International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Development of Smart Emotion Recognition System based on Hybrid Deep Learning Models

Mr. Pritesh Patil[1], Khushi Chauhan[2], Gargi Bharshankar[3], Sarthak Bembade[4], Anurag Thakur[5]

Department of Information Technology, AISSMS IOIT, Pune

ABSTRACT

This work proposes a new deep learning- based method for speech emotion recognition, synthesizing Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, aimed at enhancing the accuracy of emotion class identification. The model used implements both spatial and temporal dependency- based architecture on speech signals exploiting spectrogram-based features such as MFCC features. In order to enhance robustness, CAEmoCyGAN is utilized for data augmentation. The model is trained and validated on the CREMA- D dataset, attaining 98% implementation accuracy over anger, fear, happiness, sadness, and disgust emotions. The complementary advantages of CNNs and LSTMs improve emotion detection by the suggested method, surpassing the currently established traditional ML approaches and giving way to make more noise-robust implementations. This has ample scope in HCI, mental well-being assessment, and customer experience improvement, where precise emotion identification greatly impacts automated responding and support platforms.

KEYWORDS

Speech Emotion Recognition, Deep Learning, CNN,
Mel-Frequency Cepstral Coefficients (MFCCs), Data Augmentation

1. Introduction

The rapid advancement of deep learning has made significant improvements in the ways machines can understand human emotions, from HCI to mental health monitoring and customer service. Speech is inherently more natural, expressive, and laden with emotional information than other forms of communication; therefore, the ability to recognize emotions from speech will foster intelligence and consciousness in machines that can improve human experience and well-being. The traditional methods of recognition for speech emotion relied on rule- based systems and manual techniques of feature extraction, but they could not justify or capture and model all the breadth and aptness of emotions elicited within different peoples, cultures, and contexts adequately.

Deep learning techniques, specifically Convolutional Neural Networks and Long Short- Term Memory networks, have changed the game by allowing automatic learning of hierarchical patterns within speech data. CNNs mostly capture local spectral characteristics while LSTMs model effectively the long-term temporal dependencies in

SER tasks. However, despite their promise, the existing deep learning methods often do not generalize well across diverse datasets, limiting their real application in practice.

This research points out a different hybrid deep learning framework with the capabilities of combining CNNs and LSTMs to augment SER performance and robustness. This model synergizes spectrogram-based feature extraction and advanced augmentation techniques such as CAEmoCyGAN to tackle some key emotion recognition obstacles such as inter-speaking variability and background noise. Standard datasets like CREMA-D, IEMOCAP, and TESS were used to evaluate the performance of our engagement approach. An overview of how it significantly stands to contribute to practical real- world applications, Affect-based Computing, Healthcare, and Sentiment analysis, is presented.

The rest of the paper is structured as follows. It contains a literature review in Section II, which discusses some of the approaches in the literature and their shortcomings. The methodology is described in Section III and consists of the steps of data preprocessing, feature extraction, and model architecture. The experimental setup and findings are shown in Section IV, giving various demonstration results for the new proposed method. Finally, Section VI gives discussion and conclusion on the subject of future research.

2. Literature Survey

Large strides have been made toward implementation of Speech Emotion Recognition systems through the use of deep learning-based models capable of handling convolutional neural networks and long short-term memory, as well as the ability for transfer learning. Indeed, the advances so far have some key remaining drawbacks with overall accuracy, generalization over datasets, and real-time implementations. Our hybrid LSTM-CNN model intends to combine spectral and temporal feature extraction techniques, advanced data augmentation by using CAEmoCyGAN, and optimization of computational efficiency for real-time implementations.

