IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INNOVATION AND APPLICATION OF INTELLIGENT PROCESSING OF DATA, INFORMATION AND KNOWLEDGE AS RESOURCES IN EDGE COMPUTING

Received December 2, 2019, accepted December 11, 2019, date of publication December 20, 2019, date of current version January 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961139

Intelligent Emotion Detection Method Based on Deep Learning in Medical and Health Data

JIANQIANG XU[1], ZHUJIAO HU[2], JUNZHONG ZOU[1], AND ANQI BI[3]

[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
[2]School of Microelectronics, Fudan University, Shanghai 201203, China
[3]School of Computer Science and Engineering, Changshu Institute of Technology, Suzhou 215500, China

Corresponding author: Junzhong Zou (zoujunzhongecust@126.com)

This work was supported in part by the Fund of Minhang District Human Resources and Social Security Bureau, in part by the Wireless Intelligent Handheld Terminal Based on RFID Technology under Grant 11C26213100798 and in part by the RFID intelligent handheld mobile terminal and solution for food and drug traceability system under Grant 1401H122500, in part by the Jiangsu University Natural Science Research Project under Grant 18kjb5200001, and in part by the Project Fund of Shanghai Economic and Information Commission and the application of artificial intelligence in new retail.

ABSTRACT Emotional abnormality may be brought out by physiological fatigue. In order to solve the problem, an emotion detection method based on deep learning in medical and health data is proposed in this paper. First of all, the related content of emotional fatigue is studied. The concept and the classification of emotional fatigue are introduced. Then, a multi-modal data emotional fatigue detection system is designed. In the system, multi-channel convolutional aotoencoder neural network is used to extract electrocardiograms (ECG) data features and emotional text features for emotional fatigue detection. Secondly, the network structure of learning ECG features by multi-channel convolutional aotoencoder model is introduced in detail. And the network structure of learning emotional text features by convolutional aotoencoder model is also described in detail. Finally, multi-modal data features are combined for emotional detection. It is shown by the experimental results that the proposed model has an average accuracy of more than 85% in predicting emotional fatigue.

INDEX TERMS Emotion detection model, multi-channel convolutional aotoencoder (MCAE), medical health, deep learning, emotional text features, intelligent data analysis.

I. INTRODUCTION

In today's society, the pace of people's life is speeding up. With the stress on entering school, employment, work, and family, people may suffer from strong mental pressure for a long time. And psychological problems are resulted. Psychological abnormality is different from physical diseases, and it may have influence on people's mood or character. When people's emotions are in a state of fluctuation, over tension, depression or pessimism for a long time, psychological diseases may occur. The incidence of mental illness is increasing year by year. One of the main reasons is that people can't detect and deal with negative emotions very well. So, it's very important to detect and release emotions in time. Fatigue is the feeling of physical weakness or lack of vitality. It also refers to sleepiness, exhaustion, drowsiness and mental malaise, including physical and mental fatigue or weakness [1]–[3].

The associate editor coordinating the review of this manuscript and approving it for publication was Honghao Gao.

Although physical fatigue and psychological fatigue are two forms of human fatigue, there are many differences between them. In reality, these two states often appear simultaneously. When people feel fatigued, whether it is fatigue or sleepiness caused by depression, it is necessary to determine the type of fatigue and take corresponding mitigation measures. Physical fatigue can be relieved by rest. While for psychological fatigue, it needs to find and judge the causes, and take targeted mitigation measures [4], [5]. In literature [6], AIWAC system is proposed to serve specific groups of people. For empty nest elderly, the physiological information is collected, the physiological state is perceived, the negative emotional, such as loneliness is eliminated. For the people with autism, their abnormal psychological state can be perceived to help them out of the shadow of social phobia.

Medical and health data is a multi-modal and complex data with continuous and rapid growth. It contains rich and diverse information. Challenges related to medical and health data are as follows: how to collect and obtain medical and health

data quickly and accurately, how to use high-speed network to transmit medical and health data reliably and efficiently, and how to mine useful information from big data of health care with machine learning and deep learning related to artificial intelligence and further develop intelligent applications for medical staff and ordinary people. Aiming at the problems related to the analysis of health care data and the development of intelligent application, the risk assessment of chronic diseases that have great impact on people's health is studied in this paper. The characteristics of health care big data are analyzed, the corresponding data feature learning model is designed, and the patients are classifed according to whether they have high-risk disease or not. Because of the diversity and complexity of data types in intelligent health care applications, it is necessary to build flexible and diverse network architecture and provide safe data transmission services.

It is proposed to establish a multimodal emotional fatigue detection system in this paper. Both emotional text features and ECG data features are used to detect emotional fatigue. In the model, different feature information is combined to form different network input channels. So that the emotional information of input sentences can be learned by the network model from various feature representations in the training process. The importance of each word in the sentence is expressed effectively, and more hidden information is obtained.

II. RELATED RESEARCH

At present, the commonly used emotion detection methods include facial expression analysis, physiological information analysis, text analysis, voice analysis and body language analysis. Human physiological information is also an important way of emotion recognition. In literature [7], how to construct emotion recognition system and its main components with physiological signals is introduced. The related concepts and existing problems are discussed. In literature [8], emotion detection is based on the electrocardiograms (ECG) signal. And an empirical mode decomposition (EMD) solution for dynamic monitoring of emotion mode is proposed. In each mode, the classification features based on the instantaneous frequency (Hilbert Huang Transform) and the local oscillation are used. In literature [9], a new method based on ECG and pulse wave is proposed to measure positive and negative emotions in real time. In literature [10], the feature selection problem of emotion recognition based on ECG signals is studied. And different features for emotion recognition are selected through variance analysis and heuristic search. In literature [11], six emotion states are identified based on QRS signals in ECG signals. And two nonlinear characteristic Hurst Exponent [14] calculation methods, rescaled range statistics (RRS) [12] and finite variance scaling(FVS) [13] are used.

