Research on the Recognition of Psychological Emotions in Adults Using Multimodal Fusion

Hui Zhao
School of Education, Science and Music, Luoyang Institute of Science and Technology, Luoyang, Henan 471027, China
Email: zhaoh1977@outlook.com

Keywords: multimodal fusion, adult, psychological emotion, depression, speech, gait

Received: March 8, 2024

Recognition of psychological emotions in adults plays a pivotal role in various fields. In this study, 110 healthy people and 110 depressed people were recruited. Gait, facial, and speech feature data were collected and screened. The obtained features were utilized as multimodal features. Subsequently, attention-bidirectional long short-term memory (BiLSTM) was developed as a method for recognizing adult psychological emotions. It was observed that better recognition results were achieved through multimodal fusion than unimodal approaches. The accuracy of the attention-BiLSTM method was 0.8182, and the F1 value was 0.8125, demonstrating higher recognition accuracy compared to methods such as the recurrent neural network and LSTM. These results verify the reliability of multimodal fusion and the attention-BiLSTM method fo recognizing psychological emotions in adults, and they can be applied in practice.

Povzetek: Raziskava uporablja multimodalno zlivanje z metodo attention-BiLSTM za prepoznavanje psiholoških čustev odraslih z analizo hoje, obraznih potez in govora.

1 Introduction

Emotions reflect the physiological and psychological states resulting from an individual's thoughts, behaviors, etc., and are related to their needs and desires. Human daily communication activities rely heavily on the transmission of emotions. Healthy psychological emotions are crucial for normal life activities, yet in a fast-paced and evolving society, adults often grapple with complex employment, emotional, and interpersonal communication pressures, leading to psychological disturbances. Depression, characterized by persistent low mood, can escalate to self-injury and suicidal behavior [1]. Early detection and intervention are especially critical in such cases. Presently, depression diagnosis heavily relies on the subjective judgment of professional doctors through methods like interviews and questionnaires. However, these approaches are inefficient, and some patients resist medical treatment. The advent of artificial intelligence has introduced new avenues for recognizing psychological conditions like depression [2]. The analysis of related works is shown in Table 1. Table 1 shows that the current research mostly focuses on utilizing a single indicator for recognizing psychological emotions, such as text, facial expressions, and speech. There is relatively less research conducted on studying two or more indicators simultaneously. In this paper, multimodal fusion was proposed for depression recognition based on multiple indicators, integrating gait, facial, and speech feature data for adult psychological emotion recognition. The proposed method was validated through experiments on real data. This paper provides a novel and reliable approach to emotion recognition research and further affirms the effectiveness of multimodal fusion.

Table 1: A summary of related works

	Key indicator	Recognition method	Result
Wani et al. [3]	Text social data	Support vector machine	It achieved better results than existing techniques on the stance sentiment emotion corpus and Aman datasets.
Hossain et al. [4]	Speech and visual signals	A convolutional neural network (CNN) and an extreme learning machine	The method was evaluated based on three datasets and achieved success.
Alamgir et al. [5]	Facial features	A bi-directional Elman neural network	It achieved 98.57% and 98.75% accuracy on the JAFFE and CK+ datasets.
Tsouvalas et al. [6]	Speech features	Federated learning	The approach improved the recognition rate by 8.67% on average using only 10% labeled data in the experiment on IEMOCAP.

2 Analysis of psychological emotion features in adults

2.1 Research subjects

To obtain multimodal characteristics, 220 subjects were recruited as research subjects for the experiment. Among them, the subjects in the depressed group were all from a mental health center in Luoyang, and the subjects in the healthy group were recruited from the society. All subjects understood the purpose and process of the study. Approval was obtained from the subjects' guardiana, and they signed the informed consent form. The collected experimental data was only used for the research in this paper, and the privacy rights and image rights of the subjects were strictly protected. The screening conditions for subjects are as follows.

(1) Depression group: depression diagnosed by a specialized psychiatrist with a patient health questionnaire-9 (PHQ-9) score ≥ 10.

(2) Healthy group: no history of psychiatric illness and PHQ-9 score < 5.

For the depression group, the exclusion conditions are as follows.

(1) Persons with extreme suicidal tendencies or co-occurring mental disorders.

(2) Pregnant or lactating women.

(3) Persons with visual, motor, or cognitive impairments.

(4) Those with serious cardiovascular diseases.

(5) Those with severe alcohol or drug dependence.

