Speech Emotion Classification Using Advanced Machine Learning Models

Course Project for AI and Interactive Systems (ELEC872)

December 2024

Reza Jahantigh (Student #: 20379506)
Department of ECE
Queen's University, Kingston, Canada
reza.jahantigh@queensu.ca

Tomi Kofo-Alada (Student #: 20152073)
Department of ECE
Queen's University, Kingston, Canada
t.kofoalada@queensu.ca

Abstract

Speech Emotion Recognition (SER) plays a critical role in advancing human-computer interaction (HCI) by enabling systems to interpret emotional states through speech signals. This study investigates the effectiveness of a Convolutional Neural Network (CNN) model as the primary architecture for SER, alongside four baseline models—CNN-LSTM, LSTM, Random Forest (RF), and Support Vector Machine (SVM)—to benchmark performance. Comprehensive preprocessing, including segmentation, silence removal, MFCC extraction, and normalization, was applied to the RAVDESS dataset to ensure data quality. Experimental results demonstrate that the CNN model achieves an accuracy at 80.0%, slightly lower than a DCNN-CFC-SVM hybrid model (81.3%) and a CNN-based approach (82.0%) reported in related works. Despite this, the CNN model showcases robust spatial feature extraction, making it a competitive solution for SER tasks. This study underscores the potential of spatial feature extraction while identifying limitations in temporal modeling, offering insights for future advancements in emotion recognition through speech.

1. Introduction

Emotion plays a crucial role in human interaction, influencing communication and facilitating advancements in affective computing applications such as automated vehicles and mental health monitoring systems [11, 20, 35]. While humans naturally interpret emotions through cues like facial expressions, tone of voice, and behavior, the subjective nature of emotions makes their recognition challenging for computers [1].

Speech signals have emerged as a rich source of emotional content, making Speech Emotion Recognition (SER) a prominent area of research. SER has applications across domains, including human-computer interaction, safety monitoring, education, smart homes, and healthcare systems [17, 33, 34]. Unlike facial expression analysis, speech-based methods offer a straightforward way to convey and detect emotions, leveraging features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, intensity, and spectral information for classification [21].

To enable consistent benchmarking, various datasets such as Ryerson Audio-Visual Database for Emotional Speech and Song (RAVDESS) [18], the Remote Collaborative and Affective (RECOLA) dataset [27], the Geneva Multimodal Emotion Portrayal (GEMEP) corpus [2], the Berlin Emotional Voice Database (EMODB) [3], and the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [4] have been widely adopted in SER research. SER methods typically use a set of various features such as MFCCs, energy, pitch, intensity, length, and rhythm in speech signals to recognize emotions preceded by preprocessing techniques such as noise reduction, segmentation, silence removal, and normalization to provide high-quality data for classification [21, 35].

Traditional machine learning approaches like Support Vector Machines (SVM) [7, 22], Deep Belief Networks [15], Hidden Markov Models (HMM) [31], and Gaussian Mixture Models (GMM) [5, 19] relied on manual feature extraction, but their performance often lacked robustness to variations in speech signals[25]. Recent advances in deep learning, particularly CNN and CNN-LSTM architectures [9, 10, 14, 16, 26, 30], have significantly improved SER accuracy through automatic feature extraction and the ability to model spatial and temporal dependencies. Hybrid models like CNN-SVM [12] and attention-based CNNs [8] have further pushed the boundaries of SER performance.

In this study, various classifiers, with a focus on CNN, are implemented and evaluated on the RAVDESS dataset to identify the best-performing model for speech emotion classification through a comparative analysis. The main contribution of this work is the development of a preprocessing

pipeline to ensure high-quality input data for deep learning, using only normalized MFCCs instead of the commonly used feature sets in previous studies. This approach provides a reproducible framework leveraging the RAVDESS dataset, enabling further advancements in emotion recognition through improved preprocessing and model development techniques. The rest of the report is organized as follows:
Section 2 provides an overview of related work, Section 3 describes the preprocessing steps and classification models, Section 4 details the experimental setup, Section 5 presents the results and discussion, and Section 6 concludes the study.

2. Related Work

A combination of the RAVDESS dataset with the Toronto Emotional Speech Set (TESS) is used in [23] to diversify the training data. Then, features such as MFCC, Chroma, and Mel Spectrogram are used to train multiple traditional classifiers, including Gradient Boosting, Multi-Layer Perceptron (MLP), Random Forest (RF), and SVM. This study highlighted the effectiveness of feature diversity in improving emotion recognition performance and provided comparative insights on classifier accuracy across different datasets. Similarly, researchers in [32] used a variety of features to compare the performance of various ML methods, including SVM, RF, MLP, DT, and CNN. The computational efficiency of CNN in automatically identifying essential features without human intervention is highlighted by training CNN with two, three, and four layers using the RAVDESS dataset's versatility as a benchmark. Moreover, the same approach is used in [13] using one-dimensional CNN on various features on multiple datasets RAVDESS, EMODB, and Interactive Emotional Dyadic Motion Capture (IEMOCAP). In this work, an incremental method for modifying the initial model is used on Mel-frequency cepstral coefficients, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features extracted directly from raw sound data without the need for conversion to visual representations to improve classification accuracy.

Farooq *et al*. [6] proposed a methodology combining a deep convolutional neural network (DCNN) with correlation-based feature selection (CFS) for emotion classification. They used pre-trained AlexNet to extract features from log-mel spectrograms and refined them through CFS to reduce dimensionality and enhance model performance. The classification, applied to the RAVDESS dataset using SVM, RF, and K-Nearest Neighbors (KNN), highlighted the effectiveness of deep learning for feature extraction and traditional classifiers for emotion recognition. In this study, the leave-one-speaker out (LOSO) scheme, in which one speaker is selected for testing, and the rest of the speakers are used for training purposes, is used for speaker-independent SER, and a ten-fold cross-validation technique is applied to perform speaker-dependent experiments. The speaker-independent SER showed better results with existing handcrafted features-based SER approaches using the method introduced in this study.

The study in [29] introduced a deep framework combining CNN and BiLSTM. The CNN extracts discriminative and salient features from speech spectrograms, while the BiLSTM captures temporal information in the speech data. Instead of using the entire utterance, key segments are selected for feature extraction, followed by normalization to reduce computational complexity and ensure accurate recognition. The method was tested on various datasets, demonstrating a reduction in processing time and improved accuracy compared to state-of-the-art SER methods.

3. Methodology

3.1. Preprocessing

Preprocessing is vital for preparing speech data for effective emotion classification. In this study, the RAVDESS dataset underwent multiple preprocessing steps including segmentation, silence removal, resampling, Mel Frequency Cepstral Coefficients (MFCC) extraction, and normalization with each addressing specific challenges associated with speech data.

1. **Segmentation:** Speech signals are divided into smaller, meaningful segments of equal length to increase the number of training samples by capturing emotion-relevant features from short time frames. This segmentation also facilitates easier comparison between different approaches. Segments longer than 250 ms are known to contain sufficient meaningful information for emotion detection [24]. Accordingly, in this study, each speech signal is divided into equal segments of 270 ms in length.
2. **Silence Removal:** Segmentation may produce speech signals with redundant information, particularly in unvoiced or silent parts of the original signal, as shown in Figure 1, which can lead to various silent segments with different labels. To address this issue, silent segments are removed from the dataset. This step reduced dataset size and enhanced emotion classification accuracy [28].
3. **Resampling:** Segmentation and silence removal can introduce class imbalances, favoring majority classes. To address this, resampling techniques—oversampling and undersampling—were considered. Oversampling increases the number of samples in the minority class while undersampling reduces the size of the majority class by removing samples to achieve balance [6]. This study employs oversampling to align the size of minority classes with that of the largest class, ensuring balanced data distribution.