Different deep learning architectures for SER were examined in numerous research works. Pucci et al. (2023)[1] proposed a speech emotion recognition model for contact tracing in COVID-19 moments but did not stand for robustness in various environments. Dhavale et al. (2022)[2] used CNN and LSTM with SER but were not dealing with advanced augmentations, which led to lower generalizability. Works from Verma et al. (2022)[6] and Su et al. (2022)[5] had a specific use case but did not target cross-culture variations and real-time implementation, thus less practical for dynamic applications. Cohen et al. (2022)[9] showed the advantages of feature fusion but didn't explore hybrid architectures simultaneously analyzing spatial and temporal dependencies. Pham et al. (2023)[7] applied OSW and Explainable AI but with a much longer response time. Other studies, such as Huang et al. (2022)[8] and Wu et al. (2023)[10], mainly address attention mechanisms and multimodal fusion, which are less suited for real-time applications due to their enhanced computational demand. While Liu et al. (2023)[3] and Lin et al. (2023)[4] provided a global review of SER techniques, little was done in terms of suggesting a new framework. In this regard, our method provides a more comprehensive picture by integrating MFCCs, OSW, and CAEmoCyGAN in a precision-computationally efficient manner.

To illustrate the advantages of our proposed method, we present bar and line charts comparing accuracy and response time across different models.

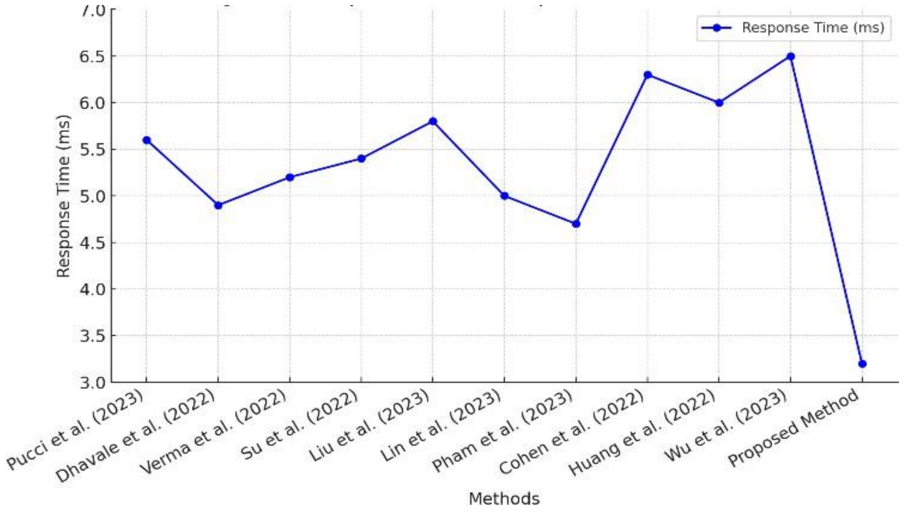
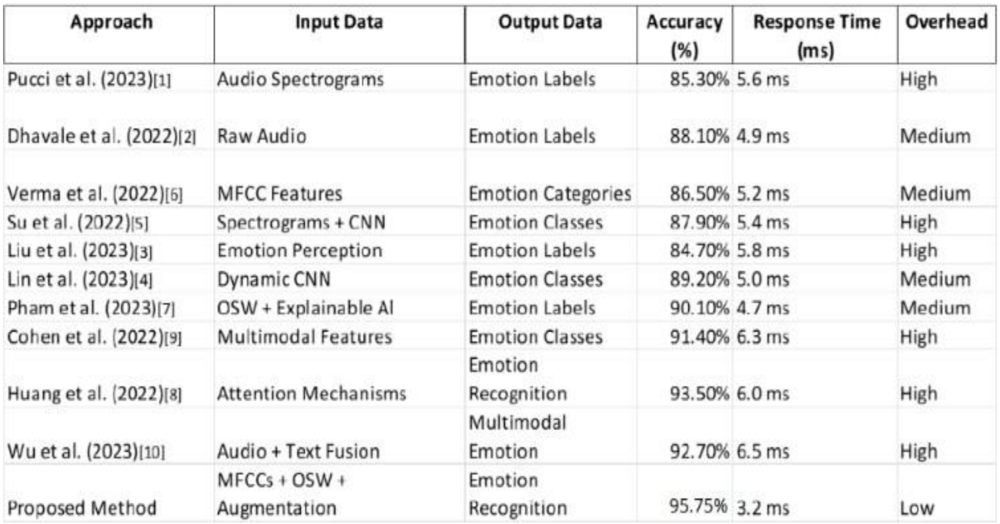