However, the relationship between emotion and people's fatigue is not explored by these emotional tests. And the different emotional roots of users are not distinguished. So, it is difficult to release emotion in a more targeted way. To solve the problem, an emotion detection method based on multi-channel convolutional neural network in medical and health data is proposed.

III. DEEP LEARNING MODEL FOR EMOTIONAL FATIGUE DETECTION

A. TYPER OF EMOTIONAL FATIGUE

When users feel tired, it is easy to lead to emotional abnormality. And the work efficiency may be reduced because of the emotional state. At this time, users are considered to be in the emotional fatigue state. Analyzing the causes of users' emotional fatigue, they are divided into three types, as shown in Figure 1.

Type 1: physiological fatigue, mainly refers to the fatigue caused by physical overdraft.

Type 2: repetitive fatigue, fatigue caused by repeated monotonous actions for a long time. For example, if a driver is driving on a highway for a long time, the driver's actions, road conditions, and visual stimuli received by the driver are single. And traffic accident caused by emotional fatigue may occur.

Type 3: environment fatigue, work in closed space for a long time. For example, the long-time underwater monitoring, scientific experiment and equipment maintenance. And the computer worker who sits in front of the computer in the room for a long time.

For type 1, the best way to deal with it is to take a rest immediately. However, for type 2 and type 3, although users' emotional state is very poor, their physical strength is still in good condition. Corresponding ''emotional feedback'' can be used to change users' emotional state and improve their interest and work efficiency.

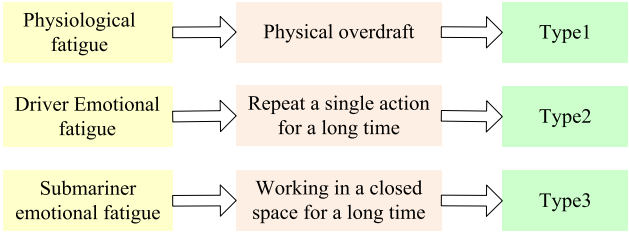


FIGURE 1. Three types of emotional fatigue.

B. CONVOLUTIONAL AOTOENCODER NEURAL NETWORK

The reasons for choosing convolution neural network to extract the features of emotional text and ECG data are as follows: (1) CNN has been widely used in computer vision and other fields. CNN has a strong ability to process images and time series signals. (2) Compared with the traditional shallow learning algorithm, deep learning algorithm has the ability of data-driven learning features. And it can better learn the feature representation of the original data [15]. In order to extract better features and enhance the fitting effect of nonlinear functions, the method of increasing the number of layers and neurons of neural network can be used [16].

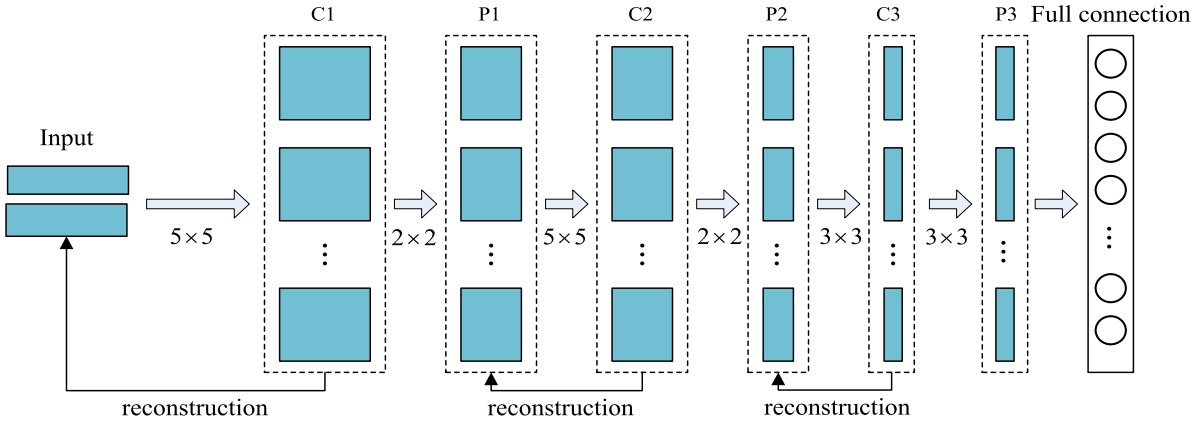


FIGURE 2. Schematic diagram of convolutional aotoencoder neural network.

Convolutional aotoencoder neural network (CAE) is a neural network structure which combines convolutional neural network and aotoencoder operation. Aotoencoder is used to learn and determine the network parameters of the convolution layer. Greedy layer-wise unsupervised training is taken by the parameters of each layer [17], [18]. The convolutional aotoencoder neural network for feature extraction of emotional text is shown in Figure 2. In the network structure, three layers are used to perform convolutional operation, three layers are used to perform pooling operation, and one layer is used to complete the full connection.

1) CONVOLUTIONAL AOTOENCODER NEURAL NETWORK

For convolutional aotoencoder layer, convolutional operation is used to obtain the characteristic representation of input data. And the optimized network parameters are determined through greedy unsupervised training. Encoding operation and decoding operation are included in convolutional aotoencoder layer.

