Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 1636855, 10 pages
https://doi.org/10.1155/2022/1636855

Research Article

Using Convolutional Neural Networks for the Assessment Research of Mental Health

Yanbing Liu

School of Management, Northwestern Polytechnical University, Xi'an 710129, Shanxi, China

Correspondence should be addressed to Yanbing Liu; liuyanbing@mail.nwpu.edu.cn

Received 17 March 2022; Revised 11 April 2022; Accepted 12 April 2022; Published 9 May 2022

Academic Editor: Gengxin Sun

Existing mental health assessment methods mainly rely on experts' experience, which has subjective bias, so convolutional neural networks are applied to mental health assessment to achieve the fusion of face, voice, and gait. Among them, the OpenPose algorithm is used to extract facial and posture features; openSMILE is used to extract voice features; and attention mechanism is introduced to reasonably allocate the weight values of different modal features. As can be seen, the effective identification and evaluation of 10 indicators such as mental health somatization, depression, and anxiety are realized. Simulation results show that the proposed method can accurately assess mental health. Here, the overall recognition accuracy can reach 77.20%, and the $F1$ value can reach 0.77. Compared with the recognition methods based on face single-mode fusion, face + voice dual-mode fusion, and face + voice + gait multimodal fusion, the recognition accuracy and $F1$ value of proposed method are improved to varying degrees, and the recognition effect is better, which has certain practical application value.

1. Introduction

With the development of economy and the acceleration of the pace of life, people's life pressure is becoming bigger and bigger, and mental health problems have become the focus of global attention. At present, the methods of mental health assessment are mainly based on experts' assessment or self-assessment, which is assessed from the perspective of the patient and the practitioner. Briggs Hannah et al. explored the thoughts, feelings, and educational requirements of nursing staff and nurses at the clinical help desk of emergency medical services, and the focus was on the classification tools used for calls and uses related to mental health. Here, quantitative data are analyzed by descriptive statistics, and qualitative data are analyzed by subject analysis. Thus, mental health assessment and triage of patients and their families are realized [1]. Scelzo Anna evaluated mental health in the form of questionnaires and believed that a good mental health assessment is conducive to promoting healthy aging [2]. Michael R. Hass et al. proposed a concept of case conceptualization and realized the assessment of students' mental health by determining students' psychological needs and writing goals. This method is better than the traditional evaluation process [3]. Fortuna Lisa R. adopted the 2.2 pros and cons method to introduce trauma narrative in the process of sheltered mental health assessment, so as to improve the accuracy of mental health assessment [4]. Scott A. Bresler, Ph. D., reviewed the mental health assessment by forensic in the digital era and believed that the rational use of Internet data is conducive to accurately assess mental health [5]. Higuchi Masakazu et al. constructed a mental health assessment system based on voice modes in a mobile device based on voice, which opens the mental health voice assessment with certain foresight [6]. Newson Jennifer J. et al. assessed the Chinese and Canadian interactive mental health by taking a pilot primary care outpatient clinic led by nurse practitioners as the research object, which is conducive to strengthening the mental health communication between clinicians and patients [7]. OReilly Michelle et al. analyzed 28 videos recording British children's psychology by using discourse psychology, established a rhetorical case to prove the clinical need, and believed that children's mental health is related to parents' teaching by words and deeds [8]. Since then, with the development of information technology,

people began to introduce computer-aided methods to evaluate psychology, such as Heesacker Martin using computer system, and CNN proposed by some scholars to evaluate psychology [9–11]. The above research indicates that most of the current mental health assessment methods are mainly based on experts' experience and analysis, and there is a certain degree of subjectivity. In order to better objectively assess mental health, an automatic intelligent assessment method of mental health based on the rapidly developing convolutional neural network is proposed.

2. Basic Methods

2.1. Introduction to OpenPose Algorithm. OpenPose algorithm is a bottom-up algorithm based on convolutional neural network, which is suitable for single and multiperson pose recognition and has good robustness. The basic structure of OpenPose algorithm is shown in Figure 1, and there are two branches and multistage convolutional neural networks [12]. Here, the yellow and blue parts represent one branch, respectively, and the left and right parts represent two phases, respectively. The yellow branch is used to describe the confidence map of face and posture key points, and the blue branch is used to describe the correlation degree of each key point. The left part is responsible for generating detection confidence maps and partial affinity domains, and the right part is responsible for connecting the prediction results of different branches of yellow and blue to improve the prediction accuracy.

2.2. Introduction to Multimodal Fusion. Multimodal fusion refers to the fusion of feature information of different modes [13], including three fusion modes, namely, data layer fusion, feature layer fusion, and decision layer fusion. Data layer fusion first combines data, extracts features from the combined data, and then inputs them into a classifier for recognition. Feature layer fusion extracts different modal information data features separately and inputs the combined features into a classifier for recognition. Decision layer fusion extracts data features separately, identifies each extracted feature, and finally fuses the recognition results. In practical applications, multimodal fusion based on data layer plays a positive role in recognition task, but its fusion efficiency is low. The multimodal fusion method based on feature layer may increase the amount and difficulty of calculation because it cannot screen effective features, thus reducing the model recognition results. The fusion method based on decision layer only combines the results of different modes but theoretically does not really integrate the information of all modes [14, 15]. According to the characteristics of mental health assessment mainly from three aspects of human face, voice and gait, as well as reference [16], the multimodal fusion method based on feature layer is adopted, and its basic process is shown in Figure 2.

2.3. Introduction to Attention Mechanism. Attention mechanism is a kind of perception mode that simulates the human brain to selectively attach importance to useful information and discard useless information, which is first applied in the field of visual images. In recent years, with the in-depth study of attention mechanism, it has been widely used in image recognition, recommendation system, and other fields. The attention mechanism usually follows the form of query (Q), keyword (K), and weight value (V). The structure of the classical attention mechanism is shown in Figure 3 [17].