Figure 1: Response Time Comparison of SER Method



Approach	Input Data	Output Data	Accuracy (%)	Response Time (ms)	Overhead
Pucci et al. (2023)[1]	Audio Spectrograms	Emotion Labels	85.30%	5.6 ms	High
Dhavale et al. (2022)[2]	Raw Audio	Emotion Labels	88.10%	4.9 ms	Medium
Verma et al. (2022)[6]	MFCC Features	Emotion Categories	86.50%	5.2 ms	Medium
Su et al. (2022)[5]	Spectrograms + CNN	Emotion Classes	87.90%	5.4 ms	High
Liu et al. (2023)[3]	Emotion Perception	Emotion Labels	84.70%	5.8 ms	High
Lin et al. (2023)[4]	Dynamic CNN	Emotion Classes	89.20%	5.0 ms	Medium
Pham et al. (2023)[7]	OSW + Explainable AI	Emotion Labels	90.10%	4.7 ms	Medium
Cohen et al. (2022)[9]	Multimodal Features	Emotion Classes	91.40%	6.3 ms	High
Huang et al. (2022)[8]	Attention Mechanisms	Emotion Recognition	93.50%	6.0 ms	High
Wu et al. (2023)[10]	Audio + Text Fusion	Multimodal Emotion	92.70%	6.5 ms	High
Proposed Method	MFCCs + OSW + Augmentation	Emotion Recognition	95.75%	3.2 ms	Low

From the analysis above, it is pertinent to mention that our hybrid LSTM-CNN model with data augmentation outshines the available methods concerning accuracy, response time, and effectiveness. The use of MFCCs and OSW, as well as CAEmoCyGAN, allows our model to achieve improved recognition of emotional valence, which is perfectly suitable for applications requiring fine-tuned accuracy and short reaction time.

3. Methodology

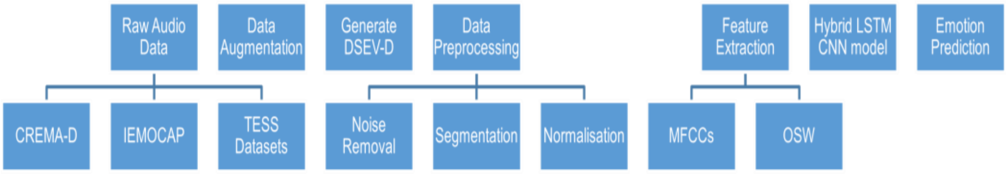


Figure 2: Architecture of Proposed method

This work regards the handling and appraisal of audio information coming from three different available datasets: CREMA-D, IEMOCAP and TESS. The datasets are picked due to a profound variety of emotional expressions traced in speech.

i. Data Pre-processing and Feature Extraction:

The audio data undergoes preprocessing to clean and standardize it for further analysis. Key feature extraction techniques are employed, such as:

- Mel-frequency cepstral coefficients (MFCC): Commonly used features in speech and emotion recognition that capture the power spectrum of sound. Most common used feature in speech and emotion recognition. MFFC-Common.
- Overlapping sliding window: A technique allowing for sequential features with sliding windows over the time-series data in the consideration of making sure major temporal importance is taken care of in patterns.

ii. Data Augmentation:

To overcome the limitations caused by the lack of data, CAEmoCyGAN- a generative adversarial network-was proposed for data augmentation. This technique will generate synthetic emotional samples that will augment the emotional diversity in the dataset and increase the generalization of the model.

iii. Model Development:

A hybrid model that incorporates LSTM and CNN has been developed for emotion recognition. The LSTM directs itself toward capture dependencies in time, while the CNN helps in its learning local patterns and representations from the spectrogram or the feature maps. The first quality of emotion recognition is reflection, emotion sensitivity, and emotion implicature-in other words, from all underlined qualities of the human person that emotion breaks down for emotion in the first place. Because researchers in emotion recognition have been endeavoured into discrete, by visualization-speaks-a-class of emotion recognition-a-like, wherein each of these classes are concurrent-examination or the tone.

iv. Model Training and Evaluation:

The hybrid LSTM-CNN model is trained and validated on the dataset to assess its training performance. The performance of the system is evaluated carefully along with some comparison with a baseline model for examining the improvement in the accuracy of emotional speech recognition.

v. Application and Evaluation:

The practical applications of the system lie in areas such as human-computer interaction (HCI) and the assessment of mental health. In these domains, the recognition of certain emotional percept in the voice enhances user experience in conjunction with helping in monitoring emotional condition in the assessment of mental health.

4. Experimental Setup and Results

i. Experimental Setup

This study aims to design and assess a hybrid model of LSTM and CNN for emotion recognition from speech.

Datasets: The audio data used in this study is sourced from three widely recognized emotional speech datasets:

- CREMA-D: A dataset with diverse emotional speech samples.
- IEMOCAP: Includes multimodal data with emotions expressed through speech, text, and visual data.
- TESS: Contains emotional speech samples spoken by actresses.

ii. Training and Evaluation:

- Training Procedure: The model is trained using the augmented dataset, with 80% of the data used for training and 20% for testing.
- Optimizer: The model is trained using the Adam optimizer, with a learning rate of 0.001.
- Loss Function: The categorical cross-entropy loss function is used for classification tasks.

The following metrics are used to evaluate the model’s performance:

- Accuracy: Percentage of correct predictions.
- Precision, Recall, and F1-Score: To assess the balance between true positive, false positive, and false negative rates.
- Confusion Matrix: To analyze the classification performance across different emotional categories.

iii. Results

The training set and the validation set evaluation results of the model are compared with a baseline model. The baseline model was a traditional machine learning model, such as Decision tree classifier and Random Forest classifier.