For the healthy group, the exclusion conditions are as follows.

(1) Persons with mental disorders within two generations and three lines.

(2) Pregnant or lactating women.

(3) Persons with visual, motor, or cognitive impairments.

(4) Those with serious cardiovascular diseases.

(5) Those with severe alcohol or drug dependence.

The subjects are summarized in Table 2.

Table 2: Basic information of the subjects

	Depression group (n=110)	Healthy group (n=110)
Number of males	47	51
Number of females	61	57
Age range	20-50	20-50
Average age	27.25	26.88
Average PHQ-9 score	13.21	2.03

2.2 Gait data collection and analysis

A study has demonstrated differences in gait between individuals with depression and those without [7]. Gait data can be collected without contact, minimizing interference with human psychological emotions as much as possible. Gait cannot be faked, making it more reliable in recognizing psychological emotions. This study utilizes a Kinect camera for gait data acquisition, enabling the capture of the 3D coordinates of 25 skeletal joint points of the human body, as illustrated in Table 3.

Table 3: Kinect-generated joint points

0.Spine base	7.Hand left	14.Ankle left	21.Hand tip left
1.Spine mid	8.Shoulder right	15.Foot left	22.Thumb left
2.Neck	9.Elbow right	16.Hip right	23.Hand tip right
3.Head	10.Wrist right	17.Knee right	24.Thumb right
4.Shoulder left	11.Hand right	18.Ankle right	
5.Elbow left	12.Hip left	19.Foot right	
6.Wrist left	13.Knee left	20.Spine shoulder	

Gait data were collected by arranging the subjects to walk in an open room with a range of 6×1 meters for about 2 minutes. The walking process was captured using a Kinect camera. The collected data were segmented, and only the part facing the Kinect camera was retained. At least one complete gait cycle was intercepted. The Kinect coordinate system was transformed using 0. Spine base as the origin of the coordinate system. The coordinates of the i-th joint point at the t-th frame were transformed to:

$$\begin{cases} x_i^{t\prime} = x_i^t - x_0^t \\ y_i^{t\prime} = y_i^t - y_0^t. \\ z_i^{t\prime} = z_i^t - z_0^t \end{cases}$$

To reduce the effect of body size differences, the joint data were normalized using the distance between 0. Spine base and 16. Hip right as the height of the body. The coordinates of the i-th joint point at the t-th frame were transformed to:

$$\begin{cases} x_i^{t\prime} = \frac{x_i^t}{|x_{16}^t - x_0^t|} \\ y_i^{t\prime} = \frac{y_i^t}{|y_{16}^t - y_0^t|}. \\ z_i^{t\prime} = \frac{z_i^t}{|z_{16}^t - z_0^t|} \end{cases}$$

For the processed gait data, the following features were selected in conjunction with the current study.

Step Speed: The change in 3. Head along the Z-axis coordinates was used to calculate the individual step speed.

Stride length: The distance traveled in a straight line along the Z-axis between two consecutive landings of the ipsilateral heel was used to calculate the individual stride length.

Step width: The distance between the feet along the X-axis when the feet are supported on the ground is used to calculate the individual step width.

Swing arm amplitude: The gap between the maximum distance along the Z-axis when the left-hand and right-hand swing in one gait cycle was used to calculate the swing arm amplitude.

Head tilt angle: The angle between the line between 2. Neck and 20. Spine shoulder and the vertical direction of the body was used as the head tilt angle.

The amplitude of joint motion: The maximum angular difference between shoulder-elbow-hip-knee movements in the radial plane during a gait cycle was used to calculate the amplitude of joint motion.

The collected data was statistically analyzed, and the results are displayed in Table 4.

Table 4: Statistical analysis of gait characteristics

Feature	Depression group	Health group	p value
Step speed (m/s)	1.25±0.13	1.41±0.05	**0.000**
Step length/m	0.99±0.21	1.12±0.16	**0.000**
Step width/mm	146.21±26.77	143.25±23.58	0.256
Swing arm range/mm	335.26±115.24	426.25±87.26	**0.000**
Head tilt angle/°	1.33±2.16	-0.25±2.77	**0.007**
The amplitude of shoulder joint motion/°	30.55±9.87	35.12±5.67	**0.032**
The amplitude of elbow joint motion/°	55.21±12.26	63.61±12.26	**0.017**
The amplitude of hip joint motion/°	45.77±7.35	46.58±8.25	0.352
The amplitude of knee joint motion/°	49.27±7.15	50.33±8.25	0.413

As per Table 4, the comparison of step width and the amplitudes of hip and knee joint motion did not reveal significant differences between the depression and healthy groups. Therefore, these three features were excluded, and the subsequent identification process focused on the remaining six features.

