RESEARCH ARTICLE

Process analysis of facial expressions, movements, and psychological changes in depression based on deep learning algorithms

Qiong Zhao*

Zhengzhou University of Industrial Technology, Zhengzhou, Henan, China.

Received: July 23, 2024; **accepted:** December 20, 2024.

Depression is a common mental health disorder, and early identification and diagnosis are crucial to improve treatment outcomes. In this study, a deep learning model called Multimodal Attention Fusion Network (MAFN) was proposed to assist in the early diagnosis of depression by integrating facial expressions, body language, and voice data. By using public datasets from AffectNet, CREMA-D, and MPII Human Pose, as well as collecting and evaluating data from 10 patients suspected of early depression, the effectiveness of the MAFN model in identifying and predicting emotional changes in patients with depression were validated. The results showed that MAFN demonstrated significant improvements in accuracy, precision, recall, and F1 value compared with the unimodal model, demonstrating its advantages in processing multimodal data. The proposed model not only improved the accuracy of early diagnosis of depression, but also provided clinicians with more reliable auxiliary tools, which helped to intervene and treat patients with depression in a timely manner, thereby improving patients' treatment outcomes and quality of life.

Keywords: deep learning; depression; multimodality; expression; movement.

***Corresponding author:** Qiong Zhao, Zhengzhou University of Industrial Technology, Zhengzhou 450000, Henan, China. Email: zhaoqiong126@hotmail.com.

Introduction

According to the World Health Organization. depression is a common mental illness that affects more than 300 million people worldwide, which not only causes great psychological pain to patients, but can also lead to serious social dysfunction including reduced work ability and impaired interpersonal relationships [1]. In recent years, with the development of computer vision and artificial intelligence technologies, especially deep learning, it has become possible to assist in the diagnosis and monitoring of depression by analyzing non-verbal behavior [2]. In the face of the global mental health crisis, the prevalence of depression has continued to increase and has become a public health issue that cannot be ignored, especially among adolescents and young adults. This trend has aroused deep concern from all walks of life. Depression not only erodes individual mental health, leading to reduced productivity and quality of life, but also increases the burden on the medical system and has significant negative socioeconomic impacts. During the COVID-19 pandemic, social isolation, life stress and inequality have been particularly prominent. People have been under unprecedented psychological pressure. The risk of depression has increased, and the number of depression episodes has increased year by year across the country [3]. The diversity of depression is

reflected in individual differences in symptoms and experiences, which means that, even under similar conditions, the manifestations of depression may vary greatly. This complexity stems from the interweaving of multiple factors such as cultural background, gender, age, and genetic predisposition, which together affect the clinical manifestations of depression and patients' treatment response. Therefore, there are many challenges in finding universally applicable diagnostic criteria and treatment methods [4].

Traditional depression diagnostic methods including clinical interviews and self-assessment questionnaires, although they play an important role in assessing patients' psychological states, have obvious limitations. The interview process may be affected by the patient's level of consciousness, expression ability, and recall bias, making it difficult to capture subtle emotional changes. The accuracy of self-assessment questionnaires relies on the patient's self-insight, and these questionnaires sometimes cannot truly reflect the patient's emotions and state because the patient may deny, feel ashamed or have cognitive biases. In view of this, exploring more objective and comprehensive assessment methods has become an urgent need in the field of mental health research [5]. Studies have shown that patients with depression exhibit specific patterns in non-verbal behavior. For example, they may smile less, make less eye contact, lower their voices, and move more slowly. These changes reflect the patient's internal emotional state and may also be related to neurobiological mechanisms [6]. In recent years, machine learning and deep learning techniques have been widely used in the automatic recognition of nonverbal behaviors including facial expression recognition, speech emotion analysis, and gesture recognition. These techniques automatically extract and learn complex features by analyzing large amounts of training data, achieving high-precision classification and prediction. The deep convolutional neural networks (CNNs) are used to recognize depressive expressions from videos, while recurrent neural networks (RNNs) are used to capture the changes in emotions over time [7]. Although deep learning has shown great potential in analyzing nonverbal behaviors related to depression, most current studies focus on the analysis of a single modality such as facial expression or speech only, ignoring the value of integrating multimodal information. In addition, there is a lack of systematic research on the psychological integration changes of nonverbal behaviors of patients with depression over time, i.e. how these behaviors evolve during treatment.