In the encoding operation, formula (1) is used for input x to obtain the output value of the convolution. Among them, n is the number of characteristic graphs of input data, w is the connection weight, i.e. convolution kernel, w_i is the i-th convolution kernel, b_i is the deviation ($i = 1, 2, \cdots m$), and $\sigma(x) = 1/\left(1 + e^{-x}\right)$ is the activation function.

$$y_i = \sigma\left(\sum_{j=1}^{n} w_{ij} \cdot x_j + b_i\right) \tag{1}$$

The decoding operation is the reverse operation of the encoding operation. The output y obtained by the encoding operation is used as the input data, and the reconstruction input is $\tilde{x}$. The formula used by the decoding operation is as follows:

$$\tilde{x}_j = \phi\left(y_t \cdot \tilde{w}_{ij} + \tilde{b}_j\right) \tag{2}$$

Where, $\tilde{b}$ represents the deviation, and the same deviation is used in each input characteristic graph of decoding operation. $\tilde{w}$ is the reconstruction weight parameter, and is the transposition of weight w. According to the original input x and the reconstructed input $\tilde{x}$, the reconstruction error can be calculated. The mean square error is used in the cost function of reconstruction error.

$$E(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 \tag{3}$$

The gradient descent method is used to update the parameter $\theta = \left\{w_i, \tilde{w}, b, \tilde{b}\right\}$. And the convolution reconstruction operation is repeated. Finally, the minimum cost function value is obtained, and the training is ended. At this time, the parameter θ of the layer CAE is determined.

2) POOLING LAYER

When the convolution feature graph is generated, the pooling operation is used for the down sampling. In general, the average and maximum value can be chosen by the pooling function. The dimension of the convolution layer can be reduced to a large extent by the pooling operation. And the occurrence of over fitting can be avoided.

In convolutional aotoencoder neural network, the maximum pooling operation is used. The sizes of the sampling area of P1, P2 and P3 are 2×2, 2×2 and 3×3, respectively.

3) FULLY CONNECTED LAYER

A fully connected layer is connected to the last pooling layer. The output of the pooling layer is the input of the fully connected layer. The neurons between the two layers are fully connected. The operation is shown in formula (4):

$$x^f = w^f x^p + b^f \tag{4}$$

Where, x^p represents the output value of the pooling operation, x^f represents the output of the fully connected layer, b^f represents the deviation, and w^f represents the weight.

C. MULTICHANNEL CONVOLUTIONAL AOTOENCODER NEURAL NETWORK

A large number of data is used by the data-driven feature learning method. And the representation method of original

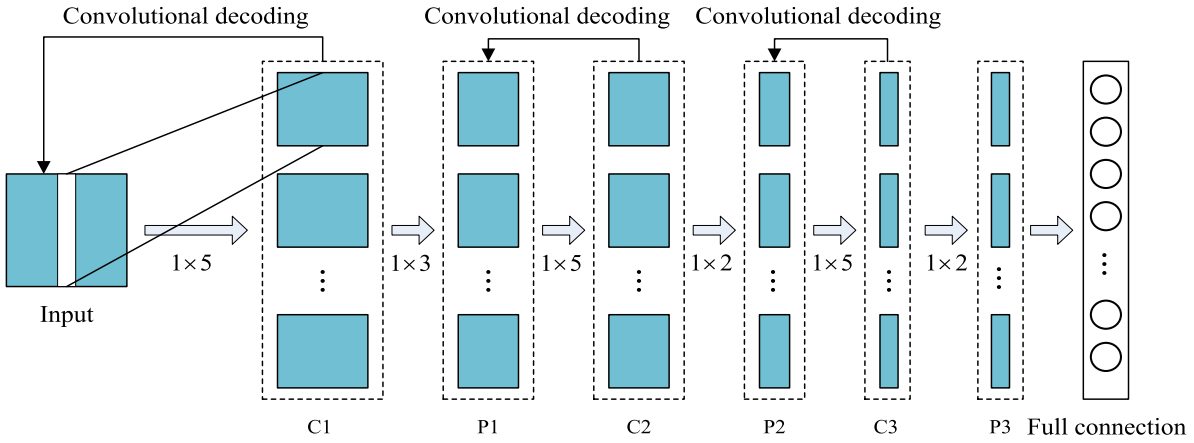


FIGURE 3. Schematic diagram of ECG-MCAE multichannel convolutional aotoencoder neural network.

data features are finally obtained. Put it simply, it "let the data speak for itself." The feature learning method using depth model is called "data-driven feature learning," because the biggest advantage of it is the ability to obtain features directly from data. It is shown by more and more studies that the features obtained by using depth learning method through data-driven learning are better than those obtained by artificial design.

ECG is time series data. Gathering relevant information and eliminating irrelevant information are the traditional methods of artificial feature extraction. The common method is to use wavelet transform to achieve good results. Wavelet transform is an integral transform based on the main information of extraction. Due to the ignorance of details, there are limitations of wavelet transform for feature extraction. In order to avoid the limitations of artificial design features, a multi-channel convolutional aotoencoder neural network structure is proposed.

ECG signals are obtained from three sensors of wearable devices. The collected data is three channel time series data. Traditional CNN method can not be used directly. The multi-channel time series data operation method used in the multi-channel convolutional neural network is applied to the convolutional aotoencoder neural network. And the multi-channel convolutional aotoencoder (MCAE) is designed to extract ECG features, referred to as ECG-MCAE, as shown in Figure 3. Time dimension of ECG data is considered by the processing unit, and the same convolution kernel is used by the multi-channel [19], [20]. For ECG-MCAE, the method of unsupervised learning network parameters of aotoencoder is used for reference. And convolution layer parameters of multi-channel convolutional neural network are trained by greedy layer-wise unsupervised training method [21, [22]. Convolutional aotoencoder operation and pooling operation with time dimension in ECG-MCAE are defined as follows.

1) MULTICHANNEL CONVOLUTIONAL AOTOENCODER

The convolutional aotoencoder operation with time characteristics is divided into two parts: convolutional coding and convolutional decoding. In convolutional coding operation, the input data is convoluted to obtain the corresponding output. In the convolutional decoding operation, the reconstruction of the input is completed.