2.3 Facial data collection and analysis

Facial and speech data acquisition was conducted by professional psychiatrists. They scored the subjects using the PHQ-9 scale [8] through interviews. The scoring process was recorded using a Canon 700D camera. The length of facial data for different subjects was aligned through frame extraction processing, ensuring a uniform video length of 5 minutes. Facial feature extraction was performed using the open-source tool Openface [9], which extracted facial action units (AUs) as features. Openface provided a total of 17 AUs, as illustrated in Table 5.

Table 5: AU and corresponding facial movements

AU01	Lift the inner corner of the eyebrow	AU09	Wrinkle the nose	AU20	Pull up the corners of the mouth
AU02	Lift the outer corner of the eyebrow	AU10	Lift the upper lip	AU23	Tighten lips
AU04	Gather eyebrows and press them down	AU12	Pull the corners of the mouth to tilt upward	AU25	Separate lips
AU05	Lift upper eyelid	AU14	Tightening cheek muscle	AU26	Suck in lips
AU06	Lift cheek	AU15	Pull down the corners of the mouth	AU45	Blink
AU07	Tightening eyelid	AU17	Lift chin		

Openface employed the detected AU intensity to reflect the activity intensity of the AU, where a larger value indicates more intense movement. The data collected were tabulated in Table 6.

Table 6: Statistical analysis of facial features

	Depression group	Healthy group	p value
AU01	0.18±0.01	0.19±0.01	0.121
AU02	0.15±0.02	0.14±0.01	0.252
AU04	0.42±0.02	0.41±0.01	0.215
AU05	0.15±0.01	0.13±0.01	0.236
AU06	0.22±0.11	0.77±0.12	**0.000**
AU07	0.52±0.21	0.79±0.22	**0.012**
AU09	0.15±0.01	0.13±0.02	0.325
AU10	0.41±0.22	0.83±0.21	**0.000**
AU12	0.27±0.21	0.88±0.33	**0.000**
AU14	0.01±0.11	0.55±0.22	**0.000**

AU15	0.23±0.01	0.22±0.01	0.521
AU17	0.57±0.07	0.55±0.08	0.214
AU20	0.16±0.01	0.15±0.01	0.362
AU23	0.18±0.01	0.16±0.01	0.412
AU25	0.58±0.12	0.61±0.08	0.251
AU26	0.51±0.02	0.48±0.03	0.236
AU45	0.28±0.01	0.25±0.01	0.274

Upon examining Table 5, significant differences between the depression and healthy groups were observed for AU06, AU07, AU10, AU12, and AU14 ($p < 0.05$). In contrast, the differences in the remaining AUs were small. Consequently, these features were excluded, and only the five AUs with significant differences were considered for further analysis.

2.4 Speech data collection and analysis

The speech data was captured through a Roland R-26 portable recorder with a sampling frequency of 44.1kHz, dual-channel, and 16-bit. The recorded audio was labeled and then converted to mono through a cool edit. Segments containing interferences, such as long pauses and coughs, were cut, and then the remaining video was pre-processed. First, the signal passed through a pre-emphasis filter: $H(z) = 1 - \mu z^{-1}$, where μ is the pre-emphasis coefficient, which was taken as 0.97. Then, the obtained signal was processed by sub-frame windowing, using a Hamming window with a window length of 25 ms and a window shift of 10 ms. Finally, the Mel-frequency cepstral coefficients (MFCC) were extracted as speech features by the open SMILE toolkit [10], which includes:

(1) 12 MFCC parameters and logarithmic energy parameters,

(2) 13 first-order differences of MFCCs,

(3) 13 second-order differences of MFCCs.

The final number of speech features obtained was 39.

3 Recognition method based on multimodal fusion

Multimodal fusion involves combining inputs from various modalities to enhance model recognition accuracy. This study analyzed a total of three modal features—gait, face, and speech—in the recognition of adult psychological emotions. Specifically, they included six gait features, five facial features, and 39 speech features, which were joined to obtain 50-dimensional features. To extract deeper features from these features, the paper chose the bidirectional long short-term memory (BiLSTM) neural network method in deep learning for the recognition process.

