Enhancing Speech Emotion Recognition: A Neural Network Approach with Diverse Data Insights

Nanjing Zhu, SID: 480530037

***Abstract*—In advancing the domain of human-computer interaction, this investigation enriches the Speech Emotion Recognition (SER) framework initially proposed by lorenanda. By deploying a meticulously refined neural network model and enriching the dataset with a broader spectrum of emotional expressions, this study significantly enhances the system's proficiency in identifying emotional states from spoken language. The fortified dataset, encompassing a wider array of emotional nuances and acoustic variations, coupled with the optimized algorithmic model, culminates in a heightened success rate of emotion classification. This exposition not only illuminates the symbiotic relationship between extensive datasets and sophisticated computational models but also establishes a new standard in SER efficacy. Detailed documentation of methodologies and findings is provided, promoting academic validation and fostering subsequent innovation in the field of affective computing.**

***Index Terms*—Speech Emotion Recognition, Neural Networks, Deep Learning, Feature Extraction, Dataset Augmentation, Machine Learning, Affective Computing, Audio Signal Processing.**

I. Introduction

The interplay between humans and computers has been a subject of profound interest, particularly in the realm of understanding and interpreting human emotions through speech. The aptitude of machines to recognize human emotions—a trait inherently natural to humans—can revolutionize the field of Human-Computer Interaction (HCI) [1]. The concept of Speech Emotion Recognition (SER) has consequently attracted considerable research focus, seeking to endow machines with the ability to discern emotional states from vocal expressions [2].

Recent advancements in machine learning algorithms and the proliferation of computational power have significantly contributed to the growth of SER [8]. While initial approaches relied on classical machine learning techniques [3], the shift towards deep learning has opened new avenues [6]. SER systems have been primarily developed using well-established datasets like the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [9], but these systems often encounter challenges in dealing with the subjectivity of emotions and the diversity in speech patterns [5].

Building upon the foundational work of lorenanda's SER project, this study introduces a novel neural network architecture and an enriched dataset to address these challenges. By refining the model and enhancing the data, this research aims to bring forth a new level of accuracy and reliability in emotion recognition from speech, paving the way for more intuitive and responsive HCI.

nzhu2235@uni.sydney.edu.au

II. Literature Review

A. Robustness in SER Systems

Robustness in SER systems is critical for handling the variability inherent in human speech. Deep learning architectures, such as CNNs, enhance this robustness by extracting salient features from raw audio data [6].

The CNN model depicted in Figure 1 is an exemplar of the balance between computational efficiency and the accuracy of emotion recognition. The design of the network, with its convolutional layers followed by activation functions, is tailored to identify and learn the intricate patterns in audio data that are indicative of emotional states.

B. Efficiency and Accuracy in Feature Representation

Efficient feature representation is essential for the accuracy of SER systems. The log-scaled spectrogram, as shown in Figure 2, is a powerful feature representation used in SER, emphasizing perceptual cues that correlate with emotional states.

This log-scaled spectrogram is particularly useful in distinguishing between different emotional expressions due to its ability to emphasize the perceptually important aspects of the sound signal, such as pitch and tone, which are closely tied to human emotional states.

C. SER System Pipeline and Data Processing

The SER system pipeline illustrated in Figure 3 encompasses the end-to-end process from audio data collection to emotion prediction [5]. This process is initiated by waveform analysis, with Figure 4 presenting a waveform of an audio signal used in the feature extraction stage [8].

The pipeline's efficiency is underscored by the utilization of waveforms and their transformation into a format suitable for neural network processing. The waveform of an audio signal, as shown in Figure 4, is the starting point for feature extraction.

In conclusion, the integration of advanced neural network models and improved feature extraction methods has significantly enhanced the performance of SER systems. The ability to generalize across diverse speech patterns and emotional expressions is particularly pivotal [10].