This study aimed to fill this gap by using deep learning algorithms to explore the integrated change patterns of facial expressions and body language in patients with depression and their relationship with the improvement of patients' psychological state. The study used deep learning algorithms, especially CNNs and RNNs to analyze the multimodal nonverbal behaviors of patients with depression. Facial expression recognition, voice emotion analysis, and gesture recognition technology were adopted to build a comprehensive model to more comprehensively assess the patient's depressive state. By integrating multimodal non-verbal behavior data, this study was able to improve the accuracy and sensitivity of depression diagnosis and provide support for clinical decision-making. In addition, the results of the study could promote the development of remote monitoring and early intervention, help early detection and treatment of depression, thereby improve patients' quality of life and social function.

Materials and methods

Data collection and processing

The data used in this study were obtained from multiple sources including AffectNet, a large-scale facial expression dataset that provides a wide range of annotated facial expressions (https://ibug.doc.ic.ac.uk/resources/affectnet/) [7]. CREMA-D, a dataset for emotional speech that includes recordings of actors expressing

various emotions (https://github.com/CheyneyComputerScience/CREMA-D) [8]. MPII Human Pose, a dataset for human posture analysis that contains annotated images of people in various poses (https://human-pose.mpi-inf.mpg.de/) [9]. To ensure diversity and representativeness, individual data from different cultural backgrounds, age groups, genders, and severity of condition were collected. These datasets provided richly labeled information for model training, enabling us to build a robust and comprehensive model for analyzing nonverbal behaviors associated with depression. Sample selection followed a rigorous screening process to ensure the quality and relevance of the data. First, data from depressed patients were screened by clinical diagnostic criteria including the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) (DSM-5) diagnostic criteria for depression. The non-depressed control group consisted of individuals in good mental health with no history of mental illness. To ensure diversity in the data, the sample was selected across different age, gender, and ethnic backgrounds, while ensuring a balanced sample size in each group to avoid problems of category imbalance [10]. There were 1,000,000 images of facial expression collected from 2017 to 2019 retrieved from AffectNet, 7,442 audio files of emotional speech recordings collected in 2014 from CREMA-D, and 25,000 human pose images from 2014 to 2016 obtained from MPII Human Pose, respectively. All procedures were approved by the Institutional Review Board (IRB) of Beijing Normal University (Beijing, China) (Approval No. IRB-2023-001). All participants signed informed consent to ensure that the use of their data complies with ethical standards and laws and regulations. Data preprocessing is a key link to ensure the efficiency and accuracy of model training. The facial expression images were preprocessed by first localizing the facial region through the facial detection algorithm, and then performing the size normalization, grayscale conversion, and luminance normalization to eliminate the effects of illumination and angle changes. The speech data were preprocessed by removing muted segments, reducing noise, and normalizing volume before converting them to spectrograms or Mel Frequency Cepstrum Coefficients (MFCCs) to capture acoustic features. The posture preprocessing included the extraction of body language data using a skeleton keypoint detection algorithm followed by coordinate normalization and translation correction to remove the effects of camera position and rotation. To increase the robustness and generalization of the model, data enhancement techniques such as randomly rotating, flipping, and scaling the images, as well as applying slight noise and speed variations were used to the speech data, which helped the model to be more stable in the face of new data and reduce the risk of overfitting [11, 12].

Deep learning models

The proposed model used Multi-modal Attention Fusion Network (MAFN) architecture, which consisted of three main modules including modality-specific encoder, attention mechanism, and fusion layer. Modality-specific encoders included that each modality (facial expression, body language, and sound) had a separate encoder for extracting its intrinsic features. For facial expression image data, a pre-trained ResNet for feature extraction was used, while, for sound data, a Long Short-Term Memory (LSTM) based network was used to capture temporal properties and, for body language, a Graph Convolutional Network (GCN) was applied to analyze the spatial relationships between key points of the skeleton. The attention mechanism covered that, considering that different modalities might contribute differently to depression diagnosis in different contexts, a multi-head self-attention mechanism was introduced to dynamically assign weights to the features of each modality, which helped the model to utilize the information more intelligently in the fusion stage and improve the overall performance. The fusion layer was responsible for integrating features from different modalities to generate a comprehensive representation. A gated fusion unit was used, which could selectively fuse features based on the attention weights of each

modality, thus reducing redundant information and improving the efficiency and robustness of the model [13, 14]. To extract features of facial expressions from image data, a pre-trained ResNet model was used to solve the gradient vanishing problem and degradation problem in deep networks by introducing a residual learning framework, which enabled the model to learn deeper features. Assuming that the input image was I, its feature vector F_I could be computed as shown in equation 1 [15].