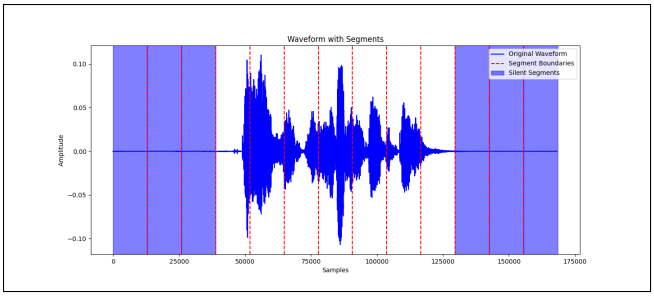


Figure 1. Segmented speech signals with silent parts

4. **MFCC Extraction:** MFCCs, known for capturing spectral properties of speech, were computed for each segment using Librosa. Each segment, processed at a consistent sampling rate of 22,050 Hz, generated 40 coefficients averaged across time frames. This dimensionality reduction preserved critical spectral variations while standardizing segment representations for efficient computation and reduced risk of overfitting.
5. **Normalization:** Z-score normalization standardized MFCC features to zero mean and unit variance, mitigating the influence of variable recording conditions. By focusing on feature differences rather than absolute magnitudes, normalization stabilized model training and improved convergence. The normalized features were stored in a CSV file for subsequent analysis and classification.

3.2. Classification

This project utilizes a CNN model as the primary architecture for SER, supported by four baseline models—CNN-LSTM, LSTM, RF, and SVM—to comprehensively evaluate performance.

The CNN model focuses on spatial feature extraction from Mel-Frequency Cepstral Coefficients (MFCC) representations. Its architecture comprises multiple convolutional layers for feature extraction, followed by pooling layers to reduce dimensionality while retaining essential features. The final classification is performed through fully connected layers with a sigmoid activation, mapping the extracted features to the numbers of emotional categories.

The CNN-LSTM model, included as a baseline, combines the spatial feature extraction capabilities of CNNs with the temporal sequence modeling strengths of LSTM layers. It includes two convolutional layers with batch normalization and pooling, followed by two LSTM layers to process temporal dependencies. Fully connected layers with a softmax activation finalize the classification. Dropout and L2 regularization are applied throughout to mitigate overfitting and stabilize the model during training.

The LSTM baseline focuses exclusively on capturing temporal patterns in audio signals. Its architecture consists of three stacked LSTM layers, with batch normalization to stabilize training, and a dense layer with softmax activation for classification.

The RF model serves as a classical machine learning baseline. It uses an ensemble of 400 decision trees to classify emotions based on handcrafted features extracted from speech signals.

Finally, the SVM model employs a radial basis function (RBF) kernel to classify emotions by separating them in a high-dimensional feature space. It provides a margin-based baseline for assessing traditional classifiers against deep learning approaches. This classification setup encompasses a diverse range of methods, emphasizing the comparison of spatial, temporal, and traditional approaches for SER tasks.

Process	Description	No. Samples
Initial Dataset	Original number of speech signals	1,440
Segmentation	After dividing signals into segments of 270 ms	20,408
Silent Removal	After removing silent segments	9,580
Resampling	After balancing the dataset through oversampling	12,056

Table 1. Number of samples after each data preprocessing step in the RAVDESS dataset.

4. EXPERIMENT SETUP

4.1. Dataset

The RAVDESS dataset [18], which is publicly available, is used in this study. It consists of recordings from 24 actors, with an equal distribution of 12 male and 12 female performers, featuring both audio and visual components. The 1,440 files in this dataset maintain a consistent level of quality, with a sample rate of 48000Hz and a resolution of 16 bits. The dataset contains a broad range of emotional expressions, including neutral, calm, happy, sad, angry, fearful, disgusted, and surprised, as shown in Figure 2. For this work, only the audio recordings of two English sentences, 'Kids are talking by the door' and 'Dogs are sitting by the door,' are utilized. This dataset was selected because it provides a comprehensive set of emotional expressions from both male and female performers, making it well-suited for classification tasks. Additionally, its wide availability facilitates comparison with existing studies. For the RAVDESS dataset, as presented in Table 1, the preprocessing begins by dividing 1,440 speech signals into 20,408 equal segments, each with a duration of 270 ms. Silent segments are then removed, resulting in 9,580 segments containing meaningful information. Subsequently, resampling is performed to balance the distribution of emotions among the samples. The distribution of emotions before and after resampling is illustrated in Figure 3. Finally, MFCC features are extracted from these segments and normalized for use in training the classification model.