LSTM has a good improvement on the gradient vanishing and explosion problem that exists in a recurrent neural network (RNN) [11] and has excellent performance in the processing of signals, text, etc. [12]. LSTM achieves the selective forgetting of information through forgetting gate f_t:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right),$$

where h_{t-1} is the previous unit output and x_t denotes the current unit input.

Input gate i_t achieves the selective recording of information, which is expressed as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),$$

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

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t,$$

where $\tilde{C}_t$ represents the vector of candidate values and C_t denotes the memory cell status.

Output layer o_t and C_t determines the final output h_t of LSTM together. The process is written as:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),$$

$$h_t = o_t \cdot \tanh(C_t),$$

where W and b are the weight and bias of each gate, σ is the sigmoid function, and $\tanh$ refers to the tanh function.

BiLSTM is able to learn bidirectional features, including a forward LSTM and an inverse LSTM. At moment t, the output h_t of BiLSTM includes forward $\overrightarrow{h_t}$ and reverse $\overleftarrow{h_t}$: $h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]$.

After learning advanced features through BiLSTM, to learn more valuable features, this paper combines the attention mechanism [13] based on BiLSTM, and the calculation process is:

$$Q_h = \overrightarrow{h_t} + \overleftarrow{h_t},$$

$$K = W_k \cdot Q_h,$$

$$V = W_h \cdot Q_h,$$

$$Attention(Q_h, H) = \sum_{i=1}^{n} softmax(Q_h \cdot K_i^T) \cdot V_i,$$

where Q_h stands for the query vector of the BiLSTM hidden layer, K is a key-value vector, V is a value vector, W_k and W_h are random matrices, H is the vector sequence of the BiLSTM hidden layer, and K_i^T is the transpose of the i-th key-value vector.

Eventually, the output of the attention-BiLSTM method is passed through the linear layer, and then the probability distribution, i.e., the final adult psychological emotional recognition results, is obtained through the Softmax function. The process is:

$$\hat{p}(y|S) = softmax(h \cdot W_h + b_h),$$

$$\hat{y} = arg\max_{y} \hat{p}(y|S),$$

where $\hat{p}(y|S)$ is the probability distribution of the labels of different classes in the set of class y and $\hat{y}$ is the predicted label.

The attention-BiLSTM-based multimodal fusion method is presented in Figure 1.

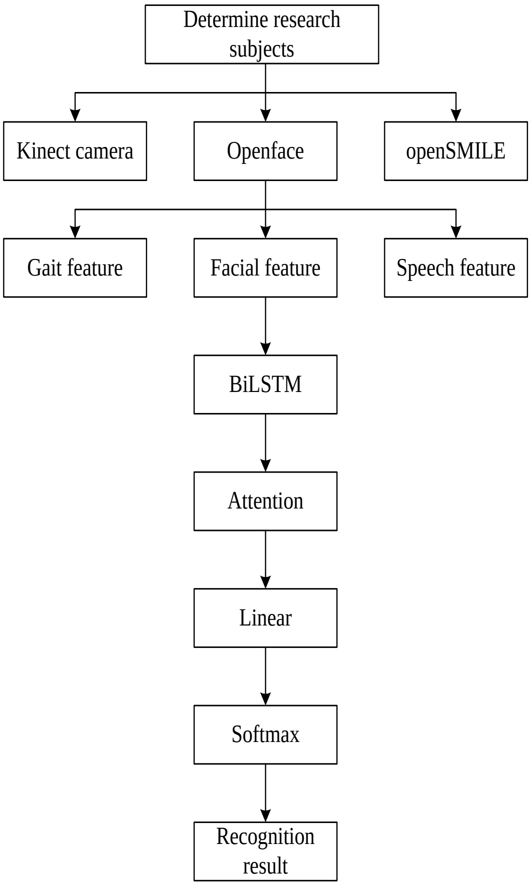

Figure 1: The attention-BiLSTM-based multimodal fusion method

As shown in Figure 1, when recognizing the psychological emotions of adults, the gait features, facial features, and speech features extracted previously were fused as inputs to BiLSTM. Then, important information was obtained through an attention layer, and finally the recognition result was obtained by output mapping in the linear and softmax layers.