$$F_I = ResNet(I) \quad (1)$$

where ResNet represented the entire network structure including multiple residual modules, each containing two layers of CNNs and a jump connection for passing the input directly to the output, thus avoiding gradient vanishing. The final feature vector v contained rich details of facial expressions. For the temporal characterization of sound data, a network-based LSTM was used, which was a special kind of RNN that remembered information in long sequences. The feature vector $F_S(t)$ of the sound data *S(t)* at each time *t* could be calculated by LSTM updating state as shown below [16].

$$F_S(t) = LSTM(S(t), h_{t-1}, c_{t-1}) \quad (2)$$

where h_{t-1} and c_{t-1} were the hidden state and unit state of the previous time step, respectively. The LSTM operation included the computation of input gates, forgetting gates, and output gates to ensure that the model effectively memorized important information and ignored irrelevant information. GCN was used for body language analysis, which was able to perform convolutional operations on graph-structured data and was particularly suitable for dealing with the spatial relationships between key points of the skeleton. Let the key points of the limb language be the graph G = (V, E), where V was the set of vertices and E was the set of edges, then the feature vector F_L could be defined as follows [17].

$$F_L = GCN(G, X) \quad (3)$$

where *X* was the node feature matrix. GCN updated the features of each node by aggregating the information of neighboring nodes, and the specific calculation process involved the Laplace matrix L and the weight matrix W, which were specified in equation 4.

$$F_L = \hat{A}XW \quad (4)$$

where $\hat{A}$ was a normalized adjacency matrix that captured the correlation between nodes in the graph. To dynamically assign weights to the features of different modalities, a multi-head self-attention mechanism was introduced. Suppose there were m modalities, and the feature vector of each modality was $F^{(i)}$, which was first converted into the query $Q^{(i)}$ with the key $K^{(i)}$, and the summation value $V^{(i)}$ as specified in equations 5 - 7 [18].

$$Q^{(i)} = F^{(i)}W_Q^{(i)} \quad (5)$$

$$K^{(i)} = F^{(i)}W_K^{(i)} \quad (6)$$

$$V^{(i)} = F^{(i)}W_V^{(i)} \quad (7)$$

where $(W_Q^{(i)}, W_K^{(i)}, W_V^{(i)})$ were the learnable weight matrix. The attention score $A^{(i)}$ was calculated as below.

$$A^{(i)} = softmax(Q^{(i)}(K^{(i)})^T / \sqrt{d_k}) \quad (8)$$

where d_k was the dimension of the key vector. The weighted feature vector was shown in equation 9.

$$F'^{(i)} = A^{(i)}V^{(i)} \quad (9)$$

The multi-head attention mechanism computed multiple attention heads in parallel by finally stitching them together and mapping them to the

output space through a fully connected layer W_O as specified in equations 10 - 11.

$$F^{att} = Concat(head_1, ..., head_h)W_O \quad (10)$$

$$head_i = Attention(Q^{(i)}, K^{(i)}, V^{(i)}) \quad (11)$$

The goal of the fusion layer was to integrate features from different modalities to generate a comprehensive representation. This task was accomplished using a gated fusion unit (GFU), which was able to selectively fuse features based on the attention weights $A^{(i)}$ to reduce redundant information. Let the weighted features be $F'^{(i)}$, the fusion operation of the GFU was shown below.

$$F_{fused} = GFU(F'^{(1)}, ..., F'^{(m)}, A^{(1)}, ..., A^{(m)}) \quad (12)$$

where the GFU controlled the fusion ratio through the gating function G as specified in equations 13 - 14.

$$G(F'^{(i)}, F'^{(j)}, A^{(i)}, A^{(j)}) = \sigma(W_g[A^{(i)}; A^{(j)}]) \quad (13)$$

$$F'^{(i)} + (1 - \sigma(W_g[A^{(i)}; A^{(j)}])) \quad (14)$$

The GFU was able to selectively fuse the features of all modes to generate the final integrated representation F_{fused}. The integrated representation F_{fused} was obtained by further processing based on the output of the gated fusion unit. The fused features could be mapped to a new representation space through a fully connected layer W_{fuse} and an activation function ϕ, which was more suitable to be used as an input to the final classifier as specified below.