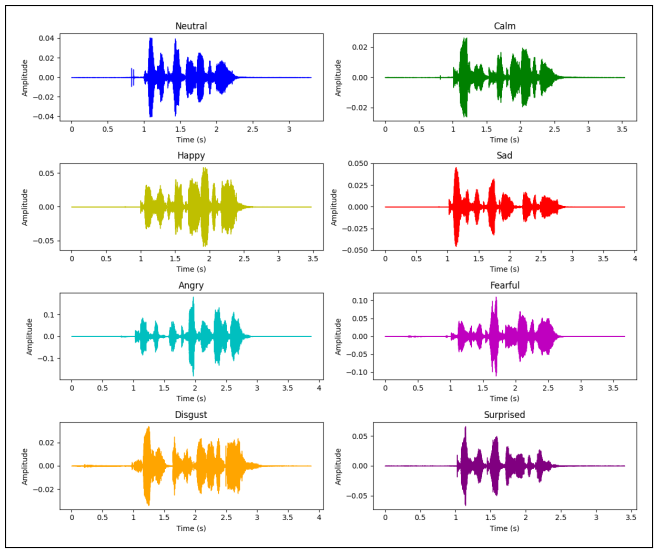
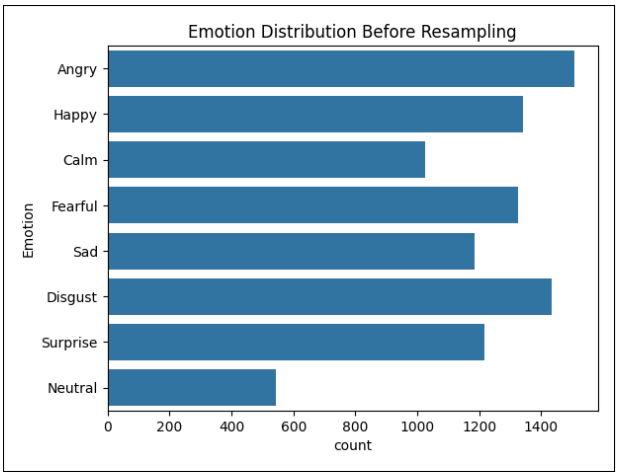
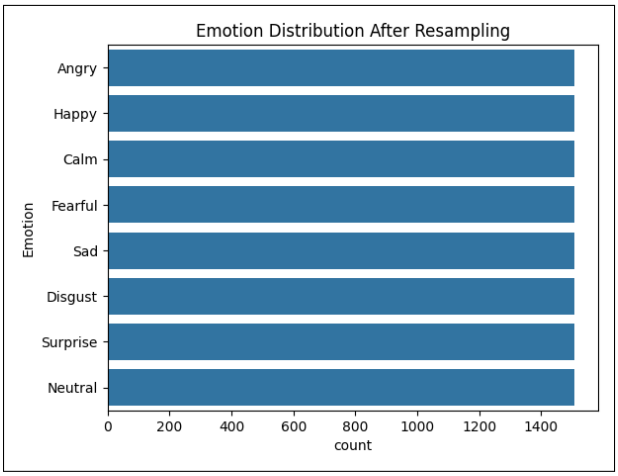


Figure 2. Sample waveforms of different emotions from the RAVDESS dataset



(a)



(b)

Figure 3. Emotion distribution for the RAVDESS dataset before and after resampling

4.2. Evaluation Metrics

To evaluate the performance of each model, we primarily use the accuracy metric. However, additional metrics such as F1-score, precision, and recall are also reported to provide a more comprehensive analysis.

- **Accuracy**: Accuracy describes the model's overall performance across all classes and is effective when all classes are equally important. It is calculated as the ratio of correct predictions (both true positives and true negatives) to the total number of predictions, as shown in Equation 1.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

- **Recall**: Recall measures the proportion of relevant cases that the model correctly identifies as relevant. It evaluates how well the model identifies positive samples and is calculated using Equation 2. A higher recall indicates better identification of positive samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

- **Precision**: Precision is the proportion of true positive predictions among all predicted positive samples. It assesses the model's ability to correctly classify positive samples, as defined in Equation 3.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

- **F1-score**: The F1-score is the harmonic mean of precision and recall, combining them into a single performance metric. It is particularly useful for evaluating models on imbalanced datasets. The F1-score is calculated as shown in Equation 4.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

4.3. Experiment Protocol

The Experiment Protocol was designed to evaluate and compare the performance of the CNN (main model) and the baseline models. The deep learning models—CNN, CNN-LSTM, and LSTM—were all trained using consistent parameters: 1,000 epochs, a batch size of 250, and standardized feature representations. These settings ensured fair and meaningful comparisons.

The CNN model employed the Adam optimizer with a learning rate of 0.0001. The CNN-LSTM model leveraged the RMSprop optimizer with the same learning rate, apply-

ing dropout (0.3) and L2 regularization throughout the architecture. The LSTM model was trained using similar hyperparameters, focusing on stabilizing temporal modeling through batch normalization.

Traditional machine learning baselines were trained differently. The RF model used 400 decision trees to combine predictions from multiple feature-based classifiers, while the SVM model used a radial basis function (RBF) kernel with a cost parameter of 30, trained on scaled MFCC features.

All models were evaluated on the RAVDESS dataset, which provides balanced classes across eight emotional categories. The experiments were conducted on Google Colab, leveraging an NVIDIA T4 GPU to ensure efficient training for the deep learning models. The software stack included TensorFlow (2.x), Keras, Scikit-learn, and Librosa for feature extraction. Using consistent hyperparameters for the deep learning models and tailored setups for the traditional models enabled a strong comparison of spatial, temporal, and feature-based approaches.

5. Result and Discussion

The results of the experiment highlight the performance of five models—CNN (the main model), CNN-LSTM, LSTM, RF, and SVM—on the RAVDESS dataset. This section presents evaluation of these models using metrics such as accuracy, precision, recall, F1-score, confusion matrices, and training/validation loss and accuracy trends over epochs.

5.1. Performance the Main Model (CNN)

According to the Table 2, the CNN model achieved the highest accuracy among all models, scoring 80.0%, with weighted precision, recall, and F1-score values of 0.79, 0.80, and 0.79, respectively. Its ability to extract spatial features from MFCC representations of speech signals underpins its superior performance. As shown in Figure 4, the model exhibited steady training and validation accuracy growth, with Figure 5 illustrating stable convergence. The normalized confusion matrix (Figure 6) highlights strong classification performance, particularly for "Neutral" (92%) and "Calm" (87%) emotions. However, misclassifications were observed in "Disgust" and "Fearful," indicating challenges in distinguishing emotions with overlapping acoustic characteristics.

5.2. Performance of Baseline Models

The baseline models—CNN-LSTM, LSTM, RF, and SVM—provided diverse insights into the effectiveness of different architectures for SER. The hybrid model, CNN-LSTM, achieved an accuracy of 77.4%, reflecting the strengths of combining spatial and temporal modeling. It excelled in classifying "Neutral" (94%) and "Calm" (86%) emotions but faced challenges with "Disgust" and "Fearful," similar to the CNN model. Exclusively focused on temporal dependencies, the LSTM model achieved 72.0%

Model	Accuracy (%)	F1-Score
CNN-LSTM	77.40	0.76
LSTM	72.00	0.72
RF	76.00	0.76
SVM	78.00	0.78
CNN	80.00	0.79

