Computers in Biology and Medicine 159 (2023) 106938

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition

Tianqi Fan [a], Sen Qiu [a,*], Zhelong Wang [a], Hongyu Zhao [a], Junhan Jiang [b], Yongzhen Wang [c], Junnan Xu [d], Tao Sun [d], Nan Jiang [e]

[a] Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Dalian, China
[b] First Affiliated Hospital of China Medical University, Shenyang, China
[c] The Army Research Institute, Beijing, China
[d] Department of Medical Oncology, Cancer Hospital of Dalian University of Technology, Shenyang, China
[e] College of Information Engineering, East China Jiaotong University, Nanchang, China

ARTICLE INFO

Keywords:
Emotion classification
Deep learning
Attention mechanisms
Convolutional neural networks
Electrocardiogram (ECG)
Convolutional Block Attention Module (CBAM)

ABSTRACT

Using ECG signals captured by wearable devices for emotion recognition is a feasible solution. We propose a deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition. In order to address the problem of individuality differences in emotion recognition tasks, we incorporate an improved Convolutional Block Attention Module (CBAM) into the proposed deep convolutional neural network. The deep convolutional neural network is responsible for capturing ECG features. Channel attention in CBAM is responsible for adding weight information to ECG features of different channels and spatial attention is responsible for the weighted representation of ECG features of different regions inside the channel. We used three publicly available datasets, WESAD, DREAMER, and ASCERTAIN, for the ECG emotion recognition task. The new state-of-the-art results are set in three datasets for multi-class classification results, WESAD for tri-class results, and ASCERTAIN for two-category results, respectively. A large number of experiments are performed, providing an interesting analysis of the design of the convolutional structure parameters and the role of the attention mechanism used. We propose to use large convolutional kernels to improve the effective perceptual field of the model and thus fully capture the ECG signal features, which achieves better performance compared to the commonly used small kernels. In addition, channel attention and spatial attention were added to the deep convolutional model separately to explore their contribution levels. We found that in most cases, channel attention contributed to the model at a higher level than spatial attention.

1. Introduction

Affective computing is a cross-fertilized research field that encompasses the disciplines of psychology, computer science and cognitive science [1]. Emotion recognition is an important part of affective computing and the recognition of an emotional state is the key for the computer to understand emotion and react. Emotion is a complex psychological phenomenon and different emotional states produce different expressions, which can be expressed through facial expressions, tone of voice, physiological signals, and other means as mediators. It has been shown that using wearable devices to capture physiological signals and combine them with deep learning or machine learning techniques for emotion recognition can achieve better results [1–3]. The data sources for affective computing are multimodal and includes facial expressions [4], voice intonation [5], gait [6], body posture [7], electroencephalogram (EEG) [8], electrocardiogram (ECG) [9], and galvanic skin (GSR) [10]. Agrafioti et al. [11] showed that there is a strong correlation between ECG signals and individual emotions, the current emotional state of an individual can be assessed by changes in ECG signals. At the present time, machine learning or deep learning methods have been applied to ECG emotion recognition [12,13]. The machine learning approach requires manual extraction of ECG features, including mean, median, standard deviation, minimum, maximum, interbeat interval (IBI), heart rate (HR), heart rate variability (HRV), total power, power spectral density (PSD) for low frequency (LF), and PSD for high frequency (HF) of the P-QRS-T complex [14]. However, using machine learning methods requires specialized knowledge of ECG

* Corresponding author.
E-mail addresses: ftq@mail.dlut.edu.cn (T. Fan), qiu@dlut.edu.cn (S. Qiu), wangzl@dlut.edu.cn (Z. Wang), zhaohy@dlut.edu.cn (H. Zhao), junhanjiang@outlook.com (J. Jiang), 308961950@qq.com (Y. Wang), 15040292193@126.com (J. Xu), jianong@126.com (T. Sun), jiangnan1018@acm.org (N. Jiang).

https://doi.org/10.1016/j.compbiomed.2023.106938
Received 26 February 2023; Received in revised form 28 March 2023; Accepted 14 April 2023
Available online 22 April 2023

and is time-consuming [1,15]. Compared to machine learning, deep learning methods avoid the process of manual feature extraction and enable end-to-end emotion recognition using only raw data [13,16].