When the attention mechanism is introduced to distribute weight, formula (1) can be used to distribute weight [18]:

$$\text{Attention}(Q, K, V) = \sum_{i=1}^{L} \text{Similarity}(Q, K_i) * V_i, \tag{1}$$

where L represents the number of keywords and *Similarity* ( ) means similarity calculation function, which usually includes the following three functions:

$$\begin{aligned} \text{Double} - \text{linear model Similarity}(Q, K_i) &= K_i^T W Q, \\ \text{Dot product model Similarity}(Q, K_i) &= K_i^T Q, \\ \text{Scale dot product model Similarity}(Q, K_i) &= \frac{K_i^T Q}{\sqrt{\mathrm{d}}}, \end{aligned} \tag{2}$$

where W represents the learnable parameter and d represents the dimension of keyword and weight value.

2.4. Introduction of SVM. SVM algorithm is a nonstatistical classification algorithm, and its basic principle is based on nonlinear transformation regression function to map sample data to high-dimensional feature space, which can realize the sample data conversion. Its kernel function is defined as follows:

$$K(x, z) = \varphi(x)^T \varphi(z), \tag{3}$$

where x and z represent data points in the original space and φ represents nonlinear transformation. In general, the kernel function of SVM is Gaussian kernel function, as shown in the following:

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right), \tag{4}$$

where z represents the center value of Gaussian function and σ represents the width parameter of function.

3. Using Convolutional Neural Networks for the Assessment Research of Mental Health

3.1. Indicators of Mental Health. In this experiment, referring to literature, there are 10 indicators for mental health, which are shown in Table 1 [19].

3.2. Overall Framework of the Model. Based on the above analysis, multiple modes of face, voice, and gait are integrated based on the OpenPose algorithm of convolutional

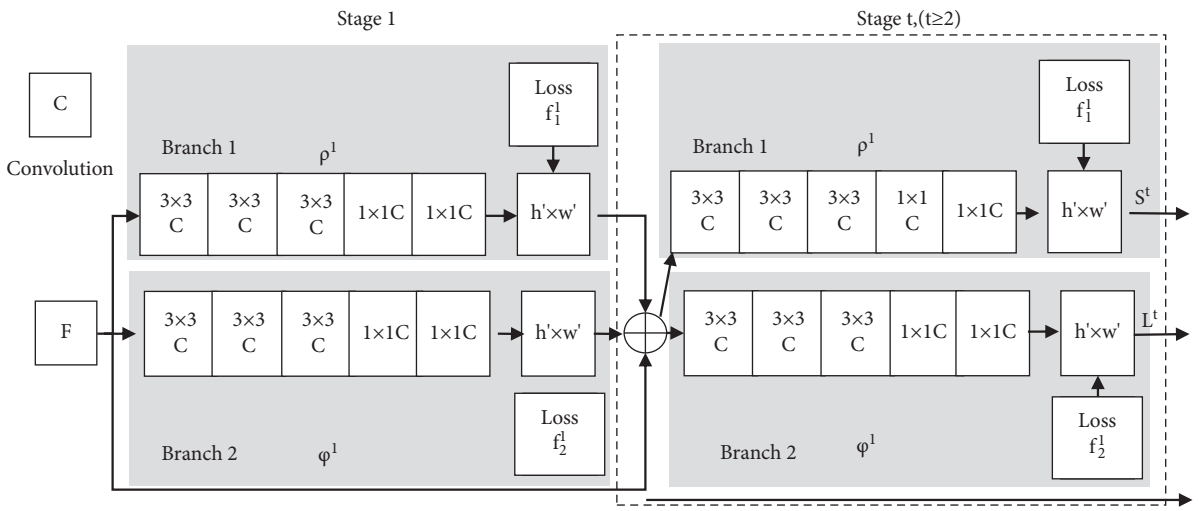

Figure 1: Flow chart of OpenPose algorithm.

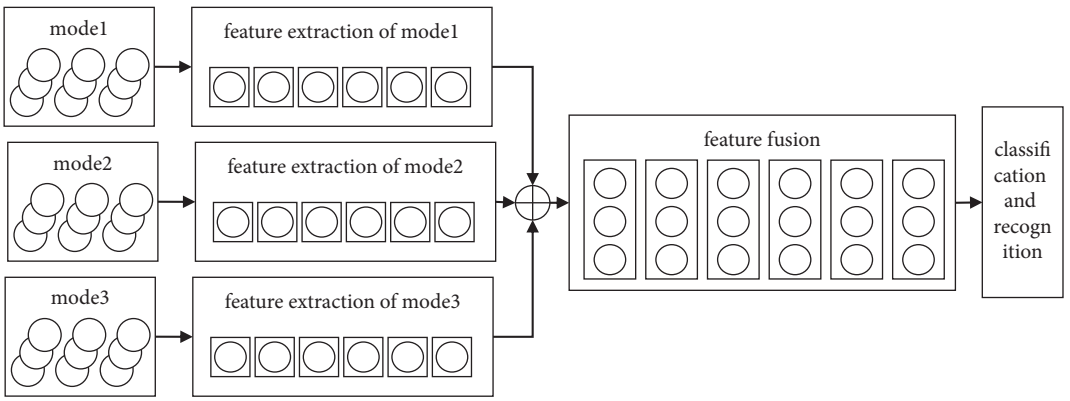

Figure 2: Multimodal fusion and recognition based on feature layer.

neural network, and attention mechanism is used to allocate the weight of different modes reasonably. A mental health assessment model of multimode fusion with the introduction of the attention mechanism is proposed, and its overall framework is shown in Figure 4. Firstly, OpenPose algorithm is used to extract key points of human face and posture, and openSMILE is used to extract low-level voice descriptors. Then, the modal characteristics can be calculated by time domain statistical parameters. Finally, attention mechanism is introduced to allocate the weight of each mode reasonably, and support vector machine (SVM) is used to classify and recognize the mental health assessment.

3.2.1. Feature Extraction. For feature extraction of face and gait images, OpenPose algorithm is adopted to extract key points of face and gait. At the same time, face data and gait data are input into the algorithm to generate detection confidence graph combination S and confidence graph unit S_j, whose calculation formulas are shown as follows:

$$S = \left(S_1, S_2, \ldots, S_j\right), \quad j \in \{1, \ldots, J\}, \tag{5}$$

$$S_j = i^{w \times h}, \tag{6}$$

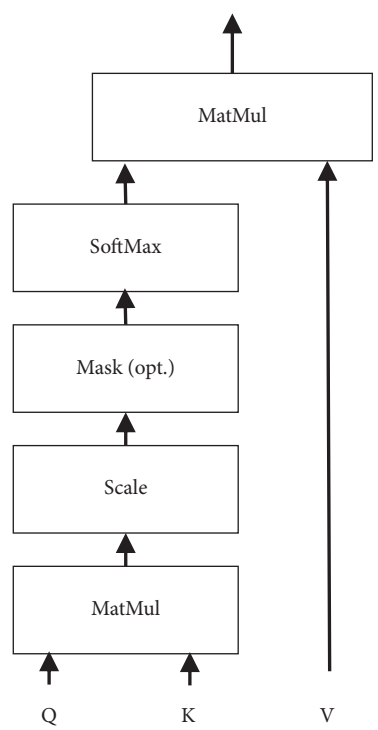

Figure 3: Structure of classical attention mechanism.

Table 1: Description of mental health indicators.

Indicators of mental health	Description
Somatization	Somatization is mainly used to express the discomfort of subjects
Obsessive-compulsive	A repeated and persistent compulsion or behavior
Interpersonal relationship sensitivity	A strong sense of inferiority and uneasiness in dealing with others
Depression	Having a lot of pessimism inside and being in a low mood
Anxiety	Showing excessive agitation, restlessness, nervousness, etc.
Hostility	Feelings of hostility or antagonism towards others
Dreadness	An intense and unnecessary fear towards certain objects or situations
Crankiness	Excessive attachment to certain things or ideas
Psychopathy	An emotional, cognitive, or behavioral disorder caused by dysfunctions in the brain
Other	Diet status, sleep quality, and so on

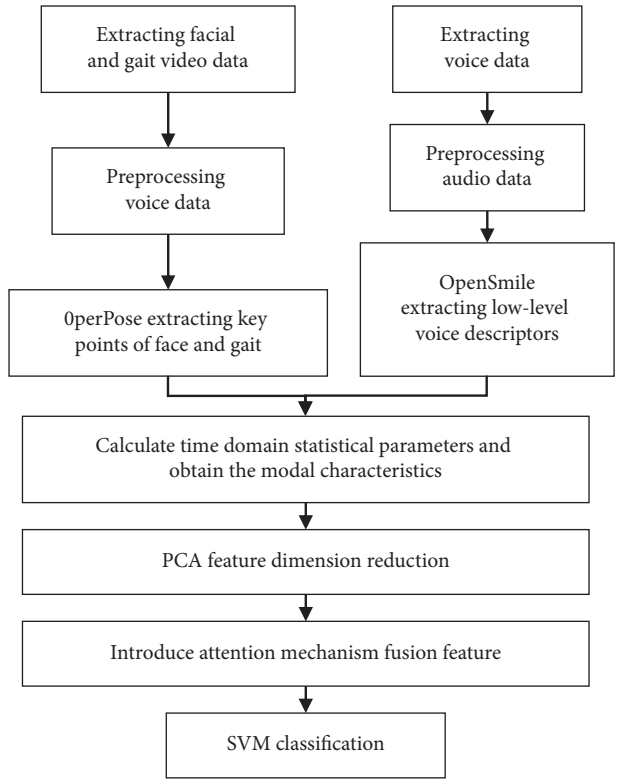

Figure 4: Structure of multimodal fusion mental health assessment model with attention mechanism.

where J values are 68 and 18, respectively; then coordinate set $Fcoo_t$ of key points of face image in frame T and coordinate set $Gcoo_t$ of key points of gait image in frame T can be expressed as follows:

$$\mathrm{Fcoo}_t = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{68}, y_{68})\}, \tag{7}$$

$$\mathrm{Gcoo}_t = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{18}, y_{18})\}. \tag{8}$$

Here, openSMILE method is used to extract the short time energy, formant, pitch frequency, and MFCC of voice features. The short-time energy $E(i)$ of frame i voice signal $y_i n$ can be calculated by formula (11):

$$E(i) = \sum_{n=0}^{L-1} y_i^2 n, \quad 1 \le i \le f_n. \tag{9}$$

By calculating data(n) of voice signal and carrying out Fourier transform, pitch frequency PF is solved.

$$\mathrm{data}(n) = u(n) \times v(n), \tag{10}$$

where $v(n)$ represents the corresponding filtering of sound channel and $u(n)u(n)$ represents the excitation response of glottis pulse.

The formant parameters For1, For2, and For3 of voice signal peak are calculated by LPC root method [20].

According to formula (11), the spectrum of voice signal is calculated. Combined with Mayer filter and discrete cosine transform, the 12th-order MFCC features are obtained, which can be expressed as formula (12):

$$P = \frac{\left|\mathrm{FFT}\left(x_i\right)\right|^2}{N}, \tag{11}$$

$$\mathrm{MFCC} = \left\{\mathrm{mfcc}_1, \mathrm{mfcc}_2, \ldots, \mathrm{mfcc}_{12}\right\}, \tag{12}$$

where P is the power spectrum; FFT is the fast Fourier transform; X_i is the voice signal; and $N = 512$.

3.2.2. Calculating Time Domain Statistical Parameter. Time domain features can describe the characteristics of different data in time dimension. In addition, arithmetic sum, mean, minimum, maximum, variance, standard deviation, skewness, kurtosis, and correlation coefficient between two axes are selected as the calculation types of time domain feature statistical parameters by referring to literature [21, 22].

3.2.3. Introducing Attention Mechanism. Similarity calculation of mental health assessment query and keyword with attention mechanism includes two aspects, namely, dot product calculation of vector check and cosine similarity calculation, as shown in the following formulas [23]:

$$\mathrm{similarity1}\left(\mathrm{Query}, \mathrm{Key}_i\right) = \mathrm{Query} \bullet \mathrm{Key}_i, \tag{13}$$

$$\mathrm{similarity2}\left(\mathrm{Query}, \mathrm{Key}_i\right) = \frac{\mathrm{Query} \bullet \mathrm{Key}_i}{\|\mathrm{Query}\| \bullet \left\|\mathrm{Key}_i\right\|}. \tag{14}$$

Then, the weight coefficient is solved through normalization operation, as shown as follows:

$$a_i = \mathrm{soft\,max}\left(\mathrm{Sim}_i\right) = \frac{e^{\mathrm{Sim}_i}}{\sum_{j=1}^{L_x} e^{\mathrm{Sim}_i}}. \tag{15}$$

Finally, formula (16) shows the weighted sum of weight coefficients, the final fusion feature Fusatt can be obtained [24]. The size of Fusatt is $103 * 4$, and its calculation method is shown in formula (17) [25]:

$$\mathrm{Attention}\left(\mathrm{Query}, \mathrm{Source}\right) = \sum_{i=1}^{L_x} a_i \bullet \mathrm{Value}_i, \tag{16}$$

$$\mathrm{Fus}_{\mathrm{att}} = \left[w1 \times F_{\mathrm{fea}}, w2 \times V_{\mathrm{fea}}, w3 \times G_{\mathrm{fea}}\right]. \tag{17}$$

In formula (17), $F_{\mathrm{fea}}, V_{\mathrm{fea}}$, and$G_{\mathrm{fea}}$ represent facial, phonetic, and gait features, respectively.

3.2.4. SVM Classification. According to the evaluation indicators of mental health, there are 10 SVM classifiers trained to judge different mental health conditions, corresponding to 10 mental health indicators such as somatization, depression, and anxiety. Each psychological index includes negative and positive two states, corresponding to not suffering from or suffering from the corresponding psychological disease of this index.

4. Simulation Experiment

4.1. Construction of Experimental Environment. This experiment is conducted on Windows7 operating system with Intel Xeon Silver4110 CPU; graphics card is NVIDIA GeForce GTX 1080Ti with of 16 G memory; the memory is 128 G; and the development language is Python.

4.2. Data Sources and Preprocessing

4.2.1. Data Sources. In this experiment, facial, gait, and voice data of 680 employees in a company collected by Guangdong Electric Power Research Institute and mental health data collected in the form of questionnaires are selected as experimental data. The basic information is shown in Table 2 [26].

4.2.2. Data Preprocessing. To avoid the impact of invalid data on model performance, this experiment deletes and processes the invalid data of some missing values contained in the data set and finally obtains 672 valid data samples. In addition, considering that there may be noise and background sound in the acquisition process of facial, voice, and gait data, the video and audio data are preprocessed, respectively.

For the video data, GaussianBlur function is called to denoise the video data, and then short time series video is generated by resampling, so as to improve the proportion of effective information. For facial video data, a video with a duration of 30 s is used as a segment; for gait video data, a video with a duration of 8 s is taken as a segment [27, 28]. The preprocessed face and gait are F and G, respectively, so the video data can be expressed as

$$\begin{aligned} F &= \left\{f_1, f_2, \ldots, f_t\right\}, \\ G &= \left\{g_1, g_2, \ldots, g_t\right\}, \end{aligned} \tag{18}$$

where t is determined by the size of video frames and f_t and g_t are single-frame facial and gait images.

For audio data, the first is to delete incomplete recording data; the second is to call wiener filtering method in Wiener function to denoise data; thus the random noise in audio is eliminated; finally, the voice is divided into multiple sequence combinations within 1 s to obtain audio set V, which can be expressed as

$$V = \left\{v_1, v_2, \ldots, v_t\right\}, \tag{19}$$

where t is determined by the length of audio and voice; v_t is the data storage format; and the storage format of v_t in this experiment is matrix storage.

Through the above preprocessing, a total of 658 valid data samples are obtained in this experiment.

4.3. Parameter Settings. In this experiment, parameters of SVM are set as follows: the kernel function is Gaussian function; degree = 3; and the penalty coefficient of error term is 1.

Table 2: Basic information of experimental data.

Information	Project	Number
Gender	Male	641
	Female	39
Age	Under 30 years old	114
	31 ~ 40 years old	146
	41 ~ 50 years old	226
	Over 51 years of age	194
Education level	Primary school or below	7
	Junior high school	44
	High school or technical secondary school	127
	Junior college	155
	Undergraduate	284
	Master's degree or above	63
Seniority	1 ~ 5 years	107
	6 ~ 10 years	96
	10 ~ 15 years	52
	16 ~ 20 years	128
	20 ~ 30 years	288
	More than 30 years	9

4.4. Evaluation Indicators. Accuracy, recall, precise, and $F1$ values are usually selected as indicators for model performance evaluation. The calculation methods are as follows:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{presion} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ F1 &= 2 \times \frac{\text{precise} \times \text{recall}}{\text{precise} + \text{recall}}, \end{aligned} \tag{20}$$

where TP and TN represent true positive cases and true negative cases and FP and FN represent false positive cases and false negative cases. According to the formulas, the higher the accuracy, accuracy, and recall, the better the model performance.

However, the accuracy and recall cannot grow at the same time. To balance the two, $F1$ value index is proposed. The higher $F1$ value indicates that the accuracy and recall are most balanced. Based on the above analysis, accuracy and $F1$ values are finally selected as indicators to evaluate the performance of model.

4.5. Experimental Results

4.5.1. Method Verification. To verify the effectiveness of proposed method, the preprocessed data are used to verify the proposed method, and identification accuracy of somatization, depression, anxiety, and other mental health indicators is used as evaluation criteria. Figures 5~7 show the recognition results based on the single mode of face, voice, and gait; Figures 8~10 show the recognition results of face + voice, face + gait, and voice + gait; Figure 11 shows the multimode recognition results of face + voice + gait; Figure 12 shows the recognition results of introducing attention mechanism based on Figure 11.

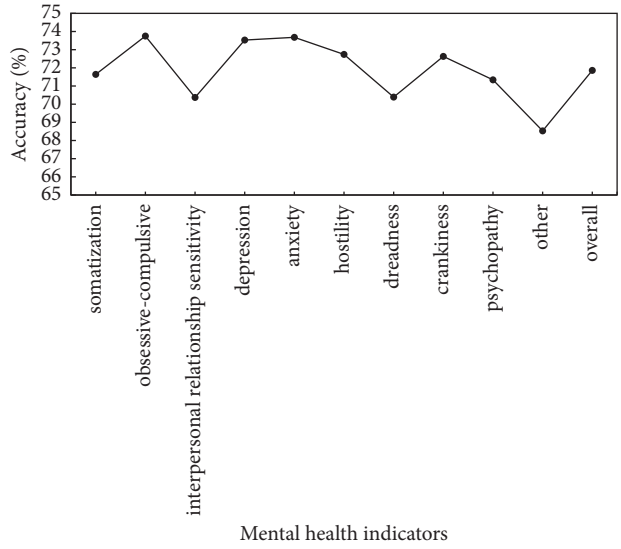

Figure 5: Recognition results based on facial single mode.

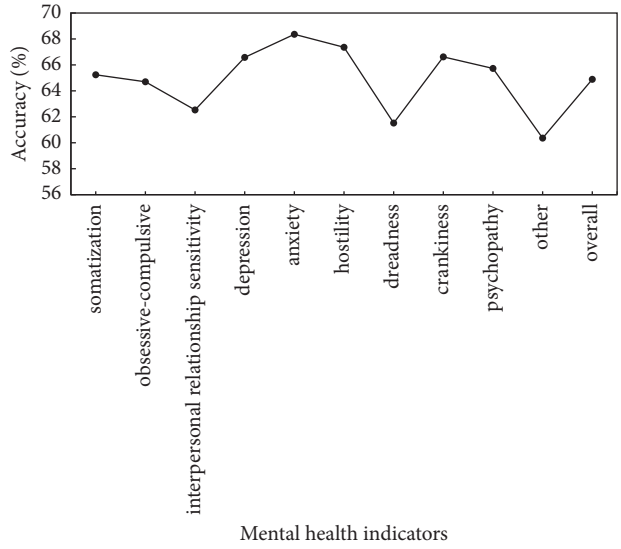

Figure 6: Recognition results based on voice single mode.

As can be seen from Figure 5, the overall recognition accuracy of recognition method based on facial single mode for all mental health indicators is 71.86%, among which the recognition accuracy of obsessive-compulsive and anxiety indicators is higher, reaching 73.75% and 73.53%, respectively. The recognition accuracy of other indicators is the lowest, reaching 68.53%. In addition, the overall $F1$ value of the method is 0.71.

As can be seen from Figure 6, the overall recognition accuracy of recognition method based on voice single mode for all mental health indicators is 64.89%. Among them, the recognition accuracy of anxiety and hostility indicators is relatively high, reaching 68.36% and 67.36%, respectively.

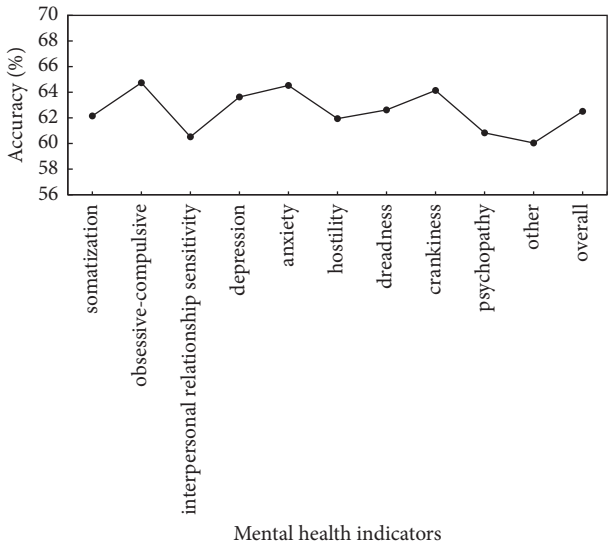

Figure 7: Recognition results based on gait single mode.

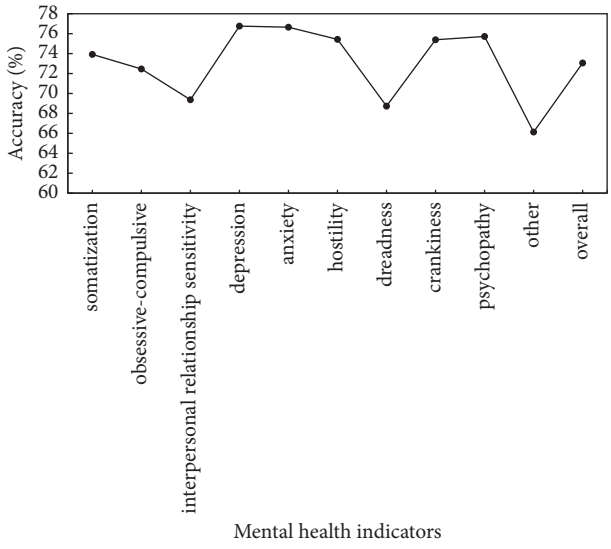

Figure 8: Recognition results based on face + speech dual-modal fusion method.

The recognition accuracy of other and dreadness indicators is relatively low, reaching 60.36% and 61.52%, respectively. In addition, the overall $F1$ value of the method is 0.65.

As can be seen from Figure 7, the overall recognition accuracy of recognition method based on gait single mode for all mental health indicators is 62.51%. Among them, the recognition accuracy of obsessive-compulsive and anxiety indicators is relatively high, reaching 64.74% and 64.53%, respectively. However, the recognition accuracy of other indicators, interpersonal relationship sensitivity, and psychopathy is relatively low, reaching 60.04%, 60.52%, and 60.83%, respectively. In addition, the overall $F1$ value of the method is 0.60.

As can be seen from Figure 8, the recognition accuracy of recognition method based on the face + voice dual-modal fusion for mental health indicators is high, and the overall recognition accuracy is 73.06%. Among them, the recognition accuracy of depression and anxiety indicators is higher, reaching 76.76% and 76.65%, respectively. The recognition accuracy rate of other and dreadness indicators is the lowest, reaching 66.13% and 68.73%, respectively. In addition, the overall $F1$ value of the method is 0.74. Compared with the recognition method based on single mode, the average recognition accuracy is improved by 1.42%, and the average $F1$ value is increased by 11.71%.

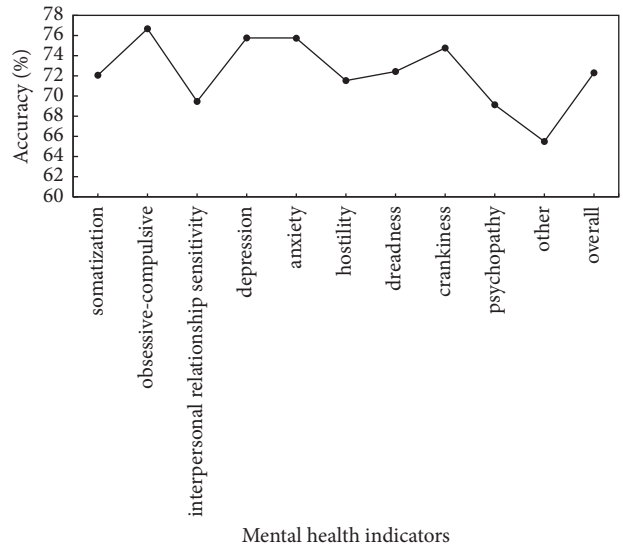

Figure 9: Recognition results based on face + gait dual-modal fusion method.

As can be seen from Figure 9, the recognition accuracy of recognition method based on the face + gait dual-modal fusion for mental health indicators is high, and the overall recognition accuracy is 72.30%. Among them, the recognition accuracy of obsessive-compulsive is high, reaching 76.67%, while the recognition accuracy of other indicators is low, reaching 65.49%. In addition, the overall $F1$ value of the method is 0.71.

As can be seen from Figure 10, the overall recognition accuracy of recognition method based on voice + gait dual-modal fusion for mental health indicators is 64.94%. Among them, the recognition accuracy of anxiety and crankiness is 69.48% and 68.53%, respectively, while the recognition accuracy of other indicators is 58.64%. In addition, the overall $F1$ value of the method is 0.66.

As can be seen from Figure 11, compared with the recognition methods based on single-modal and multimodal fusion, the overall recognition accuracy of recognition method based on facial + speech + gait multimodal fusion for various mental health indicators is improved to different degrees, reaching 73.49%. Among them, the recognition accuracy of anxiety reaches 78.72%, and the recognition accuracy of other indicators is lower, reaching 64.18%. In addition, the overall $F1$ value of the method is 0.75.

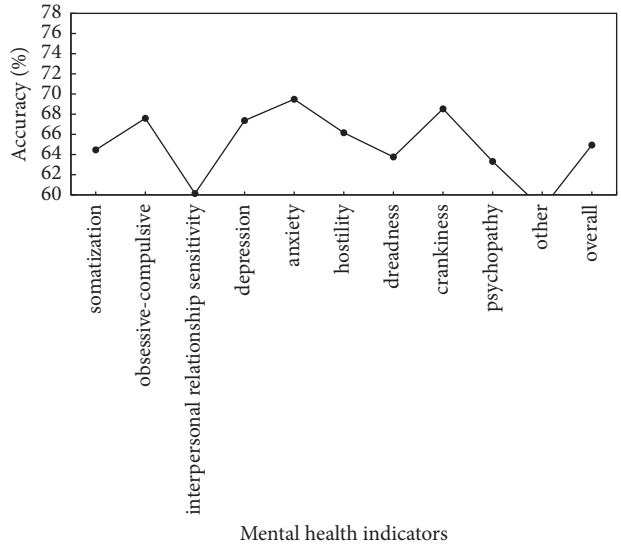

Figure 10: Recognition results based on voice + gait dual-modal fusion method.

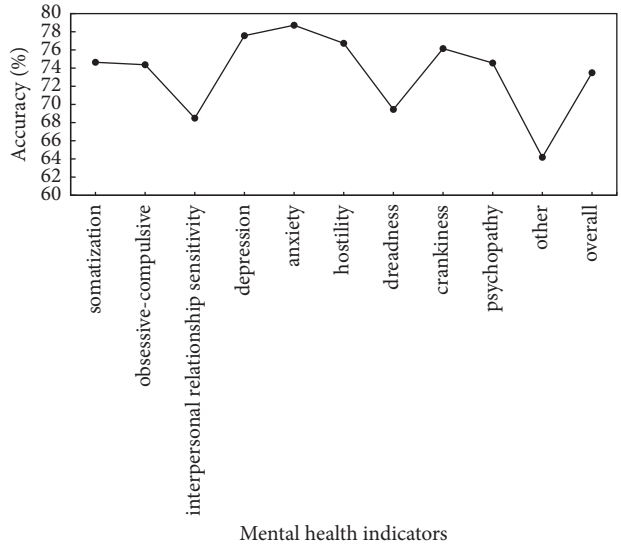

Figure 11: Recognition results based on facial + speech + gait multimodal fusion method.

As can be seen from Figure 12, the recognition method of face + voice + gait multimodal fusion with attention mechanism can accurately evaluate mental health. The recognition accuracy of anxiety and hostility can reach more than 80%, and the recognition accuracy of somatization, depression, and psychopathy can reach more than 79.3%. The overall recognition accuracy of mental health indicators is 77.20%. In addition, the overall $F1$ value of the method reaches 0.77.

In conclusion, the proposed method of mental health assessment can effectively improve the recognition accuracy of mental health indicators and $F1$ value by fusing face, voice, and gait. In addition, attention mechanism is introduced. Compared with recognition method based on single-modal, double-modal, and multimodal fusion, the proposed method of mental health assessment has better recognition effect and has certain effectiveness.

