A Deep Learning Model for Speech Emotion Recognition on RAVDESS Dataset

Zhongliang Wei[1]*, Chang Ge[2], Chang Su[3], Ruofan Chen[4], Jing Sun[5]
The First Affiliated Hospital, Anhui University of Science and Technology, Huainan, China[1, 2, 4, 5]
School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China[1, 2, 4, 5]
School of Mechatronics Engineering, Anhui University of Science and Technology, Huainan, China[3]

***Abstract*—Speech Emotion Recognition (SER), a pivotal area in artificial intelligence, is dedicated to analyzing and interpreting emotional information in human speech. To address the challenges of capturing both local acoustic features and long-range dependencies in emotional speech, this study proposes a novel parallel neural network architecture that integrates Convolutional Neural Networks (CNNs) and Transformer encoders. To integrate the distinct feature representations captured by the two branches, a cross-attention mechanism is employed for feature-level fusion, enabling deep-level semantic interaction and enhancing the model's emotion discrimination capacity. To improve model generalization and robustness, a systematic preprocessing pipeline is constructed, including signal normalization, data segmentation, additive white Gaussian noise (AWGN) augmentation with varying SNR levels, and Mel spectrogram feature extraction. A grid search strategy is adopted to optimize key hyperparameters such as learning rate, dropout rate, and batch size. Extensive experiments conducted on the RAVDESS dataset, consisting of eight emotional categories, demonstrate that our model achieves an overall accuracy of 80.00%, surpassing existing methods such as CNN-based (71.61%), multilingual CNN (77.60%), bimodal LSTM-attention (65.42%), and unsupervised feature learning (69.06%) models. Further analyses reveal its robustness across different gender groups and emotional intensities. Such outcomes highlight the architectural soundness of our model and underscore its potential to inform subsequent developments in affective speech processing.**

***Keywords*—Speech emotion recognition; deep learning; RAVDESS dataset; multi-feature fusion**

I. Introduction

As a vital subdomain of artificial intelligence, SER aims to bridge the gap between human affective expression and machine perception. By analyzing vocal cues such as tone, pitch, and speech rhythm, SER systems can identify emotional states, thereby enhancing human-computer interaction in applications ranging from virtual assistants to mental health monitoring [1]. This technology leverages machine learning algorithms to deeply analyze speech signals, extracting critical features like pitch, rhythm, tempo, and intensity to discern the underlying emotional states. The application prospects for this technology are wide-ranging, gradually transforming services and interactions across various industries. In customer service, emotion recognition enables more personalized and empathetic interactions [2]. In healthcare, it aids in remotely monitoring patients' emotional changes and evaluating treatment effectiveness. Smart home systems adjust environmental settings based on user emotions, enhancing the overall living experience. Furthermore, it is utilized in human-computer interaction [3], assistive technologies [4], market research [5], and education [6], where the SER technology serves as a pivotal component. It assists individuals with language barriers in communication and helps analyze customer feedback to gain insights into needs. With ongoing advancements in technology, the precision and breadth of SER are anticipated to increase, allowing machines to more effectively comprehend and address the emotional requirements of humans.

Despite significant progress in both theory and application, SER technology still faces a series of challenges. Firstly, emotional expression varies significantly across different cultures, contexts, and individuals, necessitating models with high generalization capabilities. Secondly, emotional cues in speech often accompany subtle variations in tone and rhythm [7], which automated systems may struggle to accurately capture. Additionally, noise interference [8], variations in recording devices, and natural variability [9] in speech can impact the accuracy of emotion recognition. Despite these challenges, advancements in technologies such as deep learning [10] and improvements in computational power offer prospects for continued expansion in the application scope of SER. The effectiveness of CNNs in identifying local structures has made them a popular choice in tasks involving both image and speech data. Meanwhile, the Transformer architecture has demonstrated outstanding performance in sequence modeling, effectively capturing long-term dependencies in speech. On this basis, a cross-modal attention mechanism is also introduced to fuse information from different network branches, thereby enhancing the model's representational capacity and emotion recognition capability.

This study proposes a parallel architecture that integrates CNN and Transformer models, with deep feature-level fusion achieved through a Cross-Attention module. Based on the RAVDESS dataset, we designed a systematic experimental framework and employed grid search to automatically optimize key hyperparameters such as learning rate, dropout rate, and batch size. Comparative experiments were conducted against four classic SER models. The results demonstrate that the proposed model exhibits significant advantages in terms of overall accuracy, recognition performance across emotion categories, generalization to different emotion intensities, and gender robustness. The results confirm the practical utility of the developed model in SER and offer meaningful directions for subsequent studies.

*Corresponding Author.

Structurally, the study begins by introducing the background, application scenarios, challenges, and opportunities of SER in Section I. We emphasize the significance and purpose of our study. In Section II, we comprehensively review existing research and methods in the emotion recognition domain, focusing on the applications and limitations of currently available models. Section III provides detailed descriptions of our research methodology, including data preprocessing, feature selection, and model construction. Subsequently, Section IV showcases the outcomes of our experiments, offering a comparative analysis of various models' performances. Section V concludes the study by presenting major insights and proposing future development paths.