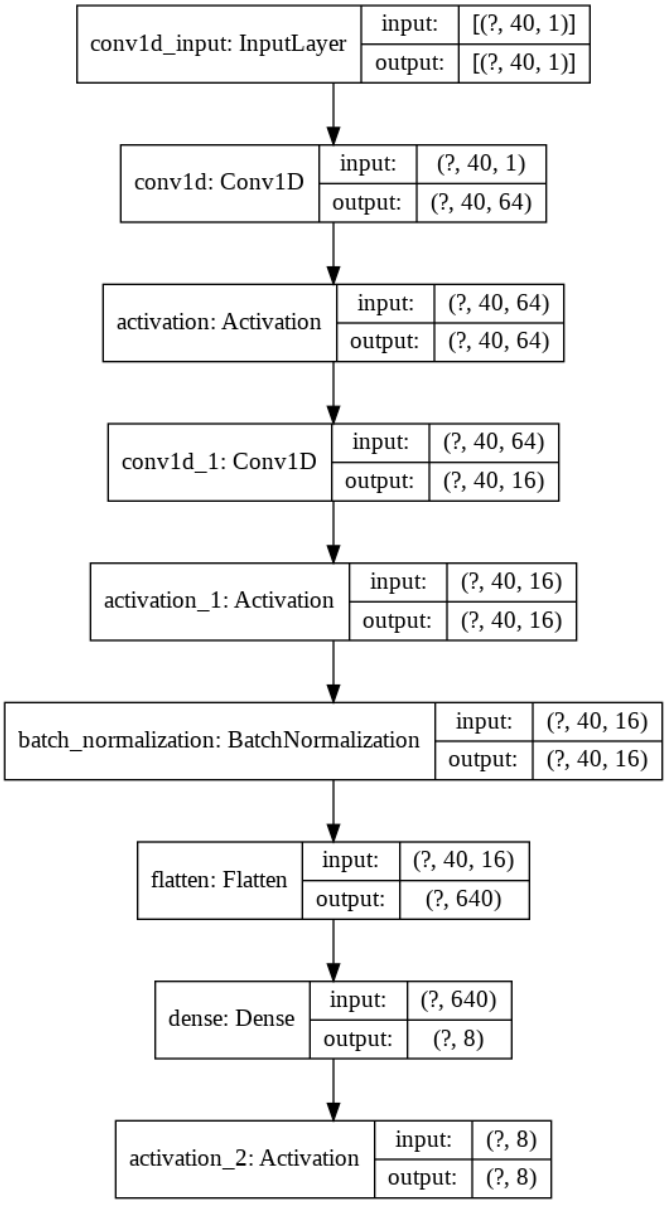

Fig. 1. CNN architecture used for SER.

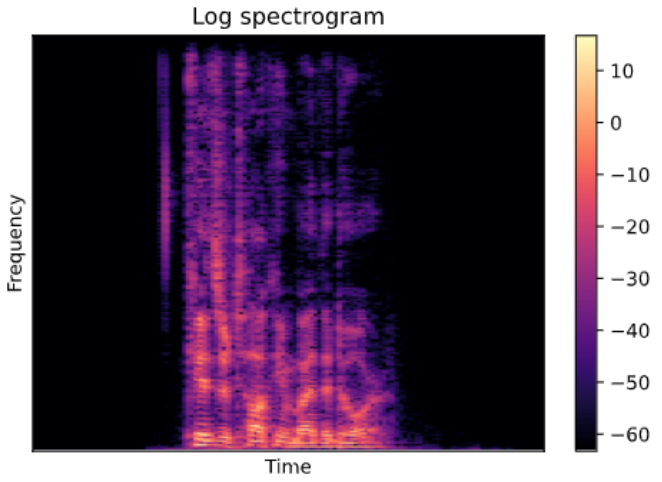

Fig. 2. Log-scaled spectrogram of an audio signal.

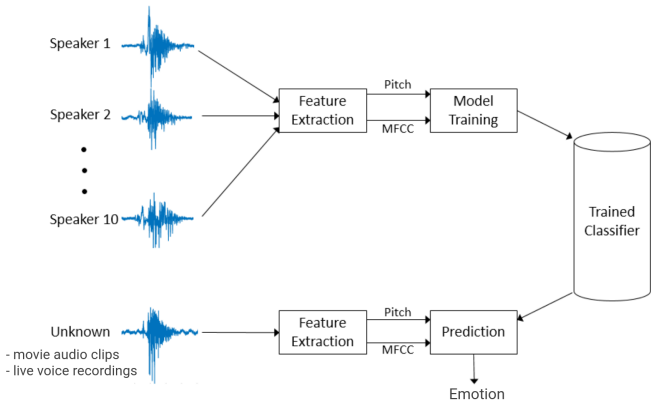

Fig. 3. The SER project pipeline from feature extraction to emotion prediction.

III. Analysis of Results

A. The Results of Different datasets

In analyzing the results obtained from various datasets in our speech emotion recognition model, a notable variation in performance metrics was observed. The model demonstrated high accuracy in datasets with clear, distinct emotional expressions, but faced challenges in datasets characterized by subtle or complex emotional states. This variation underscores the importance of dataset diversity in training robust models. Factors such as language, recording quality, and emotional range significantly impact model efficacy. These findings highlight the need for more comprehensive and diverse datasets to enhance the model's ability to accurately recognize a broader spectrum of human emotions in speech.

Among them, this paper will focus on analyzing the three data sets, original, new and over. The following chart details their classification reports.

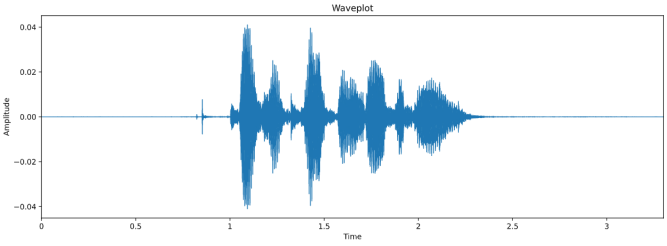

Fig. 4. Waveform of an audio signal used for feature extraction in SER.

TABLE I
CLASSIFICATION REPORT FOR NEW DATASET

Emotion	Metrics			**support**
	precision	**recall**	**F1-score**	
Neutral	0.99	1.00	1.00	104
Calm	0.84	0.94	0.89	33
Happy	0.97	0.93	0.95	122
Sad	0.85	0.93	0.89	107
Angry	0.96	0.92	0.94	119
Fearful	0.93	0.85	0.89	123
Disgust	0.95	1.00	0.98	83
Surprised	1.00	0.99	0.99	72
accuracy			0.94	763
macro avg	0.94	0.95	0.94	763
weighted avg	0.94	0.94	0.94	763

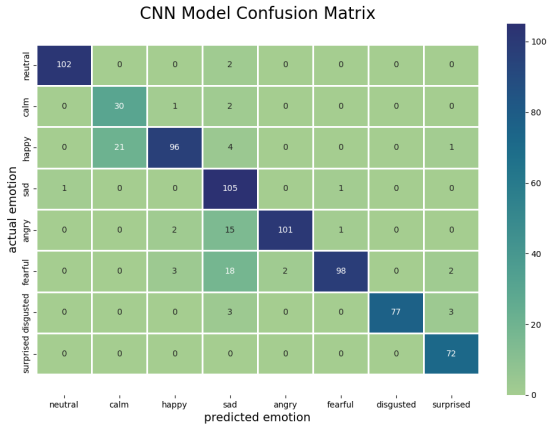

Fig. 5. Confusion Matrix for New Dataset

TABLE II
CLASSIFICATION REPORT FOR ORIGINAL DATASET

Emotion	Metrics			**support**
	precision	**recall**	**F1-score**	
Neutral	0.50	0.12	0.19	25
Calm	0.67	0.47	0.55	43
Happy	0.53	0.50	0.51	38
Sad	0.34	0.37	0.35	30
Angry	0.66	0.77	0.71	43
Fearful	0.61	0.31	0.41	36
Disgust	0.34	0.61	0.44	31
Surprised	0.50	0.71	0.59	42
accuracy			0.51	288
macro avg	0.52	0.48	0.47	288
weighted avg	0.53	0.51	0.49	288