Convolutional coding: input data x and output data y of current layer, convolutional coding operation represents the mapping from input x to data x, as shown in formula (5).

$$y_i^k = \sigma\left(\sum_{j=1}^{n} w_{ij} \cdot x_j^k + b_i\right) \tag{5}$$

Where, $y_i^k\ (i = 1, 2, \cdots, m, k = 1, 2, 3)$ represents the k-th row of the i-th output characteristic graph. $x_j^k (i = 1, 2, \cdots, m,\ k = 1, 2, 3)$ represents the k-th row of the j-th input characteristic graph. $w_i\ (i = 1, 2, \cdots, m)$ is the i-th convolution kernel, w is the connection weight between the previous convolution layer neurons and convolution layer neurons. m convolution kernels correspond to m output characteristic graphs, $b_i\ (i = 1, 2, \cdots, m)$ represents deviation. The same deviation value is used in each output characteristic graph, $\sigma\ (x) = \frac{1}{1+e^{-x}}$ is the activation function.

Convolutional decoding: the input data is the output y of the convolutional encoding operation, and the output $\tilde{x}$ is the input of the reconstructed convolutional encoding. The process is called convolutional decoding, as shown in formula (6).

$$\tilde{x}_j^k = \phi\left(y_i^k \cdot \tilde{w}_{ij} + \tilde{b}_j\right) \tag{6}$$

Where, y_i^k represents the k-th row of the i-th output characteristic graph, $i = 1, 2, \cdots m,\ k = 1, 2, 3$. $\tilde{x}_j^k$ represents the k-th row of the j-th input characteristic graph, $j = 1, 2, \cdots, n, k = 1, 2, 3$. $\tilde{w}_i$ represents the weight of convolution decoding, and is the transposition of the i-th convolution kernel. $\tilde{b}_i$ represents the deviation, and each characteristic graph uses a deviation value. $\phi\ (x)$ is the activation function, using the same function as $\sigma\ (x)$.

At this time, according to the original input x and the reconstructed input $\tilde{x}$, the reconstruction error can be calculated. The mean square error is used in the cost function of

reconstruction error.

$$E(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 \tag{7}$$

Where, i represents the number of channels of the input data. The gradient descent method is used to update parameter $\theta = \left\{w, \tilde{w}, b, \tilde{b}\right\}$. The convolution reconstruction operation is repeated. Finally, the minimum cost function value is obtained, and parameter θ is determined.

2) MULTICHANNEL POOLING

When a convolution layer completes the convolution operation, the output characteristic graph is used as the input of the pooling layer. In the pooling layer, the input characteristic graph is pooled. In general, the pooling function can choose the average or maximum.

Considering the characteristic of time series of ECG data, the average pooling operation is used in ECG-MCAE. And the pooling operation formula is designed as follows:

$$y_i^k = \underset{1\prec d\prec D}{average}\left(x_i^{k,l+d}\right) \tag{8}$$

Where, x_i represents the i-th input characteristic graph, y_i represents i-th output characteristic graph, D represents the size of the pooling area, k represents the number of rows in the characteristic graph, and l represents the number of columns in the characteristic graph. The pooling layer only performs the average pooling operation on the input data, and the number of characteristic graphs after the data passes through the current layer does not change.

3) FULLY CONNECTED LAYER

The fully connected layer is set after the last pooling layer. The neurons of the two layers are fully connected. In formula (8), x^p represents the output of the pooling layer, x^f represents the value of the fully connected layer, b^f represents the deviation, and w^f represents weight.

$$x^f = w^f x^p + b^f \tag{9}$$

D. MULTI MODAL DATA EMOTIONAL FATIGUE DETECTION

In the emotional fatigue detection system, the input value is defined as the user's data value $X = (x_1x_2, \cdots x_n)$. The extracted feature is expressed as $F = (f_1f_2, \cdots f_m)$. And the output value is c, $c \in C = \{c_i \,|i = 0, 1, 2, 3\}$ which is used to indicate whether the user is in the emotional fatigue state or not. c_0 is defined as that the user is not in emotional fatigue state. c_1 indicates that the user is in physiological fatigue Type 1 state, c_2 indicates that the user is in Type 2 state, and c_3 indicates that the user is in Type 3 state.

1) FEATURE EXTRACTION OF EMOTIONAL TEXT

Although the way of recording and language expression of emotional texts are special, they also have the commonness of texts. When analyzing and understanding emotional text data, the text analysis method can be used. And the particularity of emotional text is considered. To classify or deal with the texts data, when the natural language processing method is used, it is necessary to find the effective feature representation of input text data by numbers.

In this paper, the most important emotion feature in emotion classification task is added to explain the method of combining convolutional neural network and other features. Based on the features of the input text content, the input matrix of the network model is constructed by combining the position features of the words in the sentence, and the features of the emotional words of the emotional analysis features. Different input channels are used to receive the combination of different feature information. So that more abundant emotional feature information can be learned by the model during the training process, and the emotional polarity of the short text sentence can be identified effectively.

The feature information of input sentences can be learned by neural networks through receiving the vectorized input of texts. In the task of text classification, the most important feature information of sentences is implied in the words of the sentences. In this paper, words are used to represent sentences. By mapping each word into a multi-dimensional continuous value vector, the word vector matrix $\boldsymbol{E} \in R^{m\times|V|}$ of the word set of the whole data set is obtained. Where, m is the vector dimension of each word, $|V|$ is the size of the word set. For sentence $s = \{\omega_1, \omega_2, \cdots \omega_3\}$ of length n, each word ω_i in the sentence can be mapped into an m-dimensional vector, namely $\boldsymbol{e}_i \in R^m$.

In this paper, the common Hownet emotional word set is used to mark the part of speech of the input sentence again, as shown in Table 1. By assigning specific part of speech tagging to the special words in the sentence, the words that play an important role in emotional classification can be fully used by the model. These words include positive and negative emotional words, negative words, and degree adverbs. Therefore, during the training process, the feature information of these words is emphasized.