II. Related Work

Speech-based emotion classification typically involves extracting specific statistical parameters from speech signals and these parameters were then simplified, selected, and historically analyzed using classical machine learning techniques to detect emotional variations [11]. While this approach has yielded some progress in recognizing human emotions, accurately identifying emotions through speech remains challenging. During this process, researchers typically measure parameters such as fundamental frequency, short-term energy variations, resonance peak positions, and MFCCs, as these features are believed to be directly or indirectly related to emotional expression.

Partila et al. [12] introduced a method designed to pinpoint the most effective techniques and feature pairings for stress detection. They evaluated various feature sets, including MFCCs, and eight fundamental prosodic features. These feature sets were used in three different machine learning classifiers for classification tasks related to stress detection. Shajini-Majuran et al. [13] introduced a hierarchical classification technique based on MFCCs for emotion recognition. Specifically, they focused on statistical metrics derived from MFCCs and developed optimal fitting models using one-versus-one SVM for each decision node. Their study utilized two benchmark speech datasets: Danish and Berlin languages. Chenchah et al. [14] investigated methods for recognizing human emotions through speech, with particular attention to feature selection and the impact of classifiers on recognition accuracy. The research examined four distinct emotional states by analyzing audio features from emotional speech through LFCC and MFCCs. Following this, Hidden Markov Models and SVMs were utilized to categorize these extracted features, enabling the automatic recognition of emotions. Nalini et al. [15] introduced a music emotion recognition approach that integrates MFCCs with Residual Phase (RP) features. The study focused on identifying emotional categories such as anger, fear, joy, neutral state, and sadness. RP features, which originate from the excitation source, were employed to capture distinct emotional characteristics from music signals. Research indicates that RP signals complement MFCCs by capturing emotion-specific information. Independent models were built for each emotion using MFCCs and RP features, and evidence from these models was integrated at the score level for emotion recognition.

Deep learning methods are essential for SER. By constructing complex neural network models, these methods effectively extract features from speech signals and learn emotion-related patterns [16]. These models carry out feature learning autonomously, removing the requirement for manual feature extraction and, as a result, improving the precision and resilience of emotion recognition systems. Additionally, deep learning can handle large-scale datasets, further improving model generalization capabilities.

Using deep learning technique, Satt et al. [17] proposed a method directly applied to speech spectrograms for efficient emotion recognition. By combining convolutional and recurrent networks, they achieved higher accuracy than previous studies. The method also reduced prediction latency and handled non-speech background signals effectively. Harmonic modeling improved accuracy even with unknown noise. In [18], a comparative analysis was performed, and the CNN+LSTM model outperformed single CNN by 7% and single LSTM by 9%, demonstrating LSTM's effectiveness in SER. In [19], the authors used MFCCs, waveform, and spectrograms in parallel as inputs. Different CNN models were designed, and an attention mechanism improved classification results. The validation was conducted using the Berlin Emotional Database and the multimodal emotion dataset. In [20], the authors introduced a dialogue memory network based on emotional dynamics during conversation. Using GRUs, it processed prior utterances from both speakers, capturing context. Attention mechanisms selected relevant context for predicting current utterances, simulating dynamic emotional changes. This approach enhanced dialogue understanding and prediction.

III. Methodology

A. Datasets

The RAVDESS dataset was created to offer high-quality emotional expression recordings, facilitating research across multiple disciplines including neuroscience, psychology, mental health, hearing science, and computer technology [21]. This multimodal database includes facial and vocal expressions, with twenty-four trained actors participating in the recordings, evenly split between twelve males and twelve females. They delivered lexically-aligned phrases using a standard North American accent. The dataset encompasses eight distinct emotion categories, conveyed through both speech and singing. Each emotion is presented with two degrees of intensity—normal and heightened—along with a neutral expression as an additional category.

The RAVDESS provides audio data that is diverse, reliable, and valuable for studying emotional expression in sound. Researchers can utilize this resource for emotion recognition, sound processing, and human-computer interaction studies. In this context, we primarily focus on the audio portion of the database, which consists of 1440 English sentences. These sentences are constructed by having actors sequentially speak two lexically-matched phrases. The dataset demonstrates a relatively even distribution, containing approximately 190 samples for each emotional category except for the "disgust" category. The distribution of various emotion types in the speech portion of the RAVDESS dataset is shown in Fig. 1.

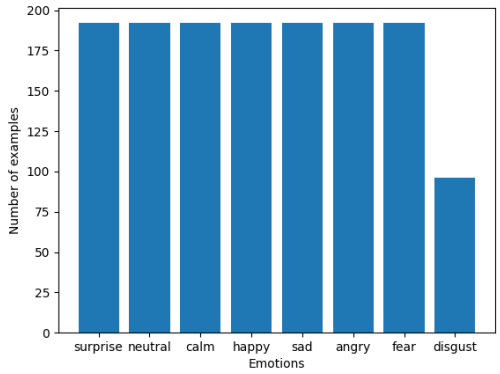

Fig. 1. Data distribution in RAVDESS.