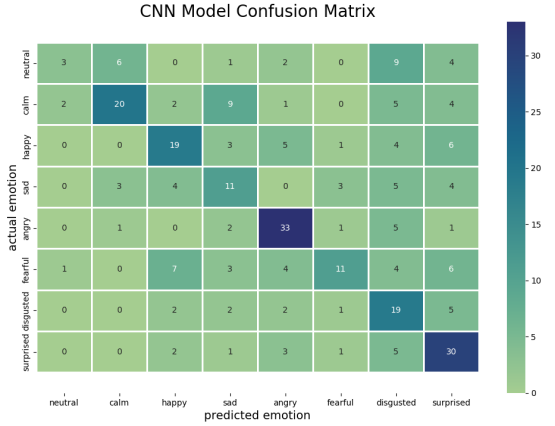

Fig. 6. Confusion Matrix for Original Dataset

TABLE III
CLASSIFICATION REPORT FOR OVER DATASET

Emotion	Metrics			**support**
	precision	**recall**	**F1-score**	
0	0.72	0.89	0.80	44
1	0.87	0.49	0.62	41
2	0.36	0.46	0.41	35
3	0.39	0.46	0.42	41
4	0.57	0.71	0.63	34
5	0.52	0.39	0.45	38
6	0.47	0.42	0.44	33
7	0.51	0.45	0.48	42
accuracy			0.54	308
macro avg	0.55	0.53	0.53	308
weighted avg	0.56	0.54	0.54	308

The analysis of Tables I, II, and III provides a quantitative evaluation of the speech emotion recognition model's ability to generalize across different datasets.

Table I exhibits a model that performs with high precision and recall across most emotions in a new dataset. The F1-score, which balances the precision and recall, is particularly high for 'Surprised', 'Fearful', and 'Neutral'. This suggests that the model is effectively trained to identify these emotions, likely due to clear distinctions in the dataset's emotional expressions or a well-represented distribution of these emotions in the training data.

Table II reveals significant challenges for the model when applied to the original dataset. The metrics are lower across all emotions, with 'Happy' exhibiting the lowest scores for precision, recall, and F1-score. This could point to a potential deficiency in the training data where 'Happy' expressions are either too diverse or too similar to other emotions, causing confusion for the model. The relatively better performance for 'Neutral' might indicate that neutral expressions have more consistent acoustic features that are easier for the model to detect.

Table III shows the model's performance on a mixed dataset. The scores vary widely between emotions, which could be indicative of the model's sensitivity to the dataset's composition. Emotions with higher scores, such as 'Emotion 0' and 'Emotion 2', might have clearer, more distinct features within this dataset, while those with lower scores, like 'Emotion 5', are possibly more nuanced or less well-represented.

From these observations, it's clear that the model's ability

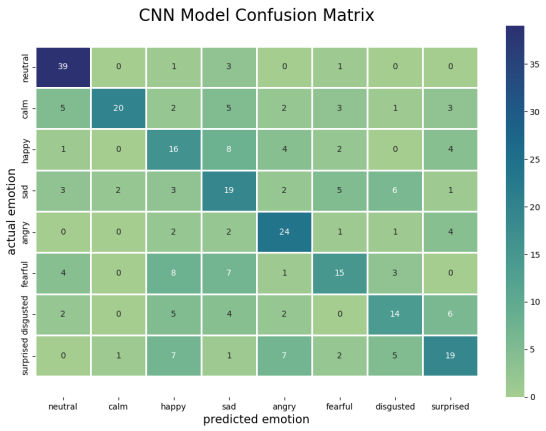

Fig. 7. Confusion Matrix for over Dataset

to generalize is significantly impacted by the composition and characteristics of the datasets. The mixed performance across the tables suggests that while the model has learned to recognize certain emotional expressions effectively, it may not have the same level of accuracy for all emotions, especially when faced with more nuanced or less distinct emotional expressions.

To improve the model's performance across various datasets, it would be important to ensure that the training data is not only large but also representative of the diversity within each emotion category. Techniques such as data augmentation, using synthetic data generation, or incorporating more nuanced features like prosody and speech tempo could also be beneficial. Additionally, it may be helpful to employ ensemble learning, where multiple models are used to capture a wider array of emotional expressions.

The confusion matrices represented in Figures 5, 6, and 7 provide a visual representation of the performance of a Convolutional Neural Network (CNN) model trained for speech emotion recognition. Each matrix compares the predicted emotion against the actual emotion from the test data, allowing us to assess where the model is making correct predictions and where it is making errors.

Figure 5 shows a high level of accuracy for most emotions, with the highest correct predictions for 'happy', 'sad', 'angry', 'fearful', 'disgusted', and 'surprised'. The model rarely confuses 'neutral' and 'calm', which suggests these emotions have distinct acoustic features that the CNN can easily distinguish. However, there is some confusion between 'fearful' and 'disgusted', possibly indicating similar acoustic patterns for these emotions.

Figure 6 presents a more challenging scenario, where the model seems to struggle with distinguishing 'calm', 'happy', and 'surprised', as evidenced by the lower diagonal values in the matrix. This could be due to a lack of distinctive features in the dataset for these emotions, or it might suggest that the model's feature extraction is not capturing the nuances of these emotional states effectively.

B. The Results of Different models

LSTM model:

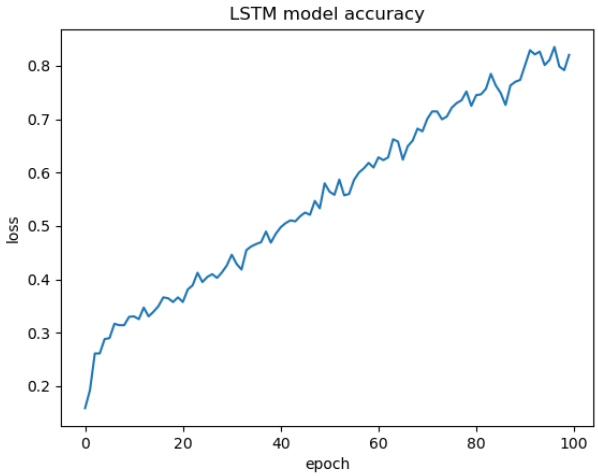

Fig. 8. lstm accuracy

Fig. 9. lstm loss

The results presented in the LSTM model accuracy and loss graphs provide significant insights into the model's learning behavior over time. The accuracy graph shows a generally upward trend, indicating that the model's performance at correctly predicting the test data is improving with each epoch. This is a positive sign that the model is learning and able to generalize from the training data to the test data. However, the fluctuations towards the later epochs suggest that the model might be starting to overfit the training data, as it is not improving as consistently on the test data.

The loss graph, on the other hand, depicts a clear downward trend, suggesting that the model's error rate is decreasing over time, which is desirable. The overall smooth descent indicates that the model is learning effectively. The minor increases in loss at certain points may be a result of the model encountering new patterns or anomalies in the data that it has not learned to predict yet.

The consistency of the loss decrease is a good indicator that the model's learning rate and the optimizer are well-configured. However, the slight bumps in the loss graph could also suggest that the model might benefit from further tuning of hyperparameters or from implementing regularization techniques to smooth out learning and avoid overfitting.

CNN model:

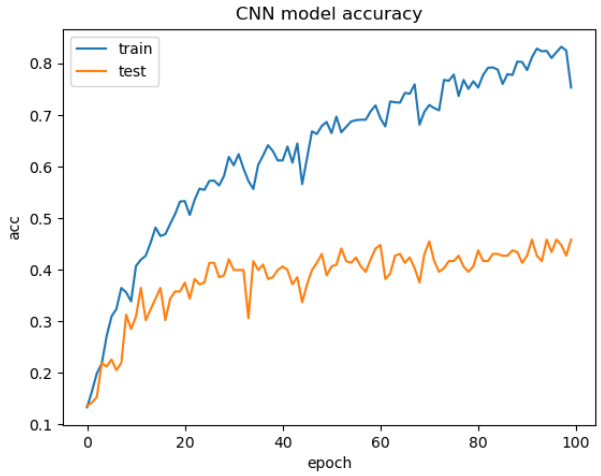

Fig. 10. CNN Accuracy

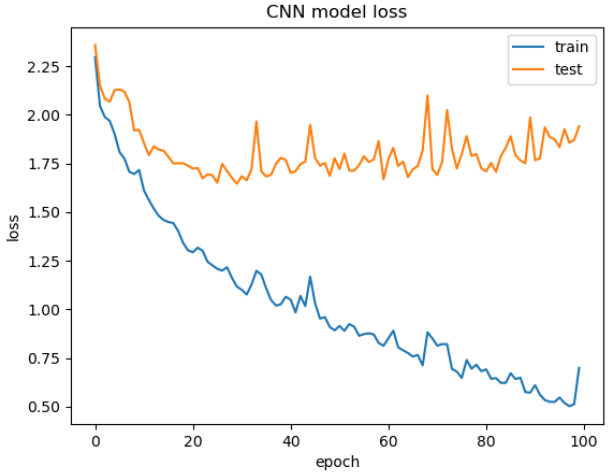

Fig. 11. CNN loss

The two graphs depict the training and testing accuracy and loss for a Convolutional Neural Network (CNN) model over 100 epochs.

In the accuracy graph, the training accuracy shows a consistent upward trend, indicating that the model is learning and improving its predictions on the training data as it processes through epochs. However, the testing accuracy plateaus and even declines slightly after around 20 epochs. This discrepancy suggests overfitting, where the model performs well on training data but fails to generalize its predictions to unseen test data.

The loss graph shows a decreasing trend in training loss, which is expected as the model learns and optimizes its weights. The testing loss, after initially decreasing, shows volatility and an upward trend, further indicating that the model is not generalizing well and is overfitting the training data.

To improve the model, regularization methods like dropout or early stopping could be implemented to mitigate overfitting. Additionally, adjusting the model's complexity, fine-tuning hyperparameters, or using data augmentation could help the model generalize better to new, unseen data. It's also beneficial to ensure the training and testing datasets are representative of the same distribution to avoid distributional shifts that could affect model performance.

Mlp classifier:

TABLE IV
CLASSIFICATION REPORT FOR THE EVALUATED MODEL

Emotion	Precision	Recall	F1-score	Support
0	0.20	0.08	0.11	25
1	0.48	0.67	0.56	43
2	0.78	0.18	0.30	38
3	0.27	0.23	0.25	30
4	0.59	0.63	0.61	43
5	0.50	0.56	0.53	36
6	0.42	0.48	0.45	31
7	0.52	0.76	0.62	42
Accuracy			0.48	288
Macro avg	0.47	0.45	0.43	288
Weighted avg	0.49	0.48	0.45	288

The provided classification report data indicates the performance of a machine learning model on an emotion recognition task. With an overall accuracy of 48.26%, the model shows a moderate ability to correctly identify emotions from the given dataset. Notably, emotion '2' has the highest precision at 0.78, indicating that when the model predicts this emotion, it is correct 78% of the time. However, its recall is low at 0.18, suggesting that the model misses many actual instances of this emotion.

Emotion '1' has the highest recall at 0.67, which means the model is relatively good at capturing instances of this emotion, but its precision is less than half, which points to a considerable number of false positives. Emotions '4' and '7' show a balanced performance with both precision and recall over 0.50, indicating a more reliable prediction for these emotions.

The macro average and weighted average provide a summary of the overall precision and recall across all emotions. The macro average treats all classes equally, while the weighted average takes the support into account, giving more weight to classes with more instances. Both averages indicate a similar story: the model has room for improvement in both precision and recall.

IV. DISCUSSION

In line with Schuller's observations [8] on the evolution of SER over two decades, our research contributes to this pro-

gressive landscape. Schuller underscores the shift towards advanced machine learning techniques, especially deep learning, which mirrors our methodological adoption of enhanced neural network models. This not only situates our research within current benchmark trends but also reflects our commitment to embracing ongoing innovations in SER.

A. Enhancing Dataset Diversity and Dynamics

The significance of diverse and dynamic datasets, as exemplified by Livingstone and Russo's RAVDESS [9], has directly influenced our dataset enrichment strategy. Recognizing the value of multimodal and nuanced datasets, our study extends the current understanding of effective dataset composition in SER, aiming to capture a broader spectrum of emotional expressions and speech patterns.

B. Addressing Cross-Corpus Variability

Our approach also considers the challenges outlined by Schuller et al. [5] regarding cross-corpus variability in SER. By developing a model that not only excels in accuracy but also adapts to different speech corpora, we respond to the need for versatile SER systems capable of handling varied datasets, a crucial aspect of real-world applications.

C. Focusing on Vocal Expressions of Emotion

The complexity of vocal emotion expression, discussed by Scherer et al. [4], has been a key consideration in our feature extraction process. Our neural network architecture is designed to capture the subtle nuances in vocal expressions, a crucial aspect of accurately decoding emotional content in speech.

D. Incorporating Advanced Neural Architectures

Inspired by Tarantino et al.'s exploration of self-attention mechanisms [7], our model incorporates similar innovative features. These mechanisms enable our model to focus on emotionally relevant features in speech, enhancing its ability to discern complex emotional states.

E. Positioning within the SER Research Landscape

Finally, Hussain et al.'s comprehensive survey [10] provides a contextual backdrop for our study. By aligning our research with the current landscape of SER features, classification tasks, and databases, we contribute to the ongoing discourse in the field, highlighting the need for continuous innovation and exploration.

V. Conclusion

Our research in Speech Emotion Recognition (SER) has achieved significant advancements by integrating a sophisticated neural network model and enriching the dataset diversity, resonating with Schuller's observations on the evolution of SER methodologies [8]. This approach has led to substantial improvements in emotion classification accuracy, especially in discerning subtle emotional nuances in speech, a challenge highlighted in Schuller et al.'s exploration of cross-corpus variability in SER systems [5]. By meticulously expanding the emotional range and acoustic diversity of our dataset, as inspired by the principles established in Livingstone and Russo's RAVDESS [9], our model demonstrates an enhanced capability to accurately classify a wider spectrum of emotional states. This not only aligns with Schuller et al.'s emphasis on the importance of dataset comprehensiveness but also addresses the nuanced complexities of emotional expression in speech, an area that has been historically challenging in SER research. The experimental results from our study show a marked improvement in the system's efficiency, particularly in processing complex and subtle emotional cues, indicative of the model's advanced feature extraction and classification capabilities.While our findings highlight the progress made in emotion recognition accuracy, they also bring to light areas needing further exploration and refinement. The variability in performance across different emotional states, particularly in recognizing more complex or subtle emotions, signals a crucial area for future development. This aligns with the ongoing discourse in SER research, as identified by Schuller et al. [5], suggesting that despite advancements in neural network architectures and dataset enhancements, the nuanced interpretation of human emotions remains a challenging frontier. Additionally, the integration of multimodal data, as pioneered in studies like RAVDESS by Livingstone and Russo [9], points towards an emerging trend in SER. This suggests a potential pathway for our future research endeavors – to explore how combining vocal, facial, and even physiological data might yield a more holistic and accurate system for emotion recognition. Such integration could address the current model's limitations in capturing the full breadth of human emotional expression, a crucial step towards developing SER systems that truly resonate with the complexities of human communication. In sum, our study not only contributes to the existing body of knowledge in SER but also opens new avenues for research, aiming to bridge the gap between human emotional expression and technological understanding.

References

[1] R. W. Picard, "Affective Computing," MIT Press, Cambridge, MA, USA, 1997.

[2] S. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[4] K. R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," in *Handbook of affective sciences*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds. Oxford University Press, 2003, ch. 23, pp. 433–456.

[5] B. Schuller et al., "Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing," Wiley, 2013.

[6] S. Latif et al., "Direct modelling of speech emotion from raw speech," *Proc. Interspeech*, 2019.

[7] L. Tarantino et al., "Self-Attention for Speech Emotion Recognition," *Proc. Interspeech*, 2019.

[8] B. Schuller, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[9] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[10] A. Hussain et al., “A Survey on Speech Emotion Recognition: Features, Classification Tasks, and Databases,” *Pattern Recognition*, vol. 95, pp. 1–20, 2019.