Table 2. Performance comparison of various trained models

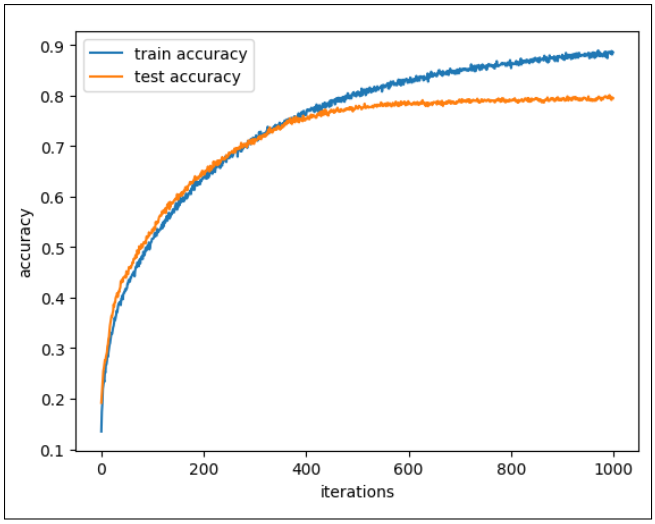


Figure 4. Training and Validation Accuracy Over Epochs for CNN

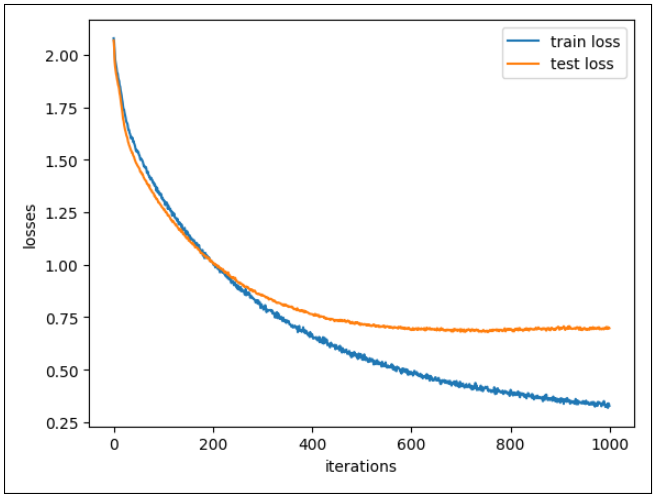


Figure 5. Training and Validation Loss Over Epochs for CNN

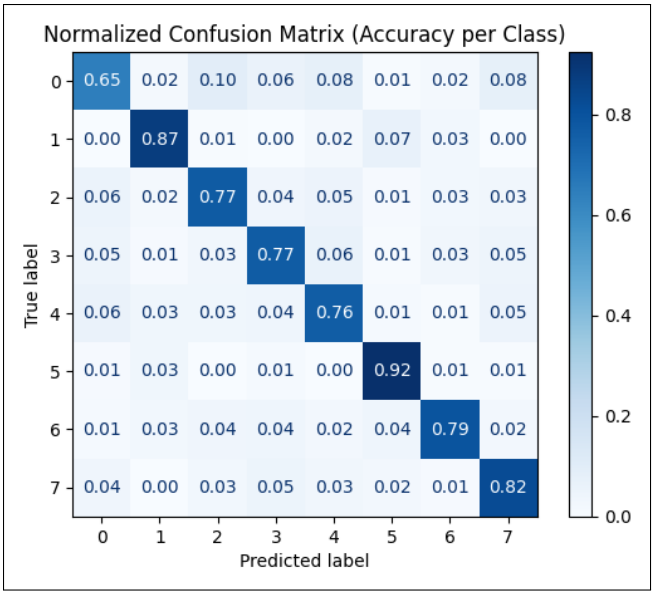


Figure 6. Normalized Confusion Matrix for CNN Model

accuracy. It performed well for "Neutral" (92%) and "Calm" (82%) emotions but struggled with "Angry" and "Fearful," highlighting the limitations of relying solely on sequential modeling.

RF achieved 76.0% accuracy. It demonstrated strong recall for "Neutral" (95%) and "Calm" (92%) but encountered misclassifications in "Disgust" and "Sad" due to overlapping acoustic features. Leveraging an RBF kernel, the SVM model achieved 78.0% accuracy. It performed strongly for "Neutral" (95%) and "Calm" (87%) but struggled to distinguish "Fearful" (75%) and "Disgust" (76%), likely due to shared acoustic traits. The varied performance across these models underscores the complementary nature of spatial and temporal feature extraction in SER. While CNN excelled as the main model, each baseline offered valuable comparative insights.