B. Data Preprocessing

To enhance the robustness and overall efficiency of the model training procedure, a structured pipeline for preprocessing and data augmentation was systematically applied to the raw speech inputs. This workflow includes four essential stages: signal normalization, dataset partitioning, data augmentation, and feature extraction. The data preprocessing workflow is depicted in Fig. 2.

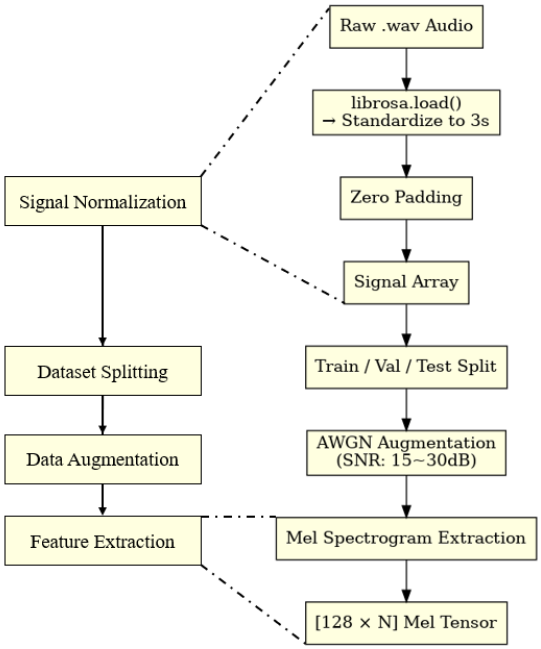

Fig. 2. Data preprocessing process.

1) Signal normalization: All original audio files were loaded using the Librosa library, with a unified sampling rate set to 48,000 Hz. A 3-second segment from the middle of each recording (starting from a 0.5-second offset) was extracted as the effective analysis region. To ensure uniform signal length in the time domain, all audio signals were padded to a fixed length L=3×48,000, with zeros added, where necessary. After this process, each speech sample was represented as a fixed-length single-channel time-domain signal.

2) Dataset splitting: To prevent model bias caused by class imbalance during dataset partitioning, the dataset is split on a per-emotion basis, assigning 80% of instances to training, 10% to validation, and the remaining 10% to testing. The indices are randomly shuffled using np.random.permutation to ensure that the samples across different subsets are independent and follow a consistent distribution.

3) Data augmentation: To improve generalization under noisy acoustic conditions, each training utterance is further expanded by synthesizing two noise-contaminated replicas. This augmentation process adds white Gaussian perturbations, where the Signal-to-Noise Ratio (SNR) is uniformly sampled from a range of 15 to 30 dB. Specifically, both the original signal and the noise are first normalized, after which a noise scaling factor is computed to achieve the target SNR. As a result, each clean utterance is transformed into a trio of instances—comprising one clean and two noise-augmented variants—thereby increasing the training set size by a factor of three. The validation and test partitions are kept intact throughout the process.

4) Feature extraction: The preprocessed time-domain signals are further converted into Mel spectrograms to more effectively capture the frequency-domain characteristics of emotional speech. To extract Mel spectrograms, the configuration employs a 1024-point FFT, a frame length of 512, a stride (hop size) of 256, and a total of 128 Mel filters. By transforming the power spectrogram into a log-scale spectrogram, the resulting 2D feature maps better align with human auditory perception.

Mel spectrograms are individually extracted for all samples in the training, validation, and test sets, and subsequently stacked to form unified datasets. The final processed Mel spectrograms are stored as 3D tensors with the following shapes:

- X_train: (sample count, 128, temporal length)
- X_val: (sample count, 128, temporal length)
- X_test: (sample count, 128, temporal length)

Since all audio clips have the same duration and processing parameters, the time dimension (i.e., number of frames) remains consistent across all samples.

This processing step provides a stable and structured feature foundation for the subsequent CNN and Transformer modules, enabling the model to accurately capture emotion-related patterns in speech signals.

C. Model Architecture

This study proposes a hybrid parallel neural network model that integrates CNN, Transformer encoders, and a Cross-Modal Attention mechanism for effective SER. As illustrated in Fig. 3, the overall architecture consists of four primary modules: a CNN branch, a Transformer branch, a fusion module based on cross-attention, and a final classification layer.

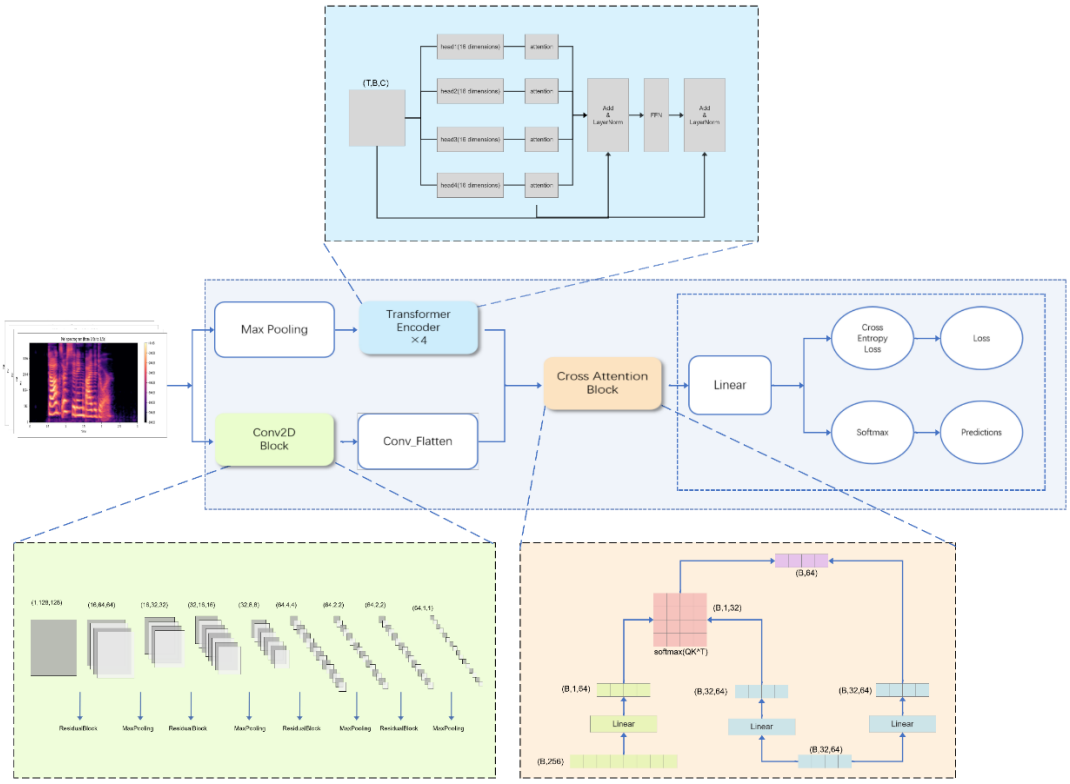

Fig. 3. Overall model architecture.

The advantages of the proposed model architecture are as follows: CNNs excel at capturing local spatial patterns, making them well-suited for extracting texture and frequency band features from Mel spectrograms. Transformers are effective in modeling long-range dependencies and temporal context, which benefits the recognition of dynamic emotional transitions. The Cross-Attention mechanism enhances the complementarity between the two modalities by introducing guided attention, thereby improving the model's overall discriminative capability.

1) CNN Branch: Local feature extraction: The input Mel spectrogram with shape (1, 128, T) is fed into a convolutional neural network composed of four Residual Blocks. Each Residual Block follows the structure [Eq. (1)]:

$$\mathbf{y} = \mathrm{ReLU}(\mathcal{F}(\mathrm{x}) + \mathrm{x}) \tag{1}$$

where, $F(x)$ denotes a pair of convolution operations, each succeeded by a normalization layer and a ReLU function. To ensure dimensional consistency between the input and output, a downsampling shortcut is introduced:

$$F(x) = BN_2(Conv_2\left(ReLu\left(BN_1(Conv_1(x))\right)\right)) \tag{2}$$

Eq. (2) defines the transformation function (x) employed in a residual unit, which comprises two sequential convolution operations. After each convolution, batch normalization (BN) is applied, followed by a ReLU activation. Specifically, the input x is first convolved using Conv_1, normalized with BN_1, and activated by ReLU. The result is then passed through a second convolutional layer Conv_2, again followed by batch normalization BN_2, but without an activation at the end. This design allows the residual block to learn complex feature transformations while maintaining training stability through batch normalization and promoting non-linearity via ReLU. The residual output (x) is combined with the shortcut path to yield the block's final representation. Subsequently, a pooling operation and a dropout layer are applied. The resulting feature map is then flattened into a global representation vector $f_{cnn} \in \mathbb{R}^D$.

2) Transformer encoder: Temporal context modeling: In order to effectively model the temporal evolution of emotional speech signals, the architecture employs a four-layer Transformer encoder. The input to this encoder is the spectrogram tensor after a 2×4 pooling operation.