$$F_{final} = \phi(W_{fuse}F_{fused} + b_{fuse}) \quad (15)$$

where b_{fuse} was the bias term and ϕ could be ReLU, tanh, or other activation functions depending on the design of the model and the task requirements. From F_{final}, the probability distribution of each category for the final decision or categorization was obtained through a categorization layer, usually another fully connected layer plus a *Softmax* function as specified below.

$$P = Softmax(W_{cls}F_{final} + b_{cls}) \quad (16)$$

where W_{cls} and b_{cls} were the weights and bias terms of the categorization layer. *P* was a vector of predicted probability distributions with each element corresponding to the predicted probability of a category.

Experimental design

The experiments were executed on a high-performance computing cluster to ensure sufficient resources to support the demands of large-scale data processing and model training. The hardware software configurations used for efficient training and inference of the model included Intel Xeon E5-2698 v4 CPUs, 256 GB of DDR4 RAM, 1 TB of NVMe SSD, NVIDIA Tesla V100 GPU, Ubuntu 20.04 LTS Operating System (Canonical Ltd, London, UK), TensorFlow 2.4.0 (Google, Mountain View, California, USA), the PyTorch 1.7.0 deep learning framework (https://pytorch.org/), and Python 3.8 (https://www.python.org/). The aim of proposed MAFN was to identify and understand complex emotional states, particularly the diagnosis of depression, by integrating facial expression, body language and voice data. This integration was accomplished through a well-designed multimodal analysis process that included cross-modal feature extraction, a multi-head self-attention mechanism, and the use of gated fusion units (Figure 1). The datasets were randomly divided into training, validation, and testing sets with the proportions being set to 70%, 15%, and 15%. In addition, a K-fold cross-validation strategy was implemented to further validate the stability of the model performance. In the training phase, the batch gradient descent method and Adam optimizer were used to

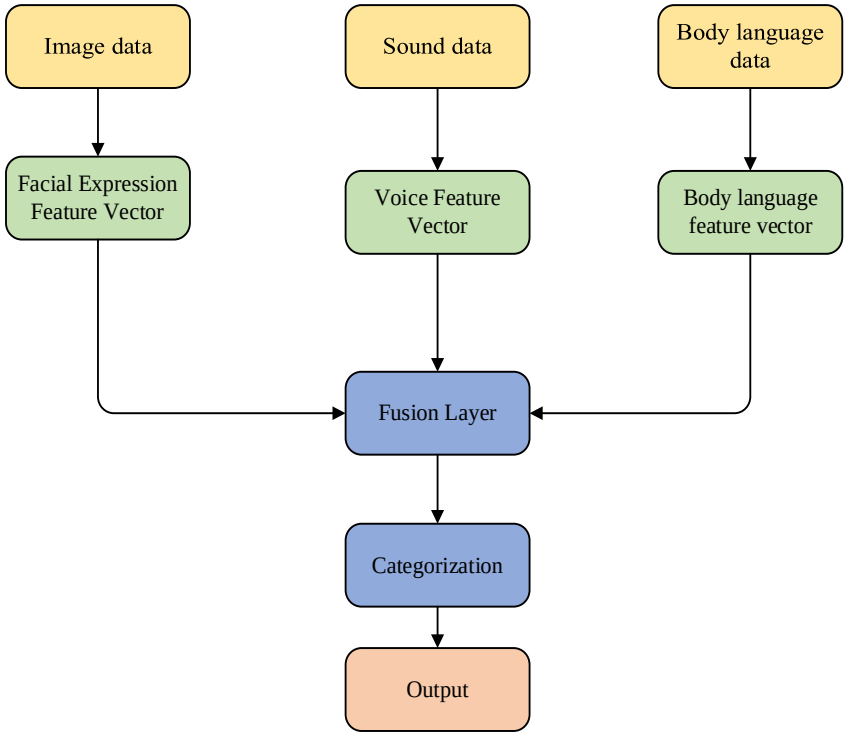

Figure 1. Multimodal framework.