4 Results and analysis

4.1 Experimental setup

The training set and test set were allocated in a ratio of 7:3. The experiment was conducted on a Linux operating system using the Python programming language and the TensorFlow deep learning framework. During training, the attention-BiLSTM model adopted a batch size of 4, a learning rate of 0.001, the Adam optimization algorithm, a dropout rate of 0.1, and a number of epochs of 100. The results were evaluated based on the confusion matrix, as outlined in Table 7.

Table 7: Confusion matrix

		Recognition label	
		Healthy	Depression
True label	Healthy	TN	FP
	Depression	FN	TP

The following indicators were used to assess the effectiveness of the attention-BiLSTM approach on the recognition of psychological emotions in adults:

(1) $Accuracy = (TP + TN)/(TP + FP + FN + TN)$,

(2) $Precision = TP/(TP + FP)$,

(3) $Recall = TP/(TP + FN)$,

(4) $F1 = (2 \times P \times R)/(P + R)$.

4.2 Analysis of results

The recognition performance of the proposed method was evaluated using a binary dataset from the UCI dataset: the heart disease dataset (https://archive.ics.uci.edu/dataset/45/heart+disease) from the UCI dataset. This dataset consisted of 13 attributes and contained a total of 3,205 data. This method was compared with the other basic recognition approaches (Table 8).

Table 8: Comparison of recognition using the heart disease dataset

	Accuracy	Precision	Recall rate	F1 value
Logic regression (LR)	0.9573	0.9722	0.8974	0.9333
Naive Bayes (NB)	0.9715	0.9561	0.9715	0.9561
Support vector machine	0.9630	0.9829	0.9127	0.9465
Random forest	0.9744	0.9621	0.9695	0.9658
Attention-BiLSTM	0.9846	0.9736	0.9855	0.9795

Table 8 shows that compared to LR, NB, and other methods, the recognition performance of the attention-BiLSTM model was superior. The accuracy on the heart disease dataset reached 0.9846, the precision was 0.9736, the recall rate was 0.9855, and the F1 value was 0.9795, all of which were the highest. This result demonstrated the excellent recognition capability of the attention-BiLSTM method.

The impact of multimodal fusion on the recognition of psychological emotions in adults was analyzed using the same recognition method. T-tests were conducted on the results of other modalities and the results of multimodal fusion (gait + facial expression + speech) proposed in this paper using SPSS 22.0 software. The significance level was 0.05. The comparison across different modalities is presented in Table 9.

Table 9: Comparison of recognition using different modes

	Accuracy	Precision	Recall rate	F1
Gait	0.7273*	0.7143*	0.7576*	0.7353*
Face	0.7424*	0.7500*	0.7273*	0.7385*
Speech	0.7121*	0.7059*	0.7273*	0.7164*
Gait + face	0.7727*	0.8000*	0.7273*	0.7619*
Gait + speech	0.7576*	0.7429*	0.7879*	0.7647*
Face + speech	0.7576*	0.7742*	0.7273*	0.7500*
Gait + face + speech	0.8182	0.8387	0.7879	0.8125

Note: * indicates the difference is statistically significant.

Table 9 shows that in unimodal recognition, facial features exhibited relatively better performance. In terms of accuracy, facial features showed a 1.51% improvement compared to gait and a 3.03% improvement compared to speech. Similarly, in terms of the F1 value, facial features demonstrated a 0.32% improvement compared to gait and a 2.21% improvement compared to speech. These outcomes suggested that among gait, facial, and speech features, facial features contained information that yielded better results in recognition of psychological emotions in adults and could effectively distinguish between depressed and healthy individuals.

In multimodal recognition, the pairwise combinations of different modalities demonstrated an increase in accuracy compared to when the modalities were individually considered: 0.7727 for gait+face, 0.7576 for gait+speech, and 0.7576 for face+speech. Ultimately, fusing the three modalities of gait, face, and speech resulted in an accuracy of 0.8182, showing an improvement of 4.55% to 6.06% compared to pairwise combinations of modalities. The F1 value reached 0.8125, reflecting a 4.78%-6.25% improvement compared to the modalities combined two by two. It can be observed from the statistical analysis results that there was a significant difference between the recognition results obtained from the multimodal fusion (gait + facial + speech) and those of other modalities. These outcomes suggested that when three modalities were integrated for adult psychological emotion recognition, the information contained in them was more comprehensive, leading to optimal recognition results.

Then, the generalization ability of the attention-BiLSTM approach was analyzed, and the inputs of the model were unified as a multimodal fusion of gait, face, and speech. T-tests were also conducted on the results. The comparison is shown in Table 10.