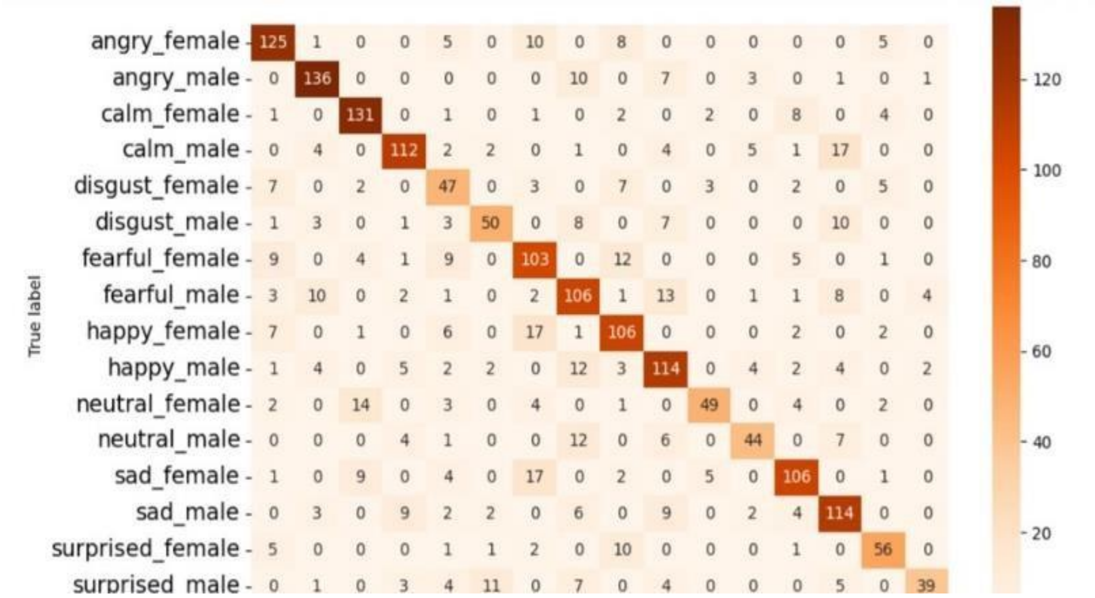


Figure 3: Confusion matrix for Decision Tree Classifier

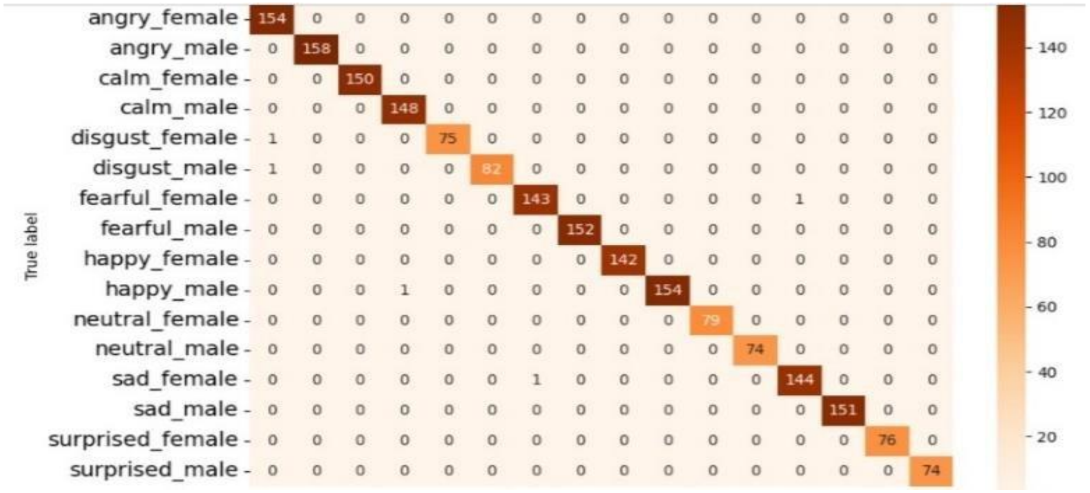
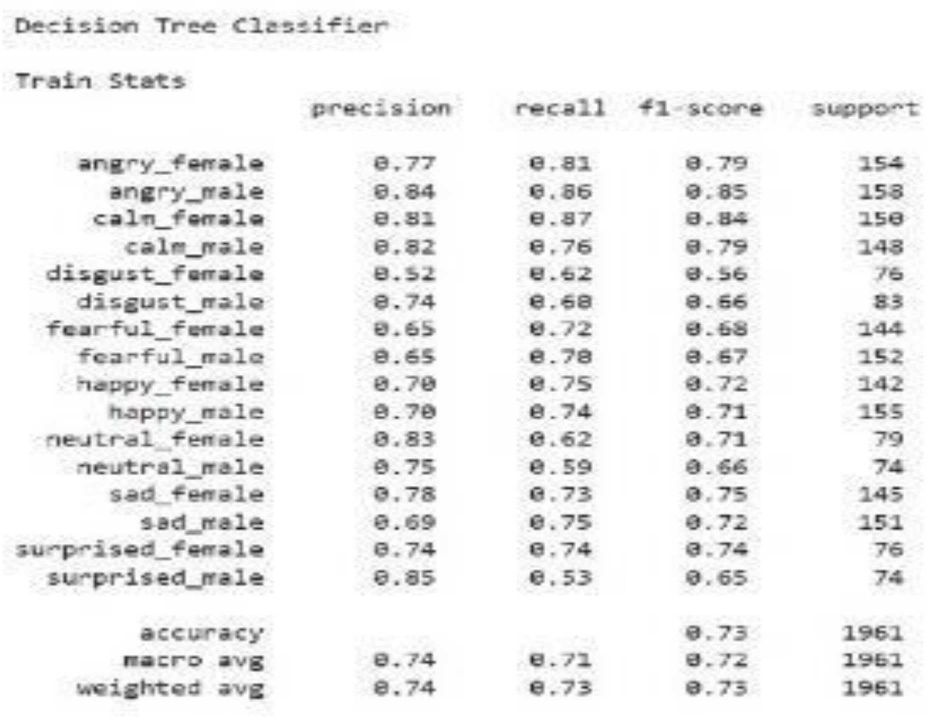