In emotional physiological signal acquisition, specific emotion-inducing material is first required to produce the appropriate emotional state in the individual, such as an edited movie clip [14]. However, a key question is: how to address the issue of individual variability in ECG signals? In ECG emotion recognition, there may be differences in the sensitivity of different subjects to emotion-inducing materials, which may lead to differences in ECG signals across subjects in the same emotional state. To this end, we designed a novel ECG signal-based emotion recognition model that combines convolutional neural network (CNN) with attentional mechanisms for emotion classification of ECG signals. We used attentional mechanisms to access key features of ECG signals in emotion expression and thus mitigate the effects of individuality differences. CNN is an efficient and widely used method in deep learning, which its efficiency is due to two aspects: local perception and weight sharing [17]. Local perception makes it possible for each neuron in the CNN to perceive only local regions, after which higher-level neurons integrate the local information to obtain global information. The principle of weight sharing is that the features of one part of the data are the same as the other parts, which indicates that the features learned on one part of the data can be applied to the other part, so the same weights can be used to learn the features for the overall data. Previous studies have demonstrated that the learned features can be enriched by stacking convolutional layers. However, too deep networks suffer from degradation, i.e., accuracy decreases instead of improving as the network deepens. He et al. [18] proposed a ResNet network with layer-hopping connections through "shortcut connections", which solves the two problems of network degradation and gradient disappearance or explosion, allowing the network to improve performance by deepening. Attention mechanisms are widely used in natural language processing and computer vision fields. The main purpose of adding the attention mechanism to the model is to make the neural network pay attention to some important and common information. In recent years, the attention mechanism has been rapidly developed and many attention networks have been derived, such as SENet [19], FcaNet [20], CBAM [21], and Self-attention [22].

In order to obtain the key ECG features in the expression of emotional, we propose to combine CBAM [21] with CNN to learn weights from feature maps through the attention mechanism and apply the learned weight information to the original feature maps, which in turn changes the distribution of the original features so that important features are attended to and redundant features are suppressed. We propose a deep convolutional neural network incorporating CBAM attention. Fig. 1 shows the whole flow of the model. Three publicly available datasets, WESAD [2], DREAMER [14], and ASCERTAIN [23] were used to implement ECG for emotion recognition.

Our contribution can be summarized as follows:

• We designed a deep convolutional network with residual structure and applied it to ECG emotion recognition. We also explored the impact of convolutional kernel size on the ECG signal emotion recognition task. The analysis shows that using a larger size of the convolutional kernel may be beneficial for the ECG emotion recognition task. However, it may have an upper limit, too large convolutional kernel size prevents model performance improvement and also increases computational resources.

• We added channel space attention to the network and implemented a detailed analysis of the network parameters. We analyzed the effects of channel attention and spatial attention on the recognition task. The results showed that channel attention had a higher level of contribution than spatial attention.

• We set new state-of-the-art results for multi-class ECG emotion recognition tasks in three datasets WESAD, DREAMER, and ASCERTAIN. We demonstrate that deep convolutional neural networks can learn ECG features well. In addition, experimental results show that the attention mechanism has an improved effect on emotion recognition.

The rest of the paper is organized as follows: Section 2 briefly summarizes previous work in the literature. Section 3 describes our proposed model. Section 4 describes our experimental procedure and the analysis of the experimental results. Section 5 provides a summary of our work.

2. Related work

2.1. Electrocardiogram and emotions

The ECG signal records the electrical activity generated by each cardiac cycle of the heart and usually consists of P waves, PR intervals, QRS wave groups, J points, ST segments, T waves, U waves, and QT intervals. ECG signals have been widely used in areas such as emotion recognition [24], stress monitoring [25], and sleep monitoring [26]. The autonomic nervous system (ANS) is part of the peripheral efferent nervous system and can regulate the activity of the heart muscle and glands. The ANS is divided into two parts: the sympathetic nerve system (SNS), which is regulated by postganglionic fibers and increases the heart rate during emotional episodes, and the parasympathetic nerve system (PNS), which is regulated by the vagus nerve and used to slow down the heart rate [27]. In other words, heart rate is influenced by the ANS, which in turn is influenced by emotion. Therefore, emotion recognition by ECG signals is feasible. The methods of using ECG signals for emotion recognition focus on machine learning [24,28,29] and deep learning [16,30–32]. Since machine learning methods require specialized ECG knowledge and manual extraction of features, end-to-end emotion recognition of ECG signals using deep learning has been gaining popularity in recent years [33].

2.2. Emotion recognition based on ECