International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Deep Learning-Based Cross-Modal Emotion Recognition for Predicting Potential Depression Patients

DiYang Xu
School of Management Science and Engineering, Anhui university of finance and economics
Email address: 20212930@aufe.edu.cn

Abstract*—*Depression is a common mental health issue. Traditional diagnostic methods for depression primarily rely on subjective psychological assessments, which often result in lower accuracy and reliability. This study proposes a deep learning-based cross-modal emotion analysis method to enhance early identification and intervention effectiveness for depression patients. The method integrates both textual and speech data modalities, employing a cascaded model to fuse cross-modal emotional features and generate comprehensive cross-modal emotion feature vectors. Results demonstrate that compared with traditional single-modality emotion analysis approaches, our cross-modal emotion analysis method can more accurately identify potential depression patients.

Keywords*—*Deep learning, Cross-modal emotion analysis, Depression Identification, Textual Data, Speech Data

I. INTRODUCTION

In recent years, with the rapid societal development and accelerated pace of life, people have faced unprecedented psychological pressures and challenges. As a prevalent psychological issue, depressive mood has garnered widespread global attention [1]. However, due to the often subtle manifestation of depressive symptoms in early stages, the optimal window for intervention is frequently missed. Therefore, developing an effective method to predict potential depression patients holds significant importance for improving early detection rates and intervention efficiency of depression.

Cross-modal emotion recognition technology offers a novel approach to address this challenge. By integrating multimodal information such as visual, speech, and textual data through deep learning models for joint processing, it enables more comprehensive and accurate acquisition of emotional state information. Particularly in cross-modal emotion recognition for depression patients, mining multiple data sources like video and speech allows deeper understanding of patients' inner emotional states, providing scientific foundations for early risk warnings of depression [2].

In deep learning-based cross-modal emotion recognition methods, single-modality data often contains limited emotional information [3]. With advancements in deep learning algorithms, researchers increasingly recognize the correlations and complementarity between emotional features across different modalities, shifting toward multimodal approaches for emotion recognition [4][5]. This study aims to explore a novel cross-modal deep learning model that enhances the accuracy and robustness of emotional state recognition in potential depression patients through comprehensive analysis of dual-modality data (speech and text) [6].

II. NETWORK STRUCTURE

This chapter elaborates on the design details of the proposed cross-modal deep learning model, including the architecture and functional roles of the Convolutional Neural Network (CNN), Long Short-Term Memory network (LSTM), and Bidirectional LSTM-CNN hybrid framework (Bi-LSTM-CNN).

A. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is one of the most widely used models in the field of deep learning. This network simulates the visual processing mechanisms of the human brain by introducing convolutional layers, enabling it to effectively extract local features from input data. It is particularly prominent in image recognition and processing tasks.

In the task of cross-modal emotion recognition, CNN is primarily used to process audio signals. For instance, by converting speech into time-frequency representations (such as Mel-spectrograms), CNN can be employed to extract acoustic features. This transformation not only preserves acoustic information but also allows the model to leverage its powerful feature extraction capabilities to identify and classify complex sound patterns, such as happiness or anxiety.

B. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of recurrent neural network that effectively addresses the issues of vanishing or exploding gradients commonly encountered in traditional recurrent neural networks when dealing with long-term dependencies. This is achieved through the introduction of gating mechanisms, allowing LSTM to maintain the continuity and accuracy of information over extended periods.

The basic structure of LSTM primarily includes: Input Gate, Forget Gate, Output Gate, and Memory unit [7]. The key formulas are as follows:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

DiYang Xu, "Deep Learning-Based Cross-Modal Emotion Recognition for Predicting Potential Depression Patients," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 10, pp. 65-68, 2025.

$$\tilde{C}_t = f_t \otimes c_{t-1} + \tanh(W_c[h_{t-1}, x_t] + b_c)$$

$$h_t = o_t \otimes \tanh(\tilde{C}_t)$$

Where:

- $x_t$represents the input to the LSTM classifier at time t
- σ denotes the Sigmoid activation function
- W and b are the weight matrices and biases for the respective gates
- $h_{t-1}$is the hidden state at time $t-1$
- i_t,f_t and o_t are the input gate, forget gate, and output gate at time t, respectively
- $\tilde{C}_t$ is the cell state at time t

In the study of cross-modal emotion recognition, LSTM can utilize historical emotional information to predict the current emotional state, which is particularly important for identifying individuals at risk of depression. It is capable of capturing the temporal characteristics of emotional fluctuations, thereby aiding in a better understanding of an individual's emotional change patterns at different points in time. For example, when analyzing speech and text data, LSTM can process both modalities simultaneously, controlling the flow of information through its internal gating mechanisms to achieve more accurate cross-modal emotion recognition.