5.3. Comparative Analysis

The comparative evaluation highlights the performance of our primary CNN model and four baseline models—CNN-LSTM, LSTM, RF, and SVM—against models proposed in related works. Table 3 summarizes the accuracy metrics for our model compared to the exsiting works. The results reveal that while our CNN model achieves an accuracy of 80.0%, it is outperformed by the CNN model from Paper [32], which reported an accuracy of 82.0%, and the DCNN + CFS + SVM model proposed by Paper [6], which achieved 81.3%. These findings emphasize that while our CNN is robust, further optimization of the architecture, such as hybrid feature selection (CFS) or deep learning enhancements, could yield superior results. However, our CNN model outperforms the CNN model in [13], demonstrating the effectiveness of our preprocessing step, which utilizes only MFCCs instead of a diverse set of features as used in [13]. Overall, while our CNN model did not achieve the highest accuracy, its competitive performance underscores the importance of spatial feature extraction in the RAVDESS dataset. The comparative analysis further demonstrates that models leveraging hybrid approaches, such as DCNN with feature selection (CFS) or advanced architectures like CNN, can deliver marginally better results. Future improvements could explore integrating such hybrid techniques to bridge this performance gap.

Model	Accuracy (%)
RF [23]	30.00
CNN [13]	71.61
CNN [32]	82.00
DCNN + CFS + SVM [6]	81.30
CNN-BiLSTM [29]	77.02
Ours	80.00

Table 3. Comparison of proposed model with existing works

Moreover, Our Random Forest baseline model achieved 76.0%, which is significantly higher than the 30.0% reported by Paper [23]. This again highlights the effectiveness of our preprocessing techniques and hyperparameter tuning in improving the performance of traditional machine learning models.

6. Conclusion

This study presents a comparative evaluation of five models for SER, with the CNN model achieving the best performance among the implemented models, reaching an accuracy of 80.0% on the RAVDESS dataset and delivering competitive results compared to related works. The findings highlight the effectiveness of our preprocessing pipeline, which utilizes MFCCs instead of a diverse feature set, thereby potentially reducing computational time. However, the models face challenges in classifying emotions like "Disgust" and "Fearful," revealing limitations in capturing nuanced emotional patterns. Addressing these challenges is crucial to enhancing the adaptability of SER systems. Future research could focus on incorporating attention mechanisms or transformer-based architectures to improve classification accuracy, particularly for complex emotions. Additionally, this study does not account for the computational cost of the various methods, as the preprocessing steps are consistent across all models. Evaluating computational efficiency could be a focus for future studies.

References

[1] R Anusha, P Subhashini, Darelli Jyothi, Potturi Harshitha, Janumpally Sushma, and Namsamgari Mukesh. Speech emotion recognition using machine learning. In *2021 5th international conference on trends in electronics and informatics (ICOEI)*, pages 1608–1612. IEEE, 2021. 1

[2] Tanja Bänziger, Hannes Pirker, and K Scherer. Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions. In *Proceedings of LREC*, pages 15–019, 2006. 1

[3] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. A database of german emotional speech. In *Interspeech*, pages 1517–1520, 2005. 1

[4] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 1

[5] Xianglin Cheng and Qiong Duan. Speech emotion recognition using gaussian mixture model. In *2012 international conference on computer application and system modeling*, pages 1222–1225. Atlantis Press, 2012. 1

[6] Misbah Farooq, Fawad Hussain, Naveed Khan Baloch, Fawad Riasat Raja, Heejung Yu, and Yousaf Bin Zikria. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors*, 20 (21):6008, 2020. 2, 6

[7] John Gideon, Emily Mower Provost, and Melvin McInnis. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2359–2363. IEEE, 2016. 1

[8] Lili Guo, Shifei Ding, Longbiao Wang, and Jianwu Dang. Dstcnet: Deep spectro-temporal-channel attention network for speech emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[9] Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, and Björn Schuller. Towards temporal modelling of categorical speech emotion recognition. 2018. 1