minimize the loss function that typically included cross-entropy loss and regularization terms to balance the accuracy and complexity of the model. During the training, the performance metrics on the validation set were monitored including accuracy, precision, recall, and F1 score to prevent overfitting and adjust the model when appropriate. Ten (10) patients from Beijing People's Hospital (Beijing, China) with suspected early depression aged from 25 to 45 years with a balanced gender distribution were selected for this study. Each patient consented to participate in the study and authorized the collection of data on their facial expressions, body movements and voice characteristics. Data collection took place in a natural environment to ensure authenticity and validity of the data. A Sony HDR-CX405 high-definition camera (Sony Corporation, Tokyo, Japan) was utilized to record each patient's facial expressions and body movements, while voice characteristics including intonation, speech rate, and pauses were captured using Boya BY-VM300 recording equipment (Shenzhen Boya Technology Co., Ltd., Shenzhen, Guangdong, China). The collected multimodal data were fed into the MAFN model, which was fully trained to recognize features associated with early depression. The model generated a risk score for each patient and the specific contribution of each modality to the risk score. The proposed MAFN was compared with the CNN, LSTM, and TMM models for its performance evaluation.

Results

The accuracy of different modalities on the expression and action recognition task

By comparing the accuracies of the training set, validation set, and test set, the recognition effect of multimodal fusion was significantly better than that of a single modality, which suggested that combining multiple sources of information including facial expressions, body movements, and audio emotions could more accurately recognize an individual's emotional state (Table 1). However, the accuracy of each modality decreased from the training set to the test set,

Table 1. Expression and motion recognition accuracy.

Modal (computing, linguistics)	Training set accuracy	Validation set accuracy	Test set accuracy
Facial expression	90.5%	88.3%	87.5%
Body movement	85.2%	83.4%	82.1%
Audio Mood	89.6%	87.9%	86.5%
Multimodal fusion	94.3%	92.7%	91.8%

Table 2. Analysis of changes in psychological integration - time series.

Point in time (math.)	Mean depression score	Amount of change in facial expression	Amount of change in body movement
Week 1	18.5	0.2	0.1
Week 2	17.2	0.1	0.05
Week 3	15.8	0.05	0.03
Week 4	14.5	0.02	0.01

which might be due to the overfitting of the model on the training set or differences in the data distribution of the test set from the training set.

Time-series analysis of changes

The psychological integration, the trends in the mean depression scores, the amount of change in facial expressions, and the amount of change in body movements of the participants over time were investigated. The results showed that depression scores gradually decreased with the increase of time, indicating an improvement in the psychological state of the subjects, while the number of changes in facial expressions and body movements decreased, which might be related to the stabilization of the psychological state (Table 2). These results implied the effectiveness of psychological intervention or treatment.

Correlation analysis of mental states and behavioral patterns

The correlation analysis between psychological states and behavioral patterns showed that, by calculating the correlation coefficients and *P* values between depression scores and different behavioral patterns, there was a significant positive correlation between negative facial expressions, reduced body movements, low audio intonation and depression scores (Table 3), which meant that, when individuals exhibited these behavioral patterns, they might be at a higher risk for depression. These findings helped to better understand the relationship between psychological states and behaviors and provided behavioral indicators for psychological assessment.

Case study results

The risk scores for early depression identified by the MAFN model and their contribution of each modality in 10 patients were shown in Table 4. Based on the assessment results of the MAFN model, five patients (P001, P003, P005, P008, P009) demonstrated higher overall risk scores than the others, suggesting that they might be in an early stage of depression. The MAFN model demonstrated strong analytical power in handling multimodal data and was effective in identifying early signs of depression. Based on this finding, physicians should consider further psychological assessment of these patients to

Table 3. Correlation analysis of mental states and behavioral patterns.

Behavioral model	Depression score correlation coefficient	*P* value
Negative facial expressions	0.78	< 0.001
Decreased body movements	0.65	< 0.001
Audio tone is low	0.72	< 0.001

Table 4. Case study results.

