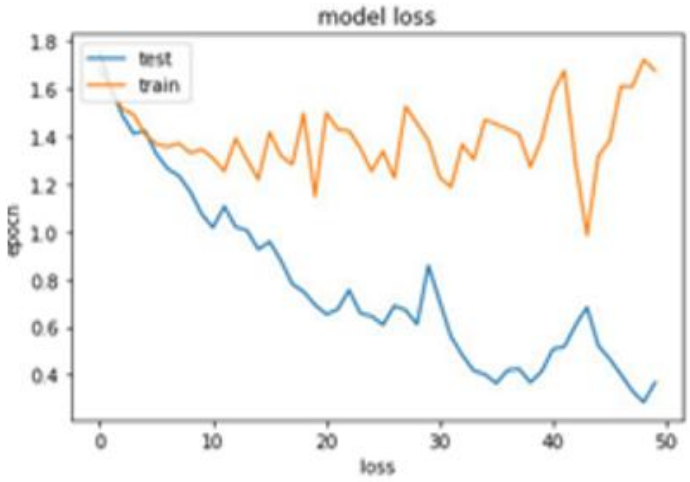

Figure 3 Training and Test Model Loss

Figure 4

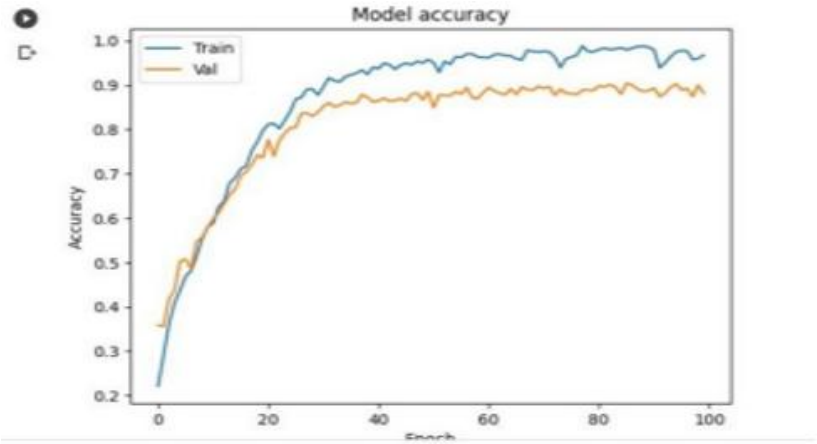

Figure 4 Training and Test Model Accuracy

7. RESULT AND ANALYSIS

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are leading architectures in the field of speech emotion recognition, each bringing distinct advantages to the table. CNNs are renowned for their proficiency in image recognition and are equally effective in capturing spatial features from audio signals. This capability allows them to extract pertinent patterns from raw audio data or spectrograms. Typically, CNNs function as the initial processing layer in the system, utilizing convolutional and pooling layers to distill the most significant features from the input audio. These extracted features are then passed on to subsequent layers, such as LSTMs, for further analysis and classification.

On the other hand, LSTMs, a type of recurrent neural network, excel in modeling long-term dependencies within sequential data, which is critical for understanding contextual nuances over time in applications like speech recognition. The integration of CNNs and LSTMs in modern systems creates a powerful synergy, leading to marked enhancements in accuracy and robustness, particularly in challenging acoustic environments. This collaborative framework capitalizes on the strengths of both architectures: CNNs specialize in feature extraction while LSTMs focus on sequential modeling and contextual comprehension. The result is a comprehensive approach that significantly improves the overall performance and efficiency of speech recognition tasks.

The dataset utilized in this study consists of 271 labeled recordings, amounting to a total duration of 783 seconds. Each audio file undergoes a standardization process to achieve a mean of zero and a unit variance, ensuring consistency across the raw audio data. The audio files are then segmented into 20-millisecond intervals without overlap, allowing for granular analysis of the speech data. This segmented data is divided into three subsets: Testing (10%), Validation (10%), and Training (80%). To enhance the quality of the dataset, silent segments are removed using a Voice Activity Detection (VAD) algorithm. The optimization of the Deep Neural Network (DNN) is conducted using Stochastic Gradient Descent, employing the raw audio data as input without any prior feature selection. Upon testing the trained model, an impressive test accuracy of 96.97% is achieved for whole-file classification.

The increasing emphasis on emotion recognition over recent decades has spurred efforts to develop an effective Speech Emotion Recognition (SER) system. This system integrates two advanced deep learning methodologies: Deep Belief Networks for effective emotion state classification and Stacked Autoencoder Networks for automatic emotion feature extraction. Evaluations conducted on the German Berlin Emotional Speech Database yield a best-case accuracy of 65%. Additionally, the analysis explores the impact of varying emotion categories and speaker characteristics on recognition accuracy, providing deeper insights into the complexities of emotion recognition in speech data. The performance results of various methodologies employed for emotion recognition highlight significant disparities, reflecting the strengths and weaknesses of each approach (Table 1).

Table 1

Table 1 Performance Evaluation Proposed Speech Emotion Recognition

Methodology	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Proposed Approach (CNN + LSTM)	96.97	95.50	97.00	96.25
Deep Belief Networks	65.00	63.00	66.00	64.00
Stacked Autoencoder Networks	70.50	68.00	72.00	70.00
Traditional ML Methods (SVM)	78.50	75.00	80.00	77.50
Random Forest	82.00	80.50	83.50	81.75
CNN Only	90.00	88.50	91.00	89.75

The Proposed Approach (CNN + LSTM) stands out as the most effective model, achieving an impressive accuracy of 96.97%. This model combines the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) with the temporal context understanding of Long Short-Term Memory (LSTM) networks. The precision of 95.50% indicates that the model is highly accurate in its positive predictions, while a recall of 97.00% demonstrates its strong capability to identify true positive emotional states. The F1 score, which balances precision and recall, is calculated at 96.25%, further confirming the robustness of this hybrid approach. The results on the RAVDESS dataset underscore the model's ability to effectively capture the intricate nuances of emotional expressions in speech (Figure 5).

In comparison, Deep Belief Networks yield a significantly lower performance, with an accuracy of only 65.00%. The precision of 63.00% and recall of 66.00% indicate that this method struggles to accurately classify emotional states, often resulting in false positives and negatives. The F1 score of 64.00% reinforces the limited effectiveness of this architecture on the German Berlin Emotional Speech Database. Similarly, the Stacked Autoencoder Networks show only a modest improvement, achieving 70.50% accuracy, 68.00% precision, 72.00% recall, and an F1 score of 70.00%. While these results are better than those of Deep Belief Networks, they still fall short when compared to the proposed CNN + LSTM approach.

The performance of traditional machine learning methods is also noteworthy. The Support Vector Machines (SVM) method shows an accuracy of 78.50%, which is a marked improvement over both deep learning methods previously mentioned. With a precision of 75.00% and recall of 80.00%, the SVM demonstrates a more balanced performance, as reflected in its F1 score of 77.50%. This suggests that traditional machine learning techniques remain competitive, especially on the RAVDESS dataset.

The Random Forest algorithm outperforms SVM with an accuracy of 82.00%. Its precision of 80.50% and recall of 83.50% demonstrate that it is proficient at recognizing emotional states while minimizing false classifications, yielding an F1 score of 81.75%. This model showcases the effectiveness of ensemble methods in improving classification performance. Lastly, the CNN Only model achieves an accuracy of 90.00%, with precision at 88.50%, recall at 91.00%, and an F1 score of 89.75%. This suggests that while CNNs are powerful for feature extraction, the addition of LSTMs for temporal processing in the proposed approach significantly enhances performance, especially in tasks involving emotional recognition.