Each layer in the Transformer encoder consists of the following components:

- Multi-head Self-Attention Mechanism [Eq. (3)]:

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}(\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{3}$$

- Feed-Forward Network [Eq. (4)]:

$$\mathrm{FFN}(\mathrm{x}) = \mathrm{ReLU}(\mathrm{x}W_1 + b_1)W_2 + b_2 \tag{4}$$

Here, $Q, K, V \in \mathbb{R}^{T \times d_k}$ correspond to the query, key, and value representations, where d_k defines the size of each attention vector. The attention mechanism calculates pairwise similarities between queries and keys, normalizes the scores via softmax, and uses them to compute a weighted combination of values. Each encoder layer also includes a feed-forward

network (FFN), which is a position-wise multilayer perceptron that improve the model's ability to learn non-linear temporal patterns in the emotional speech signal. This module produces an output sequence $f_{transf} \in \mathbb{R}^{T\times C}$, which is subsequently used for cross-modal fusion.

3) Cross-modal attention fusion: To effectively integrate features extracted from the CNN and Transformer branches, a Cross-Attention module is introduced. In this mechanism, the output from the CNN is used as the Query, while the Transformer output serves as the Key and Value. The computations are as follows [Eq. (5), (6), (7), (8), (9)]:

$$Q = W_Q \cdot f_{cnn} \tag{5}$$

$$K = W_K \cdot f_{\text{transf}} \tag{6}$$

$$V = W_V \cdot f_{\text{transf}} \tag{7}$$

$$\alpha = softmax\left(\frac{QK^{\top}}{\sqrt{d}}\right) \tag{8}$$

$$f_{\text{fused}} = \alpha \cdot V \tag{9}$$

In this module, $f_{cnn} \in \mathbb{R}^D$ denotes the global feature vector extracted from the CNN branch, and $f_{transf} \in \mathbb{R}^{T\times C}$ represents the temporal features obtained from the Transformer encoder. Through learnable projections W_Q, W_K, W_V, the attention mechanism computes the similarity between CNN-guided queries and the Transformer-derived keys, generating adaptive weights α for aggregating values. This facilitates effective cross-modal feature fusion by emphasizing complementary information between spatial and temporal representations.

This mechanism enables adaptive weighted aggregation, enhancing semantic consistency across modalities.

4) Classification output: Finally, the vectors f_{cnn} and f_{fused} are concatenated to form f_{all}, which is fed into a dynamically initialized linear layer for classification [Eq. (10)]:

$$\hat{y} = \text{Softmax}(\text{W} \cdot \text{f}_{\text{all}} + \text{b}) \tag{10}$$

The cross-entropy loss function is adopted for training [Eq. (11)]:

$$\mathcal{L} = -\sum_{\text{i}=1}^{\text{C}} \text{y}_\text{i} \log \hat{\text{y}}_\text{i} \tag{11}$$

The combined feature vector f_{all}, derived from merging the CNN and attention-based outputs, is passed through a newly instantiated dense layer and subsequently processed by a softmax activation to yield classification scores. To train the model, the cross-entropy loss is utilized, which measures the divergence between predicted distributions $\hat{\text{y}}_\text{i}$ and ground-truth labels y_i across all C emotion categories.

IV. Experimental Results and Analyses

A. Implementation Setup

The emotion recognition system was implemented by using Python 3.8, PyTorch2.0 and Cuda 11.8. All experiments ran on hardware configured with an NVIDIA 4090D-24 GPU, which accelerated both training and inference.

To achieve optimal performance, we conducted a grid search over critical hyperparameters, specifically the learning rate, dropout rate, and batch size. The final selected configuration, which yielded the best validation performance, is summarized in Table I.

TABLE I. Selected Hyperparameters after Grid Search

Hyperparameter	Value
Optimizer	SGD
Learning Rate	0.001
Momentum	0.8
Weight Decay	1e-3
Batch Size	16
Dropout Rate	0.4
Early Stopping	50 epochs
Max Epochs	1000

The model was trained using the SGD optimization algorithm, configured with a learning rate of 0.001, momentum coefficient of 0.8, and an L2 penalty term (weight decay) set to 1e-3. Dropout layers with a dropout probability of 0.4 were added after each residual block to reduce overfitting. Inputs to the model were standardized using StandardScaler, fitted on the training set and applied consistently across validation and test sets.

The checkpoint corresponding to the lowest validation loss was preserved and subsequently restored for final evaluation on the test partition. A dummy forward pass was performed to initialize the dynamically-sized fully connected layer before loading the pretrained weights.

B. Performance Metric

In this research, classification accuracy is employed as the core evaluation metric to assess the model's effectiveness in recognizing emotional speech. It measures how many test samples are accurately predicted out of the entire evaluation set. The metric is mathematically expressed as [Eq. (12)]:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\,(\hat{y}_i = y_i) \tag{12}$$

where, N represents the total number of evaluation samples, $\hat{y}_i$ is the predicted class label, and y_i is the true class label. The indicator function $\mathbb{I}(\cdot)$ returns 1 when the condition is satisfied and 0 otherwise.

The selection of accuracy as the principal evaluation metric is grounded in the balanced distribution of classes in the RAVDESS dataset. Since each emotional category contains approximately the same number of samples, accuracy effectively reflects the model's performance across all categories without bias towards more frequent labels. This makes it a suitable and meaningful indicator in our experimental setting.