In addition to the emotional words in the sentence, the negative words and degree adverbs are also re-marked in this paper. For example, ''like'' is a positive emotional word, while ''dislike'' is a negative emotional word. So, with the negative word, the sentence may contain the opposite emotional polarity of the affective word. For different part of speech tagging, each part of speech tagging is mapped into a multi-dimensional continuous value vector $\boldsymbol{tag}_i \in R$ through vectorized operation. Where, $\boldsymbol{tag}_i$ is the i-th part of speech vector and k is the dimension of the part of speech vector. In the process of training, the components of part of speech vector can be adjusted by the network model according to different part of speech tagging. Thus more detailed feature information can be learned [23], [24].

Because of the word limitation of wechat friend circle, the length of wechat text is generally short, and the emotional information contained in sentences is limited. So the position of words in the wechat friend circle is also an important feature of wechat text. The same word that appears in

TABLE 1. Part of speech tagging.

Part of speech	Tagging
Postive Sentiment Words	Pos
Negative Sentiment Words	Neg
Adverbs	Adv
Negative Words	Inver

different places may contain different information. The positional value of the i-th entry ω_i in the sentence s is calculated as follows:

$$p(\omega_i) = i = len(s) + maxlen \tag{10}$$

Where, $p(\omega_i)$ is the positional value of ω_i in the sentence s, i is the position of entry ω in sentence s, $len(s)$ is the length of sentence s, and maxlen is the maximum length of the input sentence. Like part of speech vector operation, each position value is mapped into an l dimensional vector in this paper, i.e. $\boldsymbol{position}_i \in R^l$. Where $\boldsymbol{position}_i$ is the vector of the i-th position value.

In this paper, words are used as units to convolute sentences. For sentences of length n, the features are as follows

$$\boldsymbol{e}_{1:n} = \boldsymbol{e}_1 \oplus \boldsymbol{e}_2 \oplus \cdots \oplus \boldsymbol{e}_n \tag{11}$$

$$\boldsymbol{tag}_{1:n} = \boldsymbol{tag}_1 \oplus \boldsymbol{tag}_2 \oplus \cdots \oplus \boldsymbol{tag}_n \tag{12}$$

Where, $\boldsymbol{e}$ is word vector and $\boldsymbol{tag}$ is part of speech feature. In order to simplify the structure of the network model, a feature matrix $\boldsymbol{x} \in R^{m+k}$ is formed with a simple splicing operation in this paper. And it is taken as the input of the convolution neural network

$$\boldsymbol{x} = \boldsymbol{e} \oplus \boldsymbol{tag} \tag{13}$$

Where, $\oplus$ is splicing operation. In this paper, the specific emotional words are mapped into multi-dimensional part of speech features. So that the network can optimize the classification model by adjusting the part of speech feature components in the training process. In the experiment, a maximum length maxlen is set for the input of the sentence. 0 vectors are used to complete the sentence whose length is less than maxlen. Rich local features can be extracted by the convolution layer through different convolution check input matrix. For the convolution kernel with length of h, the sentence can be divided into $\{\boldsymbol{x}_{0:h-1}, \boldsymbol{x}_{1:h}, \cdots, \boldsymbol{x}_{i:i+h-1}, \cdots, \boldsymbol{x}_{n-h+1:n}\}$. And each component is convoluted to obtain the convolution characteristic graph

$$\boldsymbol{C} = (c_1, c_2, \cdots, c_{n-h+1}) \tag{14}$$

Where, c_i is the information obtained after the convolution of component $\boldsymbol{x}_{i:i+h-1}$.

$$c_i = relu(\boldsymbol{W} \cdot \boldsymbol{x}_{i:i+h-1} + b) \tag{15}$$

Where, $\boldsymbol{W} \in R^{h\times(m+k)}$ is convolution kernel weight and $b \in R$ is bias. In this paper, the max-over-time pooling method is used to sample the feature information and extract the most important feature information:

$$\widehat{c} = \max\{\boldsymbol{C}\} \tag{16}$$

The $\hat{c}$ is the result of a convolution kernel sampling, and the feature information of the convolution kernel sampling can be expressed as

$$\hat{\boldsymbol{C}} = \left(\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_d\right) \tag{17}$$

Then, the feature information obtained from the sampling of the pooling layer is taken as the input of the fully connected layer, and the classification results are obtained:

$$y = soft\max\left(\boldsymbol{W}_f \cdot \hat{\boldsymbol{C}} + b_f\right) \tag{18}$$

Where, $b_f \in R$ is bias, $\boldsymbol{W}_f \in R^d$ is the weight of fully connected layer, and x is output result.

2) ECG FEATURE LEARNING NETWORK STRUCTURE

Based on the defined time dimension convolution and time dimension pooling operation, the ECG-MCAE network structure is built, as shown in Figure 4-2. It includes the following parts:

Input layer:

The method of sliding window is used to extract data from time series ECG data. The size of sliding window is 3×256 and the step size is 128. Two-dimensional matrix $x \in R^{3\times256}$ is used in the input layer data. Three rows correspond to the data of three sensors respectively, and 256 columns represent the number of sampling points intercepted from ECG data by sliding window. At this time, the number of neurons in the input layer is defined as 3256.

Convolution layer:

In the convolution layer C1, 50 convolution kernels with the size of 5×5 are used. Inputting a characteristic graph with a size of 128×128 pixels, and 50 output characteristic graphs with the size of 124×124 are obtained. In convolution layer C2, 40 convolution kernels with the size of 5×5 are used. Inputting 50 characteristic graphs with the size of 62×62, 40 output characteristic graphs with the size of 58×58 are obtained. In the convolution layer C3, 20 convolution kernels with the size of 3×3 are set. Inputting 40 characteristic graphs with the size of 29×29, 20 characteristic graphs with the size of 27×27 are obtained.