C. *Bidirectional LSTM - Convolutional Neural Network (Bi-LSTM-CNN)*

The Bidirectional Long Short-Term Memory-Convolutional Neural Network (Bi-LSTM-CNN) is a deep learning model that integrates bidirectional temporal modeling with local feature extraction. This model addresses the challenge of heterogeneous fusion of acoustic and textual features in cross-modal emotion recognition by simultaneously processing the spectral features of speech signals and the semantic information of text sequences. It leverages the strengths of bidirectional long-term dependency capture and local feature enhancement to achieve this goal.

Feature-level fusion technology, through multimodal feature encoding and integration, enables information synergy. It performs deep feature extraction on heterogeneous modal data separately, aligns the feature spaces, and generates a fused representation vector, which is then fed into a classifier for decision inference [8].

Feature-level fusion offers significant advantages: firstly, it fully exploits the nonlinear relationships and complementary characteristics between modalities through end-to-end joint optimization; secondly, its single-model architecture provides parameter efficiency, making it particularly suitable for small-scale multimodal datasets; and thirdly, it effectively mitigates conflicts arising from the heterogeneity of raw data through feature space projection, providing highly discriminative fused features for downstream tasks [9].

Decision-level fusion adopts a multi-classifier ensemble paradigm, where feature extraction networks and classifiers are independently constructed for each modality. A specific decision aggregation function is designed to achieve collaborative inference of multimodal prediction results. The implementation framework of this method consists of three key steps: modality-specific classifier training, decision confidence calibration, and multimodal prediction result aggregation.

The prominent advantages of decision-level fusion are as follows: firstly, it enhances robustness to missing modality data through distributed modeling; secondly, its asynchronous temporal modeling capability effectively handles mismatched sampling rates across modalities; and finally, its modular architecture design provides dynamic scalability, supporting seamless integration of new modalities [10].

III. Feature Learning of Text and Speech

A. *Feature learning of text*

This paper adopts a Bi-LSTM-CNN-based network structure for text feature extraction to fully capture the contextual information and hierarchical patterns of the text.

The Bi-LSTM-CNN network structure combines the strengths of Bidirectional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN), effectively addressing the issues of vanishing or exploding gradients commonly encountered in traditional neural networks. The Bi-LSTM captures the contextual semantic information of the current word through forward and backward LSTM units, ensuring the global nature of text features. The CNN layer then extracts local features from the output of the Bi-LSTM and further reduces dimensionality through a max-pooling layer, preserving the hierarchical patterns of the text. Finally, the extracted features are processed through a fully connected layer, and the classification results are output using the softmax function. This structure not only handles long-sequence dependencies but also enhances feature representation through convolutional operations, making it suitable for various text classification tasks.

B. *Feature learning of Speech*

This paper employs a cascaded architecture of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for speech feature encoding. In the feature extraction process, the CNN module is responsible for capturing local spatial correlation features from the speech modality, while the LSTM network focuses on modeling long-term temporal dependencies. By performing temporal modeling on the entire speech signal, it effectively captures the dynamic characteristics of psychological state changes. This dual-network collaborative mechanism achieves joint optimization of spatiotemporal features, enabling the full exploitation of deep representation information closely related to depression recognition in multimodal data.

IV. Experiment and Result Analysis

A. *Dataset*

The DAIC-WOZ multimodal database (Distress Analysis Interview Corpus Wizard-of-Oz Dataset) used in this paper is a specialized corpus for psychological disorder assessment constructed by the University of Southern California Institute for Creative Technologies (USC ICT). In 2017, the

DiYang Xu, "Deep Learning-Based Cross-Modal Emotion Recognition for Predicting Potential Depression Patients," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 10, pp. 65-68, 2025.

International Audio/Visual Emotion Challenge (AVEC) [11] included depression recognition as a sub-task of the competition, and this resource library provides data support for intelligent assessment research on mental health conditions such as anxiety disorders, depression, and post-traumatic stress disorder. The research team conducted clinical interview simulation experiments with diagnosed patients and healthy control groups through a semi-automated dialogue mode, synchronously collecting multimodal data during the interviews, including audio signals, video streams, and transcribed dialogue text [12].

The dataset includes a total of 189 participant samples, comprising 56 clinically diagnosed depression patients and 133 healthy controls (102 males and 87 females). Following standard experimental protocols, the dataset is strictly divided into training, development, and test sets to classify individuals into "healthy" and "depressed" categories.