[10] Kazi Nazmul Haque, Mohammad Abu Yousuf, and Rajib Rana. Image denoising and restoration with cnn-lstm encoder decoder with direct attention. *arXiv preprint arXiv:1801.05141*, 2018. 1

[11] Faliang Huang, Xuelong Li, Changan Yuan, Shichao Zhang, Jilian Zhang, and Shaojie Qiao. Attention-emotion-enhanced convolutional lstm for sentiment analysis. *IEEE transactions on neural networks and learning systems*, 33(9):4332–4345, 2021. 1

[12] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014. 1

[13] Dias Issa, M Fatih Demirci, and Adnan Yazici. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59: 101894, 2020. 2, 6

[14] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn Schuller. Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition. *IEEE Transactions on Affective Computing*, 14 (3):1912–1926, 2022. 1

[15] Duc Le and Emily Mower Provost. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 216–221. IEEE, 2013. 1

[16] Yuanchao Li, Tianyu Zhao, Tatsuya Kawahara, et al. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech*, pages 2803–2807, 2019. 1

[17] Zhen-Tao Liu, Meng-Ting Han, Bao-Han Wu, and Abdul Rehman. Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning. *Applied Acoustics*, 202:109178, 2023. 1

[18] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 1, 3

[19] Erfan Loweimi, Mortaza Doulaty, Jon Barker, and Thomas Hain. Long-term statistical feature extraction from speech signal and its application in emotion recognition. In *Statistical Language and Speech Processing: Third International Conference, SLSP 2015, Budapest, Hungary, November 24-26, 2015, Proceedings 3*, pages 173–184. Springer, 2015. 1

[20] Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Juan M Montero, and Fernando Fernández-Martínez. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Applied Sciences*, 12(1):327, 2021. 1

[21] GH Mohmad and Radhakrishnan Delhibabu. Speech databases, speech features and classifiers in speech emotion recognition: A review. *IEEE Access*, 2024. 1

[22] Emily Mower, Maja J Matarić, and Shrikanth Narayanan. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070, 2010. 1

[23] Abu Saleh Nasim, Rakibul Hassan Chowdory, Ashim Dey, and Annesha Das. Recognizing speech emotion based on acoustic features using machine learning. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–7, 2021. 2, 6

[24] Emily Mower Provost. Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3682–3686. IEEE, 2013. 2

[25] Ilkhomjon Pulatov, Rashid Oteniyazov, Fazliddin Makhmudov, and Young-Im Cho. Enhancing speech emotion recognition using dual feature extraction encoders. *Sensors*, 23 (14):6640, 2023. 1

[26] Thejan Rajapakshe, Rajib Rana, Sara Khalifa, Berrak Sisman, Björn W Schuller, and Carlos Busso. emodarts: Joint

optimisation of cnn & sequential neural network architectures for superior speech emotion recognition. *IEEE Access*, 2024. 1

[27] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013. 1

[28] Tushar Ranjan Sahoo and Sabyasachi Patra. Silence removal and endpoint detection of speech signal for text independent speaker identification. *International Journal of Image, Graphics and Signal Processing*, 6(6):27, 2014. 2

[29] Muhammad Sajjad, Soonil Kwon, et al. Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE access*, 8:79861–79875, 2020. 2, 6

[30] Aharon Satt, Shai Rozenberg, Ron Hoory, et al. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*, pages 1089–1093, 2017. 1

[31] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, pages II–1. Ieee, 2003. 1

[32] S. G. Shaila, A. Sindhu, L. Monish, D. Shivamma, and B. Vaishali. Speech emotion recognition using machine learning approach. In *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*, pages 592–599. Atlantis Press, 2023. 2, 6

[33] Jiaxin Wang, Hao Yin, Yiding Zhou, and Wei Xi. Advancements and challenges in speech emotion recognition: a comprehensive review. In *Fourth International Conference on Signal Processing and Machine Learning (CONF-SPML 2024)*, pages 102–109. SPIE, 2024. 1

[34] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE access*, 9:47795–47814, 2021. 1

[35] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–30, 2020. 1