Pooling layer:

The maximum pooling operation is used in Face-CAE. The size of pooling area of P1, P2 and P3 pooling layers is 2×2, 2×2 and 3×3, respectively. The number of input characteristic graph and output characteristic graph of P1 layer is 50, and the sizes are 124×124 and 62×62, respectively. The number of input characteristic graph and output characteristic graph of P2 layer is 40, and the sizes are 58×58 and 29×29, respectively. The number of input characteristic

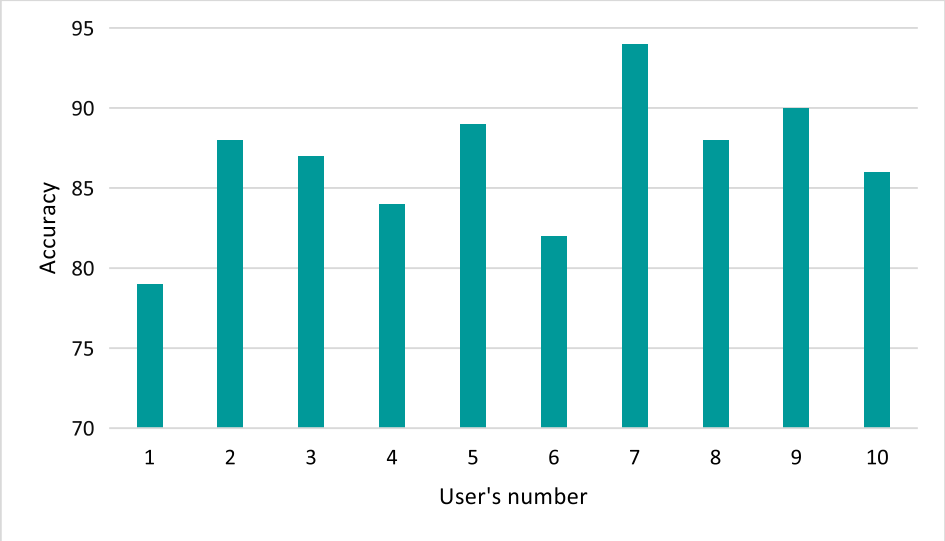


FIGURE 4. Accuracy of emotional fatigue detection in volunteers.

graph and output characteristic graph of P3 layer is 20, and the sizes are 27×27 and 9×9, respectively.

Fully connected layer:

There are 400 neurons in the fully connected layer. All of the neurons are fully connected with the P3 neurons in the last pooling layer. The 400 neurons in the fully connected layer are the feature representations of input emotional text feature data x.

3) MULTIMODAL DATA FUSION

Users' emotions can be better identified by multimodal learning. There are two main methods for multimodal data fusion: feature level fusion and decision level fusion [25], [26]. Feature layer fusion is to fuse all data features into feature vectors, which are taken as classification features. According to the data of each mode, decision level fusion classifies data separately. And then, the classification results are combined linearly, such as the use of average weight [27], [28]. In the emotional fatigue detection system, feature layer fusion method is used to fuse ECG data, emotional text and other features into one feature vector.

4) EMOTIONAL FATIGUE DETECTION MODEL TRAINING

The sample data set used in unsupervised learning is represented as $Z = (z_1, z_2, \cdots, z_n)$. It is composed of ECG data and emotional text feature data. Supervision training includes sample data set $X = (x_1, x_2, \cdots, x_n)$ and corresponding tag set Y. And all kinds of data is included in emotional fatigue detection.

The ECG data and emotional text feature data in Z are trained layer by layer through unsupervised method. After the unsupervised training, $F = (f_1, f_2, \cdots, f_m)$ inputs the fatigue detection module. The fused ECG and emotional text feature representation $F = (f_1, f_2, \cdots, f_m)$ are obtained. After the fusion of features into the last layer, the emotional fatigue of users is classified by Softmax. And the classification result c is obtained. The error between the classification result c and the tag y is calculated. It has the supervised training classifier and the fully connected layer parameters. The connection parameters between each layer of the feature extraction layer are adjusted. Using the data related to the movement operation and the corresponding tags to train the SVM, distinguish the Type 2 and Type 3 states.

After the training, the test set data is input or the user real-time data is collected. And the output emotion classification results are obtained. The identified categories include three categories: Type 1 state, Type 2 state, Type 3 state, and no emotional fatigue state. In case of Type 2 and Type 3 states, the operation data feature is used to classify the output emotional fatigue state.

IV. EXPERIMENT

A. EXPERIMENTAL DATA COLLECTION

We recruited 5 male and 5 female volunteers, 10 in total. The average age of the volunteers is 21.8 years (17-26 years). The data is collected for 10 days. Background service program and mobile App are installed in all volunteers' mobile phones to collect mobile data. ECG data is collected through sensors on wearable devices. The data is input to the mobile phone, and then sent to the back-end cloud platform through the special program on the mobile phone. At the same time, a small sensor network is deployed in the laboratory to obtain environment information. And the information is also transmitted to the back-end cloud platform in real time. The emotional fatigue status is labeled by volunteers, according to their own status. Other categories represent that the emotional data status is not labeled by users. The sample data statistics of each volunteer are shown in Table 2. Ten volunteers' wechat

TABLE 2. ECG sample data statistics.

User's number	Number of samples in different emotional states				
	Normal	Type 1	Type 2	Type 3	others
User-1	34	21	24	17	132
User-2	23	16	19	24	112
User-3	43	22	19	27	210
User-4	33	21	24	31	167
User-5	27	15	26	30	174
User-6	24	17	22	24	179
User-7	21	26	21	25	217
User-8	37	23	25	26	224
User-9	34	22	28	19	256
User-10	32	20	23	17	137

friend circle emotional texts are collected to form different datasets for experiments. Based on this, the performance of the proposed method in this paper is evaluated.