B. Experimental environment and setup

The experimental environment adopted in this study aims to ensure the stable operation and effective evaluation of the model for predicting potential depression patients through cross-modal emotion recognition. The hardware setup includes a high-performance computer equipped with an Intel Core i9 processor, 32GB RAM, and an NVIDIA RTX 4060 Ti GPU. On the software side, Python programming language combined with the PyTorch framework is used for the construction and training of deep learning models. Additionally, to handle large-scale datasets, the study leverages the powerful computational resources provided by the Google Colab platform.

Extensive preparatory work was conducted prior to the experiments to ensure accuracy and reliability. Firstly, all necessary software packages and libraries, including but not limited to TensorFlow, NumPy, and Pandas, were installed and configured, with the PyTorch version updated to the latest stable release. Secondly, the structure and parameters of the deep learning models used in the research were meticulously planned to ensure efficient operation in subsequent experiments.

In the model architecture design, the audio feature extraction module employs the Exponential Linear Unit (ELU) as the activation function for the convolutional layers, forming a cascaded structure with the Long Short-Term Memory (LSTM) network. Meanwhile, the text feature extraction module utilizes the Rectified Linear Unit (ReLU) as its activation function. This differentiated design is primarily based on the characteristic of ELU to smooth gradients in the negative value region for continuous feature processing, whereas ReLU is more suitable for handling the sparse activation properties of discrete text features. The ELU and ReLU activation functions are expressed as follows:

$$\mathrm{ELU}(x) = \begin{cases} x, & x \geq 0 \\ \alpha(e^{x} - 1), & x < 0 \end{cases}$$

$$\mathrm{ReLU}(x) = \max(0, x)$$

where α is the slope for negative values, typically set to 1.

In the single-modal experiments, the batch size was set to 30 and the number of epochs to 300; in the dual-modal fusion experiments, the batch size was set to 50 and the number of epochs to 300.

C. Experimental results and evaluation

In the evaluation of deep learning models, selecting appropriate evaluation metrics is crucial for understanding model performance. This is especially true in the field of cross-modal emotion recognition, where the unique nature of the task may render traditional evaluation metrics insufficient for comprehensively reflecting the model's performance on cross-modal data. To address this issue, this study selects the F1 score, Weighted Accuracy (WA), and Unweighted Accuracy (UA) as the evaluation metrics for the model.

WA is calculated by assigning equal weight to all samples and obtaining the overall prediction accuracy. However, when dealing with datasets where the distribution of emotion categories is imbalanced, relying solely on the WA metric can result in the category with a larger sample size dominating the evaluation outcome. To mitigate this, the UA metric is introduced to balance the recognition performance across different emotion categories: first, the independent accuracy for each emotion category is calculated separately, and then the final evaluation result is obtained by averaging these accuracies. This dual-metric design not only reflects the overall prediction performance but also avoids the evaluation bias caused by class imbalance. The specific formulas for WA and UA are as follows:

$$WA = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i}$$

$$\mathrm{Acc}_i = \frac{TP_i}{TP_i + FP_i}$$

$$UA = \frac{1}{N} \sum_{i=1}^{N} Acc_i$$

where:

- N is the emotional category
- TP_i is the correct sample size for the i-th type of emotion classification
- FP_i is the number of samples with the i-th type of emotion classification error

This approach ensures a balanced and comprehensive evaluation of the model's performance across all emotion categories.

For text-based data, this study employs two networks, Bi-LSTM and LSTM-CNN, to compare their effectiveness in emotion recognition. For audio-based data, three networks—LSTM, CNN, and LSTM-CNN—are utilized to evaluate their emotion recognition performance. Finally, the single-modal emotion recognition results are compared with the multimodal emotion recognition model constructed in this study, Bi-LSTM-CNN, which integrates both text and audio data for emotion recognition. The comparative results are presented in

DiYang Xu, "Deep Learning-Based Cross-Modal Emotion Recognition for Predicting Potential Depression Patients," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 7, Issue 10, pp. 65-68, 2025.

Table 1, and the confusion matrix of the model is illustrated in Figure 1.