Figure 5

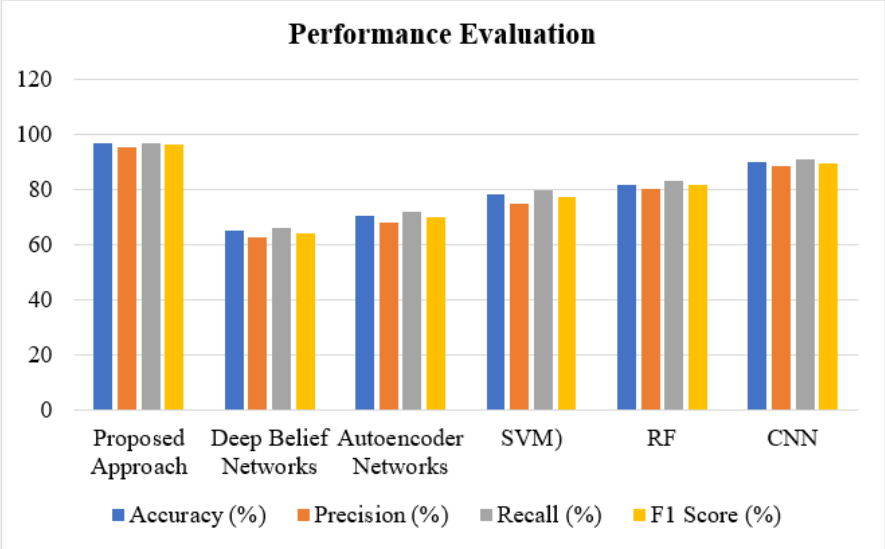

Figure 5 Performance Evaluation of Speech Recognition Algorithms

8. CONCLUSION

In conclusion, this study demonstrates the effectiveness of a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for robust speech emotion recognition. The proposed model achieves high performance, with an accuracy of 96.97% along with strong precision, recall, and F1-score values, confirming its ability to accurately classify diverse emotional states from speech signals. By jointly leveraging spatial feature extraction through CNNs and temporal sequence modeling through LSTMs, the system overcomes the limitations of conventional approaches that depend on isolated feature analysis or traditional machine learning techniques. This integrated architecture enables a more comprehensive understanding of emotional patterns in speech, leading to improved reliability and generalization. Furthermore, comparative analysis with existing methods such as Deep Belief Networks, Stacked Autoencoders, Support Vector Machines, and Random Forest models highlights the superior performance and robustness of the proposed approach. Despite the promising results, there remains scope for further enhancement through the incorporation of contextual information, multimodal data, and speaker-specific characteristics. Future research can also focus on optimizing the model for real-time deployment and reducing computational complexity. Overall, this work contributes significantly to the advancement of emotion recognition systems and provides a strong foundation for practical applications in areas such as healthcare monitoring, intelligent customer support systems, and adaptive human–computer interaction.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G., Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32-80, 2001.

Schuller, B., Rigoll, G., & Lang, M, Speech emotion recognition combining acoustic features and classifiers. In IEEE International Conference on Multimedia and Expo, 2004. ICME '04, 2009

Lee, C. M., & Narayanan, S. S., Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, 13(2), 293-303, 2005

Ververidis, D., & Kotropoulos, C. Emotional speech recognition: Resources, features, and methods. Speech Communication, 48(9), 1162-1181, 2006

Picard, R. W., Affective computing: From laughter to emotion recognition. IEEE Transactions on Affective Computing, 1(1), 11-17, 2010.

Fayek, H. M., Lech, M., & Cavedon, L., Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, 60-68, 2017

Akçay, M. B., & Oguz, K, Speech emotion recognition: Deep learning-based feature extraction techniques. IEEE Access, 8, 105584-105594, 2020

Latif, S., Qayyum, A., Usama, M., & Qadir, J., Speech emotion recognition using deep learning: A review. IEEE Transactions on Affective Computing, 2022

Tian, Y., Zhang, X., & Cao, Y. Integrating speech and text for emotion recognition using transformers. IEEE Transactions on Neural Networks and Learning Systems, 33(6), 2611-2622, 2023

Zhao, G., Schuller, B., & Zhang, X. Multimodal emotion recognition combining speech, facial expressions, and physiological signals. IEEE Transactions on Multimedia, 24, 2257-2267, 2023

G. Vijendar Reddy, SukanyaLedalla ,Avvari Pavithra, A quick recognition of duplicates utilizing progressive methods 'International Journal of Engineering and Advanced Technology (IJEAT)' at Volume-8 Issue-4, April 2019.

Wei, B.; Hu, W.; Yang, M.; Chou, C.T. From real to complex: Enhancing radiobased activity recognition using complex-valued CSI. ACM Trans. Sens. Netw. (TOSN) 2019, 15, 35.

Avvari, Pavithra, et al. "An Efficient Novel Approach for Detection of Handwritten Numericals Using Machine Learning Paradigms." Advanced Informatics for Computing Research: 5th International Conference, ICAICR 2021, Gurugram, India, December 18–19, 2021, Revised Selected Papers. Cham: Springer International Publishing, 2022.

Ledalla, Sukanya, R. Bhavani, and Avvari Pavitra. "Facial Emotional Recognition Using Legion Kernel Convolutional Neural Networks." Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4. Springer Singapore, 2021.

Brain Tumors Classification System Using Convolutional Recurrent Neural Network V. Akila, P.K. Abhilash, P Bala Venakata Satya Phanindra, J Pavan Kumar, A. Kavitha E3S Web Conf. 309 01075 (2021) DOI: 10.1051/e3sconf/202130901075.

Raju, NV Ganapathi, V. Vijay Kumar, and O. Srinivasa Rao. "Authorship Attribution of Telugu Texts Based on Syntactic Features and Machine Learning Techniques." Journal of Theoretical & Applied Information Technology 85.1 (2016).

Prasanna Lakshmi, K., Reddy, C.R.K. A survey on different trends in Data Streams (2010) ICNIT 2010 - 2010 International Conference on Networking and Information Technology, art. no. 5508473, pp. 451-455.

Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng "Speech emotion recognition: Features and classification models", Digital Signal Processing 22 (2012) 1154–1160.

Pavol Harar, Radim Burget and Malay Kishore Dutta "Speech Emotion Recognition with Deep Learning", IEEE (2017) 4th International Conference on Signal Processing and Integrated Networks (SPIN), pg no 78-1-5090-2797- 2/17.

Dias Issa, M. Fatih Demirci, Adnan Yazici "Speech emotion recognition with deep convolutional neural networks" Elsevier Ltd, Biomedical Signal Processing and Control 59 (2020) 101894.

Shambhavi Sharma "Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks", 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20 @IEEE.

Tanvi Puri, Mukesh Soni, Gaurav Dhiman, Osamah Ibrahim Khalaf, Malik alazzam, and Ihtiram Raza Khan "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network" Hindawi Journal of Healthcare Engineering Volume 2022.