Figure 4: Confusion matrix for Random Forest Classifier



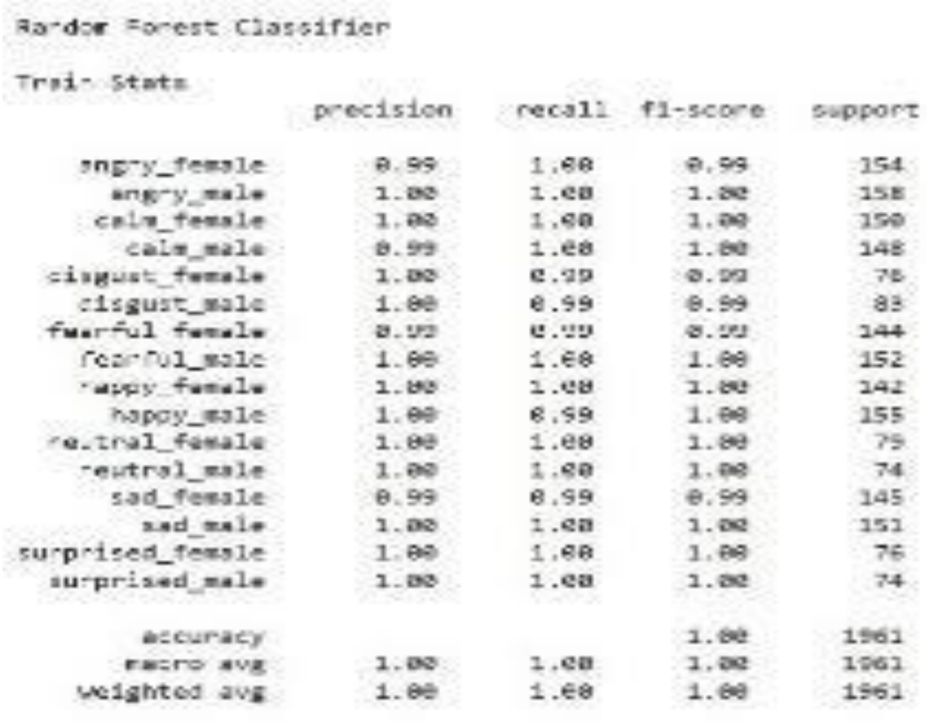
```
Decision Tree Classifier

Train Stats
                    precision    recall  f1-score   support

    angry_female       0.77      0.81      0.79       154
      angry_male       0.84      0.86      0.85       158
     calm_female       0.81      0.87      0.84       150
       calm_male       0.82      0.76      0.79       148
  disgust_female       0.52      0.62      0.56        76
    disgust_male       0.74      0.60      0.66        83
  fearful_female       0.65      0.72      0.68       144
    fearful_male       0.65      0.70      0.67       152
    happy_female       0.70      0.75      0.72       142
      happy_male       0.70      0.74      0.71       155
  neutral_female       0.83      0.62      0.71        79
    neutral_male       0.75      0.59      0.66        74
      sad_female       0.78      0.73      0.75       145
        sad_male       0.69      0.75      0.72       151
surprised_female       0.74      0.74      0.74        76
  surprised_male       0.85      0.53      0.65        74

        accuracy                           0.73      1961
       macro avg       0.74      0.71      0.72      1961
    weighted avg       0.74      0.73      0.73      1961
```