C. Comparison Models

1) Issa et al. [22] incorporated five distinct low-level descriptors such as cepstral and harmonic representations (e.g., MFCC, chroma), and these features were integrated as

input to a convolutional neural network for classifying eight emotional states in the RAVDESS dataset.

2) Bhattacharya et al. [23] introduced a multilingual speech emotion classification approach utilizing a one-dimensional CNN architecture. Their method focused on extracting time-dependent acoustic patterns from speech to identify emotion-related cues.

3) Jin et al. [24] introduced a bimodal emotion recognition strategy that leverages both audio and facial expressions. Their architecture integrates convolutional layers for audio MFCC features and a dual-layer LSTM network for facial features, subsequently applying a multi-head attention module to enhance feature fusion. Although the method is inherently multimodal, this study only adopts the audio stream's performance on the RAVDESS dataset for comparison in the ablation analysis.

4) Smietanka et al. [25] proposed an approach that enhances feature learning through the integration of unsupervised learning and hand-crafted prosodic descriptors for SER. Their model employs time-frequency representations, specifically Mel-spectrograms and CQT-spectrograms, to capture spectral-temporal dynamics. Its performance on the RAVDESS dataset serves as a baseline in our comparative analysis.

D. Results

In this study, the proposed model—integrating Convolutional Neural Networks (CNN) with a Transformer architecture—was evaluated on the RAVDESS speech emotion dataset. Fig. 4 illustrates a comparison of classification accuracies between our model and several benchmark methods from the literature. As shown, our model achieved the highest accuracy of 80.00%, significantly outperforming the methods of Issa et al. (71.61%), Bhattacharya et al. (77.60%), Jin et al. (65.42%), and Smietanka et al. (69.06%).

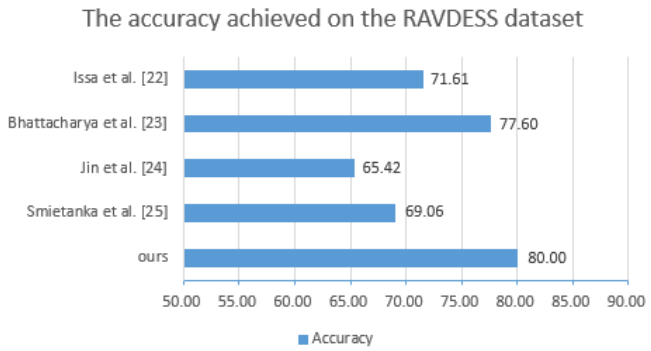

Fig. 4. Accuracy comparison.

To further evaluate the model's classification performance, Fig. 5 presents the corresponding confusion matrix. This visualization effectively reflects how the system performs across different emotional categories and offers deeper understanding into its capability to differentiate between various affective states.

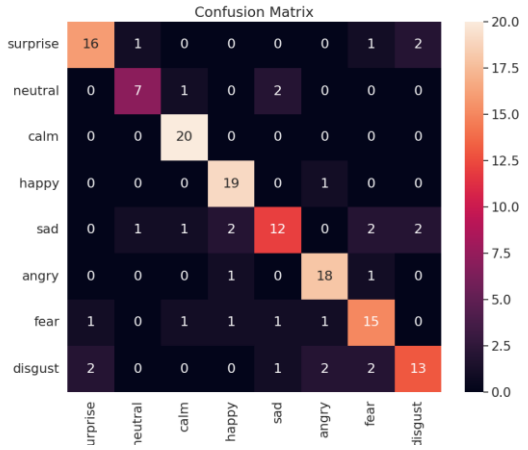

Fig. 5. Confusion matrix of emotion classification.

The classification accuracy corresponding to each emotional class, as derived from the confusion matrix, is reported in Table II. The results indicate that the model performs notably well in identifying emotions like calm, happy, and surprise. However, certain categories such as sad, fear, and disgust still exhibit a degree of misclassification.

TABLE II. CLASSIFICATION ACCURACY STATISTICS

Emotion Category	Correct Predictions	Total Samples	Accuracy (%)
surprise	16	20	80.00
neutral	7	10	70.00
calm	20	20	100.00
happy	19	20	95.00
sad	12	20	60.00
angry	18	20	90.00
fear	15	19	78.95
disgust	13	18	72.22

To investigate the effect of emotional intensity (normal versus strong) on recognition performance, the distribution of correctly and incorrectly predicted samples is visualized in Fig. 6.

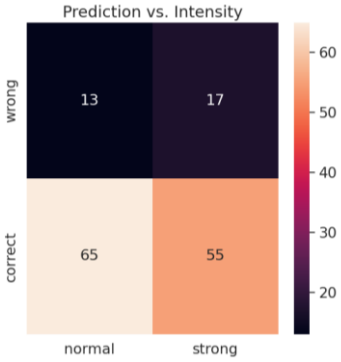

Fig. 6. Confusion matrix of intensity classification.