B. EXPERIMENTAL RESULTS AND ANANLYSIS

The unlabeled data in the collected samples is defined as the sample set U. And it is used for unsupervised training network. ECG data is input into the ECG-MCAE network training convolutional aotoencoder layer parameters, and the emotional text feature data is input into the Text-CAE network aotoencoder layer parameters. The collected labeled sample data set is recorded as D, 80% of which is taken as training set D1 and the remaining 20% as test set D2. After the unsupervised training, the parameters of ECG-MCAE network, Text-CAE network, the final fully connected layer and the classifier are adjusted by the supervised training. The environment data and mobile phone location data of Type 2 and Type 3 in D1 are extracted respectively. The model is trained, and the generated model is tested with D2.

The accuracy of emotional fatigue detection results of 10 groups of users is shown in Figure 4. Only the accuracy of User_1 and User_6 is relatively low (78.34% and 83.88% respectively). While the average accuracy of other users' emotional fatigue prediction is more than 85%. And the average accuracy of emotional fatigue detection is shown in Figure 5. In the collected sample data, there are differences in the number of samples of several emotional fatigue states, and in the standards of users' emotional fatigue state. So that there are great differences in the performance of emotional fatigue detection. Therefore, in the follow-up study, it need to be considered to improve the selection of data characteristics of emotional fatigue detection and increase the amount of sample data collection. And the accuracy and reliability of emotional fatigue detection can be further improved.

In order to realize the interaction with users, the trained emotional fatigue model is saved in the cloud. And a unique identification is established for each user's emotional model.

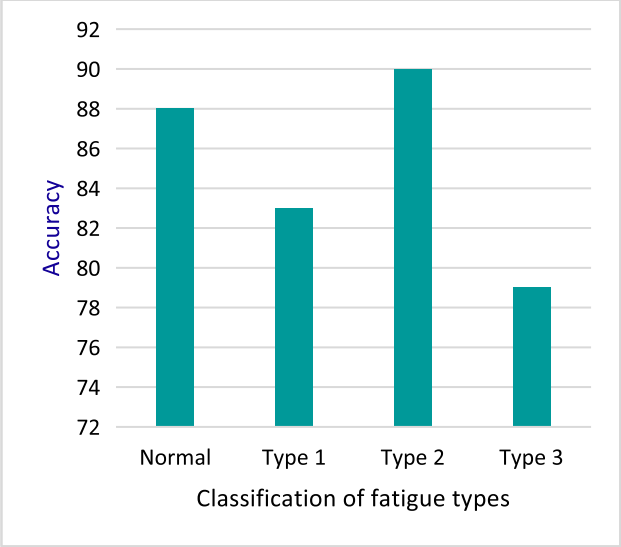


FIGURE 5. Average accuracy rate of emotional fatigue detection.

The ECG data, emotional text feature data, physiological data and environmental data related to user fatigue detection are transmitted to the cloud. Then, the received multimodal data is analyzed by the cloud in real time. And then, the emotional state of users is predicted. When it is detected that the user is in the state of emotional fatigue, personalized feedback shall be given to the user according to the type of emotional fatigue. Favorite music, suitable videos, or the personalized adjustment plans, such as rest, sports, relaxed working mode are recommended to the user.

V. CONCLUSION

When medical health data is used for disease risk assessment, whether it is structured data, unstructured data, or time series data, data features need to be extracted. That is, it needs to produce feature representation of the original data. Feature representation is the core content of machine learning.

How to select and represent features is the difficulty and hotspot of machine learning research and application. In fact, the feature representation method, which is used to obtain the features of the original input, is a representation of searching the original input data. The purpose of using the feature representation to represent the original input data is to accurately express the original input data efficiently when completing a specific task.

In this paper, a multi-channel convolutional aotoencoder model is proposed. It is used to design the ECG-MCAE network structure to learn the ECG characteristics of time series data. The convolutional aotoencoder model is used to design the Text-CAE network structure and learn the emotional text features. A multi-modal data-driven emotional fatigue detection model is proposed, the relationship between fatigue and emotional abnormality is analyzed, the concept of emotional fatigue is defined, and emotional fatigue is divided into three types according to the causes. In the emotional fatigue detection system, ECG features and emotional text features are combined to detect emotional fatigue. However, the study is still in the preliminary stage, and there are many problems need to be further studied. In the follow-up work, different activation functions can be used according to different channels. So that more feature information can be learned by the model, and the multi-channel convolution neural network model proposed in this paper can be improved.