TABLE I. Comparison of the effectiveness of single modal emotion recognition and multimodal emotion recognition

Model	F1	WA	UA
Bi-LSTM(text)	61.7	57.42	56.81
LSTM-CNN(text)	64.2	61.85	61.06
LSTM(speech)	53.9	50.76	52.05
CNN(speech)	56.3	52.17	52.98
LSTM-CNN(speech)	62.8	59.53	58.2
Bi-LSTM-CNN(text+speech)	69.2	69.38	70.71

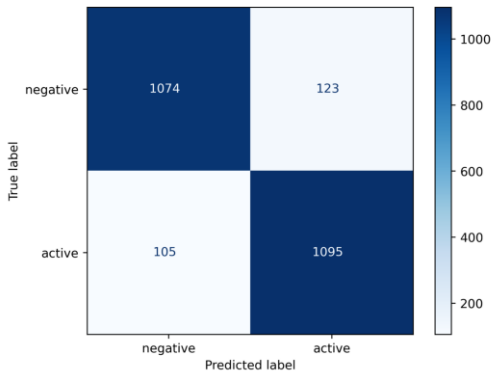

Fig. 1. Bi-LSTM-CNN multimodal emotion recognition model confusion matrix

From the comparison in Table 1, it can be observed that in the single-modal scenario, for the text modality, LSTM-CNN (F1=64.2) outperforms the pure Bi-LSTM (F1=61.7), indicating that the local feature extraction capability of CNN complements the sequence modeling ability of LSTM. In the audio modality, LSTM-CNN (F1=62.8) significantly surpasses both the standalone LSTM (53.9) and CNN (56.3), validating the effectiveness of the hybrid architecture in handling complex temporal signals (such as speech spectrograms and temporal dynamics). Furthermore, the performance of the multimodal model shows a notable improvement. Compared to other single-modal models, the Bi-LSTM-CNN multimodal emotion recognition model demonstrates significantly enhanced classification capabilities.

As shown in Figure 1, the confusion matrix of the Bi-LSTM-CNN multimodal emotion recognition model reveals that the probabilities of correctly predicting statements labeled as "depressed" and "healthy" are both above 80%. This indicates that the Bi-LSTM-CNN bimodal emotion recognition model can effectively achieve cross-modal feature complementarity, thereby exhibiting strong classification performance.

V. Conclusions

In this paper, we propose a deep learning-based cross-modal emotion recognition method aimed at predicting potential depression patients. By constructing a comprehensive model framework, including an input layer, Convolutional Neural Network (CNN), and Long Short-Term Memory Network (LSTM), this study successfully achieves efficient and accurate recognition and analysis of emotional states from different modalities (text and speech). Through a series of experimental validations, our model demonstrates strong performance across evaluation metrics. Notably, in comparative experiments, the Bi-LSTM-CNN bimodal emotion recognition model effectively achieves cross-modal feature complementarity, significantly enhancing its predictive capabilities compared to single-modal emotion recognition. This holds significant importance for early identification and intervention, as the model can comprehensively integrate various types of information when using multimodal data, thereby improving diagnostic accuracy and reliability.

In this study, the bimodal approach focuses on speech and text. Future work could explore incorporating non-verbal information, such as images and videos, to combine more emotional feature information for comprehensive and accurate predictions, further enhancing the model's precision.

Acknowledgments

Fund Project: 2023 Undergraduate Scientific Research and Innovation Fund Project of Anhui University of Finance and Economics: Research on Deep Learning-Based Cross-Modal Emotion Recognition for Predicting Potential Depression Patients(202310378168).

References

[1] Liu Hao. Research on Early Depression Recognition Based on Deep Learning [D]. Taiyuan Normal University, 2023.
[2] Hao Wei, Lu Lin. Psychiatry [M]. Beijing: People's Medical Publishing House, 2018: 78.
[3] Cohan A, Young S, Goharian N. Triaging mental health forum posts[C]//Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. 2016: 143-147
[4] LEE C W, SONG K Y, JEONG J, et al.Convolutional attention networks for multimodal emotion recognition from speech and text data[C]//Grand Challenge and Workshop on Human Multimodal Language, 2018: 28-24.
[5] GU Y, CHEN S, MARSIC I.Deep multimodal learning for emotion recognition in spoken language[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.
[6] Xia Yujing. Research on Depression Recognition Model Based on Multimodal Fusion and Few-Shot Learning [D]. Yunnan Normal University, 2023.
[7] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.
[8] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2):423-443.
[9] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 2539-2544.
[10] Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends[J]. IEEE signal processing magazine, 2017, 34(6): 96-108.
[11] Ringeval F, Schuller B, Valstar M, et al. Avec 2017: Real-life depression, and affect recognition workshop and challenge[C]//Proceedings of the 7th annual workshop on audio/visual emotion challenge. 2017: 3-9.
[12] DeVault D, Artstein R, Benn G, et al. SimSensei Kiosk: A virtual human interviewer for healthcare decision support[C]//Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. 2014: 1061-1068.