4.5.2. Comparison of Methods. To further verify the effectiveness and superiority of proposed method, the evaluation effect is compared with that of the commonly used mental health assessment method. The recognition accuracy of different methods is shown in Figure 13(a), and the $F1$ value is shown in Figure 13(b). In the figures, F, V, and G are single-modal recognition methods based on face, voice, and gait, respectively. $F + V$, $F + G$, and $V + G$ are dual-modal fusion recognition methods based on face + voice, face + gait, and voice + gait, respectively. $F + V + G$ is face + voice + gait multimodal fusion recognition method. $(F + V + G)$ attention is a multimodal fusion recognition method that introduces attention mechanism. Figure 13(a) shows that the recognition accuracy of multimodal fusion method is higher than that of single-modal and dual-modal fusion recognition methods. The proposed multimodal fusion method with attention mechanism has the highest recognition accuracy,

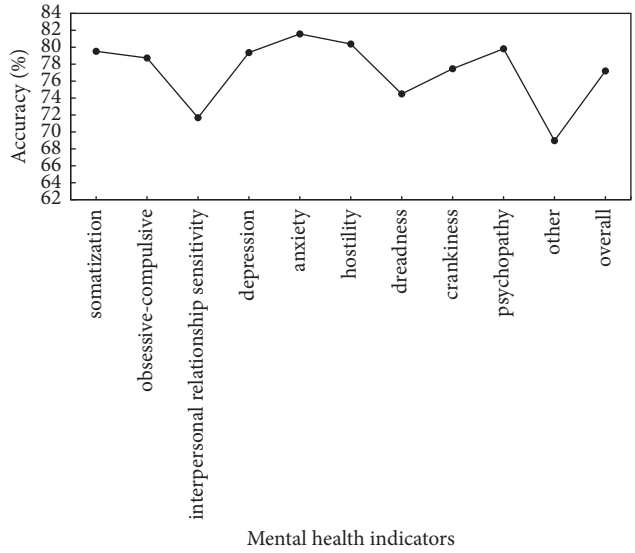

Figure 12: Recognition results of face + speech + gait multimodal fusion method with attention mechanism.

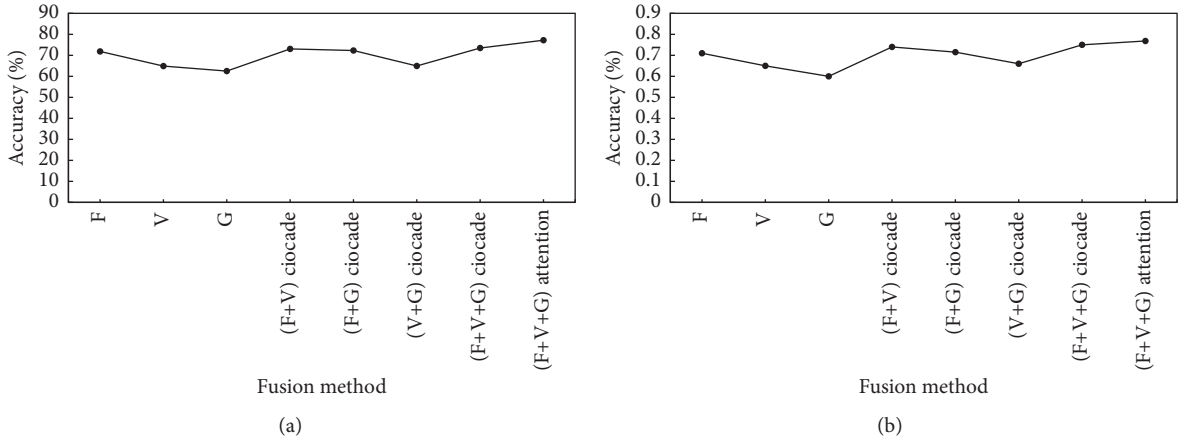

Figure 13: Results of different methods to assess mental health. (a) Comparison of recognition accuracy. (b) Comparison of $F1$ values.

reaching 77.2%. As can be seen from Figure 13(b), the proposed multimodal fusion recognition method based on the attention mechanism has the highest $F1$ value, reaching 0.77. Compared with the recognition method based on face single-modal fusion, $F1$ value is increased by 9.10%. Compared with the recognition method based on dual-modal fusion, the average $F1$ value is improved by 11.53%. Compared with the multimode fusion recognition method without attention mechanism, the $F1$ value is improved by 2.60%. The experimental results show that the proposed method has certain effectiveness and superiority in mental health assessment and solves the problem of insufficient information based on single mode and double mode. Meanwhile, the attention mechanism is introduced to reasonably allocate the weight of face, voice, and gait modes and improve the model performance. Compared with the recognition method based on single-modal fusion and dual-modal fusion, and the multimodal fusion recognition method without attention mechanism, the recognition accuracy and $F1$ value of the proposed method are improved to varying degrees, and the recognition effect is better.

5. Conclusion

To sum up, the proposed mental health assessment method based on convolution neural network can realize effective identification and evaluation of somatization, depression, anxiety, and other mental health indicators, where modal characteristics of face, voice, and gait are fused. In addition, attention mechanism is introduced to allocate different modal weights. The overall accuracy can reach 77.20%, and $F1$ value can reach 0.77. Compared with the recognition methods based on face single-modal fusion, face + voice dual-modal fusion, and face + voice + gait multimode fusion, the recognition accuracy and $F1$ value of the proposed method are improved to varying degrees, and the recognition effect is better, which has certain practical application value. However, due to the limitation of conditions, there are

still some deficiencies to be improved, mainly focusing on the construction of data set. At present, there are few data sets about mental health in China, and the size of data sets has a great influence on mental health assessment model, so the number of data sets selected in this paper is far from the requirements. Therefore, it is necessary to build a mental health database with large amount of data and high quality. The next step is collect more original data to enhance the model performance and improve the recognition accuracy of model.

Data Availability

The experimental data used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The author declares no conflicts of interest regarding this work.

References

[1] H. Briggs, S. Clarke, and N. Rees, "Mental health assessment and triage in an ambulance clinical contact centre," *Journal of Paramedic Practice*, vol. 13, no. 5, pp. 196–203, 2021.

[2] A. Scelzo, "Importance of good mental health assessment to promote healthy aging," *International Psychogeriatrics*, vol. 23, pp. 1–5, 2021.

[3] R. Michael, "Hass and zack maupin and michael doria. Case conceptualization as an alternative to educationally related mental health assessments," *Contemporary School Psychology*, vol. 15, pp. 1–9, 2021.

[4] L. R. Fortuna, "2.2 pros and cons: eliciting the trauma narrative during the asylum mental health evaluation," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 59, no. 10, pp. S3–S4, 2020.

[5] A. Scott and Ph.D. Bresler, "Review of: forensic mental health evaluations in the digital age: a practitioner's guide to using internet-based data," *Journal of Forensic Sciences*, vol. 65, no. 5, pp. 1799–1802, 2020.

[6] H. Masakazu, N. Mitsuteru, S. Shuji et al., "Effectiveness of a voice-based mental health evaluation system for mobile devices prospective study," *JMIR formative research*, vol. 4, no. 7, Article ID e16455, 2020.

[7] J. J. Newson, D. Hunter, and T. C. Thiagarajan, "The heterogeneity of mental health assessment," *Frontiers in Psychiatry*, vol. 11, pp. 76–86, 2020.

[8] M. Tom, K. Khalid, and N. L. Jessica, "Parents' constructions of normality and pathology in child mental health assessments," *Sociology of Health & Illness*, vol. 42, no. 3, pp. 544–564, 2020.

[9] H. Martin, P. Caroline, S. Q. Molly, and B. Sherry, "Computer-assisted psychological assessment and psychotherapy for collegians," *Journal of Clinical Psychology*, vol. 76, no. 6, pp. 952–972, 2020.

[10] A. M. G. Poorthuis and A. v. Dijk, "Online study-aids to stimulate effective learning in an undergraduate psychological assessment course," *Psychology Learning and Teaching*, vol. 20, no. 2, pp. 236–249, 2021.

[11] U. Granziol, A. Brancaccio, G. Pizziconi et al., "On the implementation of computerized adaptive observations for psychological assessment," *Assessment*, vol. 29, no. 2, pp. 225–241, 2020.

[12] A. M. Mayworm, B. M. Kelly, M. T. Duong, and A. R. Lyon, "Middle and high school student perspectives on digitally-delivered mental health assessments and measurement feedback systems," *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 47, no. 4, pp. 531–544, 2020.

[13] H. Zhang, H. Xu, X. Tian, J. Jiang, and J Ma, "Image fusion meets deep learning: a survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021.

[14] S. Chebbi and S. Ben Jebara, "Deception detection using multimodal fusion approaches[J]," *Multimedia Tools and Applications*, vol. 45, pp. 1–30, 2021.

[15] M. S. S. Syed, E. Pirogova, and M. Lech, "Prediction of public trust in politicians using a multimodal fusion approach," *Electronics*, vol. 10, no. 11, pp. 1259–1269, 2021.

[16] X. Zhang, Z. Li, X Gao, D. Jin, and J. Li, "Channel attention in LiDAR-camera fusion for lane line segmentation," *Pattern Recognition*, vol. 118, pp. 108020–108025, 2021.

[17] K. Huang, W. Zhou, and M. Fang, "Deep multimodal fusion autoencoder for saliency prediction of RGB-D images," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–10, Article ID 6610997, 2021.

[18] W. Hu, X. Meng, Y. Bai et al., "Interpretable multimodal fusion networks reveal mechanisms of brain cognition," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1474–1483, 2021.

[19] R. J. Deligani, S. B. Borgheai, J. McLinden, and Y. Shahriari, "Multimodal fusion of EEG-fNIRS: a mutual information-based hybrid classification framework," *Biomedical Optics Express*, vol. 12, no. 3, pp. 1635–1650, 2021.

[20] X. Liu and Y. Jiang, "Aesthetic assessment of website design based on multimodal fusion," *Future Generation Computer Systems*, vol. 117, pp. 433–438, 2021.

[21] Y. Suparat and C. Kosin, "3D point-of-intention determination using a multimodal fusion of hand pointing and eye gaze for a 3D display," *Sensors*, vol. 21, no. 4, pp. 1155–1165, 2021.

[22] Y. Wu and L. Huang, "An intelligent method of data integrity detection based on multi-modality fusion convolutional neural network in industrial control network," *Measurement*, vol. 175, pp. 109013–109015, 2021.

[23] W. Zhang, J. Yu, Y. Wang, and W. Wang, "Multimodal deep fusion for image question answering," *Knowledge-Based Systems*, vol. 212, pp. 106639–106649, 2021.

[24] N. Phong Thanh, V. Dang Bich Huynh, K. Dang Vo, P. Thanh Phan, M. Elhoseny, and D. N. Le, "Deep learning based optimal multimodal fusion framework for intrusion detection systems for healthcare data," *Computers, Materials & Continua*, vol. 66, no. 3, pp. 2555–2571, 2021.

[25] X. Lv, "Chinese description generation of dual attention images based on multi-modal fusion," *Journal of Physics: Conference Series*, vol. 1735, no. 1, pp. 012004–012006, 2021.

[26] I. Hupont, E. Cerezo, S. Ballano, and S. Baldassarri, "On the origin of the methodology for the scalable fusion of affective channels in a continuous emotional space and the "emotional kinematics" filtering technique-a correction," *Information Fusion*, vol. 67, pp. 1-2, 2021.

[27] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," *Information Fusion*, vol. 65, pp. 58–71, 2021.

[28] C. Che, H. Wang, X. Ni, and R. Lin, "Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis," *Measurement*, vol. 41, pp. 108655–108665, 2020.