Patient ID	Gender	Age	Facial expression contribution	Body Movement Contribution	Intonation contribution	Overall risk score
P001	F	28	40%	30%	30%	0.5
P002	M	35	35%	25%	40%	0.45
P003	F	32	50%	20%	30%	0.6
P004	M	42	30%	40%	30%	0.4
P005	F	30	45%	25%	30%	0.55
P006	M	40	30%	35%	35%	0.4
P007	F	27	40%	30%	30%	0.45
P008	M	38	45%	25%	30%	0.5
P009	F	33	50%	20%	30%	0.6
P010	M	45	30%	40%	30%	0.45

start intervention and treatment as early as possible. In addition, the wide application of the MAFN model would help to promote the development of precision medicine in mental health and improve the early identification rate of depression, thereby improving the overall treatment outcome and life quality of patients.

Comparison of different models

The performances of proposed MAFN model and other different models on the multimodal recognition task were compared and demonstrated that MAFN model outperformed the other models in terms of accuracy, precision, recall, and F1 score, showing its superiority in handling multimodal data (Figure 2). In contrast, unimodal models of CNN and LSTM performed poorly, which might be due to their inability to efficiently integrate information from different modalities.

Discussion

This study revealed the remarkable potential of multimodal attention fusion networks (MAFN) in recognizing and predicting mood changes in depressed patients. By comparing the recognition accuracies of different modalities, the consistently superior performance of the multimodal fusion model was observed across all datasets, which demonstrated the importance of integrating multiple sources of information such as facial expressions, body movements, and audio emotions. The results of this study suggested that the joint use of multimodal information significantly improved the accuracy of emotion recognition, and MAFN model was

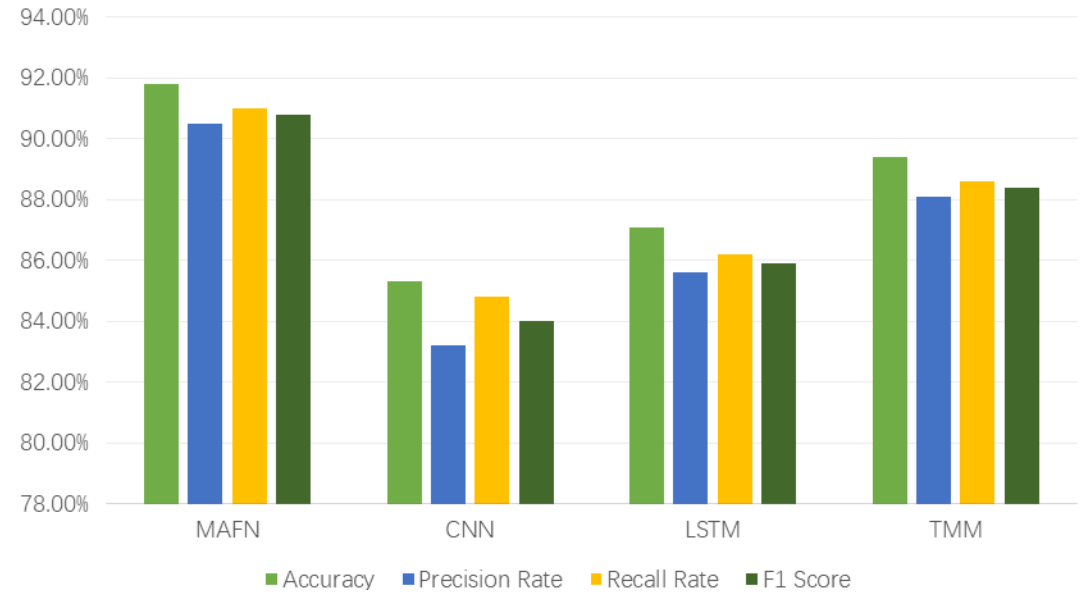

Figure 2. Comparison of MAFN model performance with other models.

equipped to handle complex scenarios and improve diagnostic accuracy. This research confirmed that MAFN model could assist in the diagnosis of early depression by analyzing multimodal data to improve the accuracy and timeliness of identification. Time-series analysis demonstrated a clear trend in the changes of depression scores and related behavioral patterns with psychological interventions, indicating that MAFN model was able to sensitively capture small changes in patients' psychological states. In addition, correlation analysis found a significant association between behavioral patterns and depression scores, further confirming the value of nonverbal behavior in the early identification of depression. This study delved into the application of multimodal deep learning in the early diagnosis of depression, proposing an innovative multimodal attention fusion network (MAFN) model. By integrating facial expression, body language, and voice data, the MAFN model demonstrated superior ability to recognize and predict mood changes in depressed patients, especially in processing complex and diverse data. The results showed that MAFN significantly outperformed traditional unimodal models on multimodal datasets. Due to its unique multi-head self-attention mechanism and gated fusion unit, the model was enabled to intelligently distribute and integrate information from different modalities. This research analyzed a large amount of public data and conducted model training and validation on a high-performance computing cluster to ensure the scientific validity of the experiments and the reliability of the results. In addition, through case studies, the power of MAFN model to identify early signs of depression in real-world scenarios were demonstrated, which provided a powerful tool to the physicians with a view to achieving earlier intervention and treatment.

Acknowledgements

This research was supported by 2022 Ministry of Education's Industry-University Cooperative Education (Grant No. 220702860220432), 2023 Henan Province University Humanities and Social Sciences Research Project (Grant No. 2023-ZDJH-608), 2022 Zhengzhou Social Science Research (Grant No. ZSLX20220114).

References

1. Farré A, Tirado J, Spataro N, Alías-Ferri M, Torrens M, Fonseca F. 2020. Alcohol induced depression: Clinical, biological and genetic features. J Clin Med. 9(8):2668.
2. Kim DH, Son WH, Kwak SS, Yun TH, Park JH, Lee JD. 2023. A hybrid deep learning emotion classification system using multimodal data. Sensors. 23(23):9333.
3. Jin X, Sun WY, Jin Z. 2020. A discriminative deep association learning for facial expression recognition. Int J Mach Learn Cybern. 11(4):779–793.
4. Chai WH, Wang GA. 2022. Deep vision multimodal learning: Methodology, benchmark, and trend. Appl Sci. 12(13):6588.
5. Akbar MN, Riaz F, Awan AB, Khan MA, Tariq U, Rehman S. 2022. A hybrid duo-deep learning and best features based framework for action recognition. Comput Mater Contin. 73(2):2555–2576.
6. Wang DH, Zhao T, Yu WH, Chawla N, Jiang M. 2023. Deep multimodal complementarity learning. IEEE Trans Neural Netw Learn Syst. 34(12):10213–10224.
7. Le Bivic G, Limosin F, Lemogne C, Hoertel N. 2023. Depression in older adults: What are the differences in clinical practice? Gériatrie et Psychologie Neuropsychiatrie du Vieillissement. 21(2):268-276.
8. Fang B, Zhao Y, Han G, He J. 2023. Expression-guided deep joint learning for facial expression recognition. Sensors. 23(16):7148.
9. Ahmad M, Saira, Alfandi O, Khattak AM, Qadri SF, Saeed IA, *et al*. 2023. Facial expression recognition using lightweight deep learning modeling. Math Biosci Eng. 20(5):8208–8225.
10. Morden E, Byron S, Roth L, Olin SCS, Shenkman E, Kelley D, *et al*. 2021. Health plans struggle to report on depression quality measures that require clinical data. Acad Pediatr. 22(3):S133–S139.
11. Guo WZ, Wang JW, Wang SP. 2019. Deep multimodal representation learning: A survey. IEEE Access. 7:63373–63394.
12. Liang J, He PJ, Wu HT, Xu XJ, Ji CH. 2022. Characteristics of depression clinical trials registered on ClinicalTrials.gov. Int J Gen Med. 15:78–96.
13. Addington J, Farris MS, Liu L, Cadenhead KS, Cannon TD, Cornblatt BA, *et al*. 2021. Depression: An actionable outcome for those at clinical high-risk. Schizophr Res. 227:38–43.
14. Hieronymus F, Jauhar S, Ostergaard SD, Young AH. 2020. One (effect) size does not fit all: Interpreting clinical significance and effect sizes in depression treatment trials. J Psychopharmacol. 34(10):1074–1078.
15. Parker G. 2021. Clinical depression: The fault not in our stars? Australas Psychiatry. 29(6):652–654.
16. Choi JH, Lee JS. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. Inf Fusion. 51:259–270.
17. Shehzad F, Khan MA, Yar MAE, Sharif M, Alhaisoni M, Tariq U, *et al*. 2023. Two-stream deep learning architecture-based human action recognition. Comput Mater Contin. 74(3):5931–5949.
18. Tsai JK, Hsu CC, Wang WY, Huang SK. 2020. Deep learning-based real-time multiple-person action recognition system. Sensors. 20(17):4758.