Table 10: Comparison of different LSTM models

	Accuracy	Precision	Recall rate	F1 value
RNN [14]	0.6818*	0.6875*	0.6667*	0.6769*
LSTM [15]	0.7121*	0.7188*	0.6970*	0.7077*
BiLSTM	0.7727*	0.7818*	0.7576*	0.7692*
Attention-LSTM [16]	0.8030	0.8125	0.7879	0.8000
Attention-BiLSTM	0.8182	0.8387	0.7879	0.8125

Table 10 reveals that the traditional RNN method performed poorly in psychological emotion recognition, with all indicators below 0.7. The LSTM method exhibited an accuracy of 0.7121, representing a 3.03% improvement compared to the RNN method, and an F1 value of 0.7077, indicating a 3.08% improvement compared to the RNN method. These findings verified the effectiveness of the LSTM method over the RNN method. Comparing the LSTM method with the BiLSTM method, the latter demonstrated a 6.06% improvement in accuracy and a 6.15% improvement in F1 value, indicating the enhancement achieved by the BiLSTM method relative to the LSTM method. Further, in the comparison between the LSTM and attention-LSTM methods, the latter exhibited a 9.09% improvement in accuracy and a 9.23% improvement in F1 value, showing the positive impact of the attention mechanism on recognition efficacy. Finally, the attention-BiLSTM approach outperformed the attention-LSTM approach with a 1.52% improvement in accuracy and a 1.25% improvement in F1 value. These results verified the reliability of the method proposed in this paper for recognizing psychological emotions in adults. From the statistical analysis of the results, it can be seen that there were significant differences between the recognition results of the attention-BiLSTM and the RNN, LSTM, and BiLSTM methods. Compared the attention-BiLSTM method with the attention-LSTM method, although there was no significant difference in accuracy and recall rate, there was a significant difference in precision and F1 value, further proving the superiority of the proposed method.

5 Discussion

With the advancement of technology, there have been increasing applications of methods such as machine learning and deep learning in recognizing psychological emotions in adults. The analysis of emotional features has also become more profound. However, current research shows that most studies on psychological emotion recognition focus only on single-modal features, such as text, speech, facial expressions, gait patterns, etc., with limited research on multi-modal fusion features. Nevertheless, there may be complementary relationships between different modalities' features. By integrating these features together, better recognition results can be achieved. Therefore, this study focuses on the recognition of psychological emotions in adults under multi-modal fusion by selecting gait, facial, and speech feature data. An

attention-BiLSTM method was designed to achieve identification between healthy and depressed groups.
The attention-BiLSTM model demonstrated good recognition performance on a binary benchmark dataset, outperforming methods such as NB and SVM. Furthermore, based on the experimental results obtained from the collected dataset, the fusion of gait, facial, and speech modalities showed better discrimination between depressed and healthy groups compared to single-modal approaches. The F1 value reached 0.8125, and this result exhibited significant differences when compared to other single-modal or multimodal recognition outcomes. The features from these three modalities complement each other, providing more comprehensive information. Therefore, the attention-BiLSTM model also achieved better effectiveness in feature learning and improved its ability to recognize depression and healthy populations. Comparing different LSTM models, it can be observed that the introduction of attention significantly distinguished the recognition results between the attention-BiLSTM and BiLSTM methods ($p < 0.05$). Furthermore, there was also a significant difference in precision and F1 value comparison between the attention-LSTM and attention-BiLSTM methods, further confirming the reliability of the attention-BiLSTM method.

Although some achievements have been made in the research on adult emotional recognition, there are still some shortcomings. For example, the methods used for extracting gait, facial, and speech features are relatively simple, and there is room for further optimization. The feature fusion method employed also uses a direct concatenation approach, which requires further discussion. In future work, more in-depth research will be conducted on feature extraction and fusion, and more comprehensive and extensive datasets will be obtained to validate the reliability of the proposed method further.

6 Conclusion

In this study, data comprising gait, facial, and speech features from depressed and healthy groups were collected for the recognition of adult psychological emotions. Subsequently, an attention-BiLSTM method was developed for recognition. Experimental analysis revealed that multimodal fusion yielded a superior recognition effect compared to unimodal approaches. Furthermore, compared with the RNN method and similar methods, the attention-BiLSTM approach proposed in this paper exhibited higher accuracy in adult psychological emotion recognition. These findings suggest the potential for further promotion and practical application of the attention-BiLSTM method.