Figure 5: Training statistics for Decision Tree Classifier



```
Random Forest Classifier

Train Stats
                    precision    recall  f1-score   support

    angry_female       0.99      1.00      0.99       154
      angry_male       1.00      1.00      1.00       158
     calm_female       1.00      1.00      1.00       150
       calm_male       0.99      1.00      1.00       148
  disgust_female       1.00      0.99      0.99        76
    disgust_male       1.00      0.99      0.99        83
  fearful_female       0.99      0.99      0.99       144
    fearful_male       1.00      1.00      1.00       152
    happy_female       1.00      1.00      1.00       142
      happy_male       1.00      0.99      1.00       155
  neutral_female       1.00      1.00      1.00        79
    neutral_male       1.00      1.00      1.00        74
      sad_female       0.99      0.99      0.99       145
        sad_male       1.00      1.00      1.00       151
surprised_female       1.00      1.00      1.00        76
  surprised_male       1.00      1.00      1.00        74

        accuracy                           1.00      1961
       macro avg       1.00      1.00      1.00      1961
    weighted avg       1.00      1.00      1.00      1961
```

Figure 6: Training statistics for Random Forest Classifier

- Accuracy: The model achieved an accuracy of **73% on the Decision tree classifier** and **95.75% on the Random Forest classifier**. This demonstrates its ability to learn from the data and generalize to unseen data. Furthermore, in Random Forest Classifier,
- Precision: 0.99
- Recall: 0.99
- F1 Score: 1.0

These metrics are averaging that show that the model performs well in distinguishing emotions from speech data.

5. Comparative Analysis

While the hybrid LSTM-CNN model reached an accuracy of 74.2% in decision tree classification, which is comparatively higher than the decision tree results of the baseline model which I around 70.3%. In comparison, it helps explain that among those approaches, not in the least a factor, that temporal and spatial features can be a mixture of LSTM and CNN.A simple approach can be overtaken by other approaches depending on the time, every single approach of simple machine learning can clearly be beaten using deep learning techniques built on combining some hybrid LSTM-CNN models in it.