In addition, to evaluate the model's generalization capability across different genders, prediction outcomes for male and female speakers are separately analyzed and presented in Fig. 7.

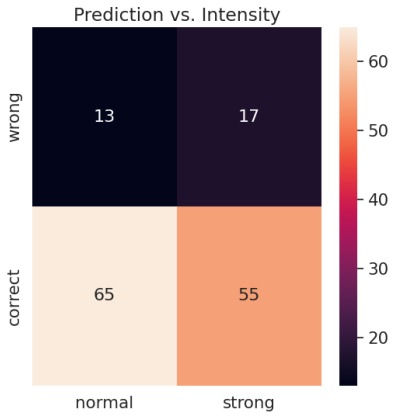

Fig. 7. Confusion matrix of gender classification.

E. Analysis

This section conducts a comprehensive examination of the model's recognition effectiveness from three perspectives: emotion classification capability, robustness to emotion intensity, and generalization across gender, based on the confusion matrix and cross-tabulation plots of emotional attributes.

1) Analysis of emotion classification capability: The confusion matrix shown in Fig. 5 clearly illustrates the model's prediction distribution across the eight emotion categories. According to the accuracy statistics in Table II:

- The model achieves the highest and most stable performance on calm, happy, and angry, with accuracy rates reaching or exceeding 90%;
- Emotions like surprise and fear also show relatively high accuracy (80.00% and 78.95%, respectively);
- However, performance drops for categories like sad and disgust, with lower accuracy (60.00% and 72.22%, respectively), and noticeable confusion—sad samples are frequently misclassified as calm or fear.

These observations indicate that the model performs well for high-energy emotions (e.g., happy, angry), but struggles with subtler, low-arousal emotions like sad and disgust. Future work could explore fine-grained emotion modeling or multi-scale feature extraction to improve discrimination for subtle emotional expressions.

2) Impact of emotion intensity on recognition performance: Fig. 6 compares the model's recognition accuracy across different emotion intensities (normal versus strong).

- It is evident that the model achieves notably higher accuracy for strong emotion samples.
- This may be attributed to stronger emotional speech containing more pronounced pitch variations and energy dynamics, making the Mel spectrogram features more distinctive and easier for the model to learn;
- In contrast, normal emotion expressions are more subdued and harder to recognize.

To enhance recognition of normal intensity emotions, future research could focus on better modeling techniques, such as emotion style transfer or sample reweighting strategies to boost sensitivity to subtle expressions.

3) Analysis of gender generalization: Fig. 7 presents the model's prediction results for male and female speakers. The analysis reveals:

- The model achieves comparable recognition accuracy for both male and female voices, showing no significant gender bias;
- This implies that the model effectively extracts gender-independent emotional cues from Mel spectrograms, demonstrating strong generalizability across different genders.

Overall, the proposed fusion model exhibits consistent and reliable performance across a wide range of speaker demographics, highlighting its potential for deployment in practical applications.

F. Discussion

The experimental findings clearly highlight the advantages of integrating CNN and Transformer architectures, particularly in capturing local acoustic features and modeling temporal dependencies, respectively. The cross-attention mechanism has effectively improved feature fusion, leading to higher emotion recognition accuracy compared to traditional and simpler architectures. However, challenges remain in accurately classifying subtle emotions, such as sadness and disgust, suggesting limitations in distinguishing nuanced emotional cues. Additionally, the significant performance gap observed between strong and normal emotional intensity indicates that the model is more sensitive to pronounced emotional expressions. Practical deployment scenarios such as healthcare and customer service could benefit significantly from this model, provided that further optimization is conducted to enhance its sensitivity to subtle emotional nuances. Future studies should explore finer-grained feature extraction and consider multi-modal data integration to address these challenges comprehensively.

V. Conclusion and Future Work

This study introduces a parallel neural network framework that integrates CNN and Transformer encoders for SER. The architecture takes advantage of CNNs for extracting local acoustic features and leverages Transformers to model temporal dependencies in speech. A cross-attention mechanism is employed to enable deep-level fusion, allowing the network to dynamically integrate information from both branches.

We utilized the RAVDESS dataset for training and evaluation, applying a standardized preprocessing pipeline involving normalization, noise augmentation, and Mel spectrogram generation. To enhance model performance, grid

search was applied for tuning key hyperparameters such as learning rate, dropout, and batch size.

Extensive experimental comparisons against four benchmark models demonstrated that our method consistently achieves higher accuracy, better per-class emotion recognition, and stronger robustness across different emotional intensities and gender categories. These outcomes substantiate the effectiveness of the proposed cross-branch fusion strategy and affirm the model's generalization potential on balanced datasets.

Looking ahead, our future work will focus on expanding the evaluation to more complex and imbalanced real-world corpora, thereby examining the model's adaptability and robustness. We also plan to investigate multi-modal strategies that combine speech with facial expressions, textual cues, or physiological indicators to further refine emotional inference. Additionally, we will explore lightweight variants optimized for real-time deployment on edge devices or mobile platforms.