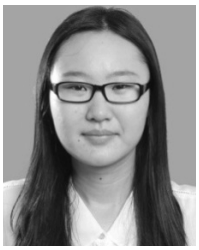
REFERENCES

[1] K. Martin, R. Meeusen, and K. G. Thompson, "Mental fatigue impairs endurance performance: A physiological explanation," *Sports Med.*, vol. 48, no. 3, pp. 1–11. 2018.
[2] E. Franchini, M. Y. Takito, and E. D. Alves, "Effects of different fatigue levels on physiological responses and pacing in judo matches," *J. Strength Conditioning Res.*, vol. 33, no. 3, pp. 783–792, 2019.
[3] A. Tsujimoto, W. W. Barkmeier, and R. L. Erickson, "Shear fatigue strength of resin composite bonded to dentin at physiological frequency," *Eur. J. Oral Sci.*, vol. 126, no. 4, pp. 316–325, 2018.
[4] W. C. Chen, Y.-J. Hsu, and M.-C. Lee, "Effect of burdock extract on physical performance and physiological fatigue in mice," *J. Veterinary Med. Sci.*, vol. 79, no. 10, pp. 1698–1706, 2017.
[5] Y. Hao, M. Chen, and L. Hu, "Energy efficient task caching and offloading for mobile edge computing," *IEEE Access*, vol. 6, pp. 11365–11373, 2018.
[6] M. Chen, Y. Zhang, Y. Li, M. M. Hassan, and A. Alamri, "AIWAC: Affective interaction through wearable computing and cloud technology," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 20–27, Feb. 2015.
[7] H. Y. Ping, L. N. Abdullah, and A. A. Halin, "A study of physiological signals-based emotion recognition systems," *Int. J. Comput. Technol.*, vol. 11, pp. 2189–2196, Sep. 2013.
[8] S. Wioleta, "Using physiological signals for emotion recognition," in *Proc. IEEE 6th Int. Conf. Hum. Syst. Interact. (HSI)*, Jun. 2013, pp. 556–561.
[9] T. Kato, H. Kawanaka, M. S. Bhuiyan, and K. Oguri, "Classification of positive and negative emotion evoked by traffic jam based on electrocardiogram (ECG) and pulse wave," in *Proc. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 1217–1222.
[10] L. Xun and G. Zheng, "ECG signal feature selection for emotion recognition," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 11, no. 3, pp. 1363–1370, 2013.
[11] S. Jerritta, M. Murugappan, and K. Wan, "Emotion detection from QRS complex of ECG signals using hurst exponent for different age groups," in *Proc. IEEE Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2013, pp. 849–854.
[12] Y. L. Zheng and S. C. Lee, "The law of iterated logarithm of rescaled range statistics for AR (1) model," *Acta Math. Sinica English*, vol. 22, no. 2, pp. 535–544, 2006.
[13] S. Jerritta, M. Murugappan, and K. Wan, "Emotion detection from QRS complex of ECG signals using hurst exponent for different age groups," *Affect. Comput. Intell. Interact.*, to be published.
[14] R. F. Ceballos and F. F. Largo, "The estimation of the hurst exponent using adjusted rescaled range analysis, detrended fluctuation analysis and variance time plot: A case of exponential distribution," *Imperial J. Interdiscipl. Res.*, vol. 3, no. 8, pp. 424–434, 2017.
[15] H. Gao, W. Huang, X. Yang, Y. Duan, and Y. Yin, "Toward service selection for workflow reconfiguration: An interface-based computing solution," *Future Gener. Comput. Syst.*, vol. 87, pp. 298–311, Oct. 2018.
[16] K. Xia, H. Yin, and Y.-D. Zhang, "Deep semantic segmentation of kidney and space-occupying lesion area based on SCNN and resnet models combined with SIFT-flow algorithm," *J. Med. Syst.*, vol. 43, no. 1, 20119, Art. no. 2.
[17] L. Zhao, H. Huang, and X. Li, "An accurate and robust approach of device-free localization with convolutional autoencoder," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5825–5840, Jun. 2019.
[18] C. Zuo, Q. Chen, and A. Asundi, "Boundary-artifact-free phase retrieval with the transport of intensity equation: Fast solution with use of discrete cosine transform," *Opt. Express*, vol. 22, no. 8, pp. 9220–9244, 2014.
[19] X. Li, S. Gannot, L. Girin, and R. Horaud, "Multichannel identification and nonnegative equalization for dereverberation and noise reduction based on convolutive transfer function," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1755–1768, Oct. 2018.
[20] C. Hao, G. Wei, and J. Yu, "Multichannel convolution blind separation algorithm for MIMO DSSS/CDMA system," *Circuits Syst. Signal Process.*, vol. 26, no. 2, pp. 249–262, 2007.
[21] I. T. Podolak and A. Roman, "CORES: Fusion of supervised and unsupervised training methods for a multi-class classification problem," *Pattern Anal. Appl.*, vol. 14, no. 4, pp. 395–413, 2011.
[22] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.
[23] H. Gao, W. Huang, and X. Yang, "Applying probabilistic model checking to path planning in an intelligent transportation system using mobility trajectories and their statistical data," *Intell. Automat. Soft Comput.*, vol. 25, no. 3, pp. 547–559, 2019.
[24] H. Gao, W. Huang, Y. Duan, X. Yang, and Q. Zou, "Research on cost-driven services composition in an uncertain environment," *J. Internet Technol.*, vol. 20, no. 3, pp. 755–769. 2019.
[25] S. M. Razavi, M. Taghipour-Gorjikolaie, and N. Mehrshad, "Multimodal biometric identification system based on finger-veins using hybrid rank–decision-level fusion technique," *IEEJ Trans. Electr. Electron. Eng.*, vol. 12, no. 5, pp. 728–735, Sep. 2017.
[26] L. Guan, Y. Tong, and J. Li, "An Online surface water COD measurement method based on multi-source spectral feature-level fusion," *RSC Adv.*, vol. 9, no. 20, pp. 11296–11304, 2019.

JIANQIANG XU received the M.B.A. degree from the East China University of Science and Technology, in 2009. He is currently pursuing the Ph.D. degree in control science and engineering with the East China University of Science and Technology. His research interests include the Internet of Things, artificial intelligence, and cloud computing.

ZHUJIAO HU received the Ph.D. degree in microelectronics and solid-state science from Fudan University, and the M.Sc. degree in engineering from Wuhan University, in 2005. She was with Shanghai Fine Electronics Company, Ltd. Her research interests include the Internet of Things, artificial intelligence, and big data.

JUNZHONG ZOU received the B.S. degree in electrical engineering from Chongqing University, Sichuan, China, in 1982, and the M.S. and Ph.D. degrees in advanced systems control engineering from Saga University, Japan, in 1995 and 1998, respectively. He joined the East China University of Science and Technology, Shanghai, China, in 2001. He is currently a Professor and also a Thesis Adviser of M.S. and Ph.D. graduate students with the East China University of Science and Technology (ECUST). He is also the Director of the Texas Instrument Digital Signal Processing Associated Lab, ECUST. His research interests are dynamical control and automation, biomedical signal processing, robotics, and mechatronic servo systems.



ANQI BI received the Ph.D. degree from Jiangnan University, Wuxi, China, in 2017. Her research directions are medical information and medical image proceeding, and so on. She was with the School of Computer Science and Engineering, Changshu Institute of Technology.