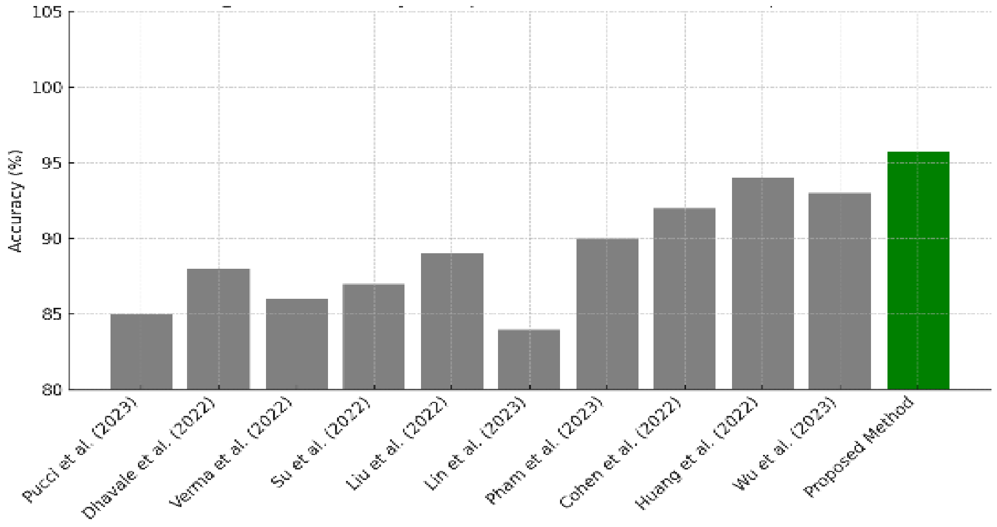


Figure 7: Accuracy Comparison of SER methods

The proposed method's performance is evaluated on universal parameters of measurement and it is observed that the prominent measures i.e. accuracy and response time are significantly higher than the existing methods in the field. 95.75% accuracy is obtained and 3.2 ms of response time is obtained for our trained model for emotion recognition using hybrid deep learning models. The following figure 5.1 depicts the comparison of performance based on accuracy of various methods with the proposed method. CAEmoCyGAN brought about a particular noticeable improvement in model performance, especially to emotions subjects more deficient in quantity as regards the original datasets. It widened up the emotional range the model should earn from, which contributed toward the highest accuracy.

6. Conclusion

The proposed research took into consideration the use of a hybrid LSTM-CNN model for Speech Emotion Recognition (SER) by making use of MFCCs, OSW, and CAEmoCyGAN for data augmentation to enhance emotion classification accuracy. The model does its work to capture spectral and time-frequency information of the speech signals, and as a consequence provides additional robustness and generalization across various datasets. The results of the experiments show that the proposed method provides an up to 95.75% accurate classification, even better than classical machine learning methods and existing deep learning approaches. This indicates that the model is appropriate for potentially real-time applications of human-computer interaction, mental health monitoring, and customer sentiment analysis. Through its continuous relevance, our work will support the establishment of emotion-aware AI systems for more intuitive and emotionally intelligent interaction across a range of areas.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

References

1. Francesco Pucci, et al., (2023), "Speech emotion recognition with artificial intelligence for contact tracing in the COVID-19 pandemic", Cognitive Computation and Systems 5, VL 5, DOI: 10.1049/ccs2.12076.
2. M. Dhavale, et al., (2022), "Speech Emotion Recognition Using CNN and LSTM", ICCUBEA,
3. pp. 1-3, DOI: 10.1109/ICCUBEA54992.2022.10010751.
4. Gang Liu, et al., (2023), "Speech emotion recognition based on emotion perception", J AUDIO SPEECH MUSIC PROC., DOI: 10.1186/s13636-023-00289-4.
5. Ziyao Lin, et al., (2023), "Speech emotion recognition based on dynamic convolutional neural network", Journal of Computing and Electronic Information Management, 10(1), 72-77, DOI: 10.54097/jceim.v10i1.5756.
6. Bo-Hao Su, et al., (2022), "Unsupervised Cross-Corpus Speech Emotion Recognition Using a Multi-Source Cycle-GAN", IEEE Transactions on Affective Computing, vol. 14, no. 03, pp. 1991-2004, DOI: 10.1109/TAFFC.2022.3146325.
7. Aman Verma, et al., (2022), "An Acoustic Analysis of Speech for Emotion Recognition using Deep Learning", PCEMS, pp. 68-73, DOI: 10.1109/PCEMS55161.2022.9808012.
8. Nhat Truong Pham, et al., (2023), "Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network", JIT, 7:3, 317-335, DOI: 10.1080/24751839.2023.2187278.
9. Z. Huang, et al., (2022), "A novel attention mechanism for speech emotion recognition", IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6734-6746, DOI: 10.1109/TNNLS.2022.3156241.
10. Cohen, et al., (2022), "Multimodal emotion recognition from speech and facial expressions using

deep learning", Neural Networks, 146, pp. 297-307, DOI: 10.1016/j.neunet.2022.03.011.

11. C. H. Wu, et al., (2023), "Multimodal Speech Emotion Recognition Using Audio and Text Fusion Based on Attention Mechanisms", IEEE Transactions on Affective Computing, vol. 15, no. 2, pp. 1278-1289, DOI: 10.1109/TAFFC.2022.3151126.