References

[1] Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, Sheth A, Welton R, Pathak J (2019). Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. *The World Wide Web Conference*, pp. 514-525. https://doi.org/10.1145/3308558.3313698.

[2] Zhong Y, Sun L, Ge C, Fan H (2021). HOG-ESRs Face Emotion Recognition Algorithm Based on HOG Feature and ESRs Method. *Symmetry*, 13(2), pp. 1-18. https://doi.org/10.3390/sym13020228.

[3] Wani AH, Hashmy R (2023). A supervised multinomial classification framework for emotion recognition in textual social data. *International Journal of Advanced Intelligence Paradigms*, 24(1/2), pp. 173-189. https://doi.org/10.1504/IJAIP.2018.10027081.

[4] Hossain MS, Muhammad G (2019). An Audio-Visual Emotion Recognition System Using Deep Learning Fusion for a Cognitive Wireless Framework. *IEEE Wireless Communications*, 26(3), pp. 62-68. https://doi.org/10.1109/MWC.2019.1800419.

[5] Alamgir FM, Alam MS (2022). A Novel Deep Learning-Based Bidirectional Elman Neural Network for Facial Emotion Recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 36(10), pp. 1-37. https://doi.org/10.1142/S0218001422520164.

[6] Tsouvalas V, Ozcelebi T, Meratnia N (2022). Privacy-preserving Speech Emotion Recognition through Semi-Supervised Federated Learning. *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Pisa, Italy*, pp. 359-364. https://doi.org/10.1109/PerComWorkshops53856.2022.9767445.

[7] Murri MB, Triolo F, Coni A, Tacconi C, Nerozzi E, Escelsior A, Respino M, Neviani F, Bertolotti M, Bertakis K, Chiari L, Zanetidou S, Amore M (2020). Instrumental assessment of balance and gait in depression: A systematic review. *Psychiatry Research*, 284, pp. 112687. https://doi.org/10.1016/j.psychres.2019.112687.

[8] Inegbenosun HE, Tlasek-Wolfson M (2021). QIM21-086: Implementation of Depression Screening With the Patient Health Questionnaire-9 (PHQ-9) at a Radiation Oncology Department. *Journal of the National Comprehensive Cancer Network: JNCC*N, 19(3.5), pp. QIM21-086. https://doi.org/10.6004/jnccn.2020.7713.

[9] Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 59-66. https://doi.org/10.1109/FG.2018.00019.

[10] Eyben F, Wöllmer M, Schuller B (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. *ACM International Conference on Multimedia*, pp. 1459-1462. https://doi.org/10.1145/1873951.1874246.

[11] Balaji E, Brindha D, Elumalai VK, Vikrama R (2021). Automatic and non-invasive Parkinson's disease diagnosis and severity rating using LSTM

network. *Applied Soft Computing*, 108(4), pp. 1-14. https://doi.org/10.1016/j.asoc.2021.107463.

[12] Khademi Z, Ebrahimi F, Kordy HM (2022). A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals. *Computers in Biology and Medicine*, 143, pp. 105288. https://doi.org/10.1016/j.compbiomed.2022.105288.

[13] Yang G, Liu S, Li Y, He L (2023). Short-term prediction method of blood glucose based on temporal multi-head attention mechanism for diabetic patients. *Biomedical Signal Processing and Control*, 82, pp. 1-12. https://doi.org/10.1016/j.bspc.2022.104552.

[14] Liu Y, Zhang Q, Song L, Chen Y (2019). Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction. *Computers and Electronics in Agriculture*, 165, pp. 104964. https://doi.org/10.1016/j.compag.2019.104964.

[15] Shang S, Luo Q, Zhao J, Xue R, Sun W, Bao N (2021). LSTM-CNN network for human activity recognition using WiFi CSI data. *Journal of Physics: Conference Series*, 1883(1), pp. 1-9. https://doi.org/10.1088/1742-6596/1883/1/012139.

[16] Muhammad K, Mustaqeem, Ullah A, Imran AS, Sajjad M, Kiran MS, Sannino G, de Albuquerque VHC (2021). Human action recognition using attention-based LSTM network with dilated CNN features. *Future Generation Computer Systems*, 125, pp. 820-830. https://doi.org/10.1016/j.future.2021.06.045.