ACKNOWLEDGMENT

This research work was supported in part by Medical Special Cultivation Project of Anhui University of Science and Technology (No. YZ2023H2C011), the National Natural Science Foundation of China (Grant NO. 62476005).

REFERENCES

[1] Z. Yang, Z. Li, S. Zhou, L. Zhang, S. Serikawa, "Speech emotion recognition based on multi-feature speed rate and LSTM," Neurocomputing, vol. 601, pp. 1-12, 2024.

[2] Y. Feng, L. Devillers, "End-to-End Continuous Speech Emotion Recognition in Real-life Customer Service Call Center Conversations," in Proc. 11th Int. Conf. Affective Comput. Intell. Interact. Workshops and Demos, ACIIW 2023, pp. 1-6, 2023.

[3] N. Grágeda, C. Busso, E. Alvarado, R. Mahu, N. B. Yoma, "Distant speech emotion recognition in an indoor human-robot interaction scenario," in Proc. Int. Speech Commun. Assoc., INTERSPEECH 2023, pp. 3657-3661, 2023.

[4] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask Learning From Augmented Auxiliary Data for Improving Speech Emotion Recognition," IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3164–3176, 2023.

[5] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in Proc. Int. Speech Commun. Assoc., INTERSPEECH 2018, pp. 3673–3677, Sep. 2018.

[6] Vyakaranam, T. Maul, and B. Ramayah, "A review on speech emotion recognition for late deafened educators in online education," International Journal of Speech Technology, vol. 27, no. 1, pp. 29–52, 2024.

[7] Guidi, J. Schoentgen, G. Bertschy, C. Gentili, E. P. Scilingo, and N. Vanello, "Features of vocal frequency contour and speech rhythm in bipolar disorder," Biomedical Signal Processing and Control, vol. 37, pp. 23–31, 2017.

[8] Y. Liu, H. Sun, G. Chen, Q. Wang, Z. Zhao, X. Lu, and L. Wang, "Multi-Level Knowledge Distillation for Speech Emotion Recognition in Noisy Conditions," arXiv, 2023.

[9] K. Lalonde, "Effects of natural variability in cross-modal temporal correlations on audiovisual speech recognition benefit," in Proc. Int. Speech Commun. Assoc., INTERSPEECH 2019, pp. 2260–2264, Sep. 2019.

[10] L. Yunxiang and K. Zexin, "Design of Efficient Speech Emotion Recognition Based on Multi Task Learning," IEEE Access, vol. 11, pp. 5528–5537, 2023.

[11] J. Tao, J. Chen, and Y. Li, "A Review of Speech Emotion Recognition," Signal Processing, vol. 39, no. 04, pp. 571-587, 2023.

[12] P. Partila, M. Voznak, and J. Tovarek, "Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System," The Scientific World Journal, vol. 70, 2015.

[13] S. Majuran and A. Ramanan, "A feature-driven hierarchical classification approach to emotions in speeches using SVMs," in 2017 IEEE International Conference on Industrial and Information Systems, ICIIS 2017 - Proceedings, pp. 1–5, Jan. 2018.

[14] F. Chenchah and Z. Lachiri, "Acoustic Emotion Recognition Using Linear and Nonlinear Cepstral Coefficients," International Journal of Advanced Computer Science & Applications, vol. 6, no. 11, 2015.

[15] N. J. Nalini and S. Palanivel, "Music emotion recognition: The combined evidence of MFCC and residual phase," Egyptian Informatics Journal, vol. 17, no. 1, pp. 1-10, 2016.

[16] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning — A systematic review," Intelligent Systems with Applications, vol. 20, 2023.

[17] Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in Proc. Int. Speech Commun. Assoc., INTERSPEECH 2017, pp. 1089–1093, Aug. 2017.

[18] J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," Sensors (Switzerland), vol. 21, no. 4, 2021.

[19] F. Makhmudov, A. Kutlimuratov, F. Akhmedov, M. S. Abdallah, and Y. Cho, "Modeling speech emotion recognition via attention-oriented parallel cnn encoders," Electronics, vol. 11, 2022.

[20] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 2122–2132, 2018.

[21] S. Livingstone and F. 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, vol. 13, no. 5, 2018.

[22] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, 101894, 2020.

[23] S. Bhattacharya, S. Borah, B. K. Mishra, and A. Mondal, "Emotion detection from multilingual audio using deep analysis," Multimedia Tools and Applications, vol. 81, pp. 41309–41338, 2022.

[24] Z. Jin and W. Zai, "Audiovisual emotion recognition based on bi-layer LSTM and multi-head attention mechanism on RAVDESS dataset," The Journal of Supercomputing, vol. 81, 31, 2025.

[25] L. Smietanka and T. Maka, "Enhancing embedded space with low–level features for speech emotion recognition," Applied Sciences, vol. 15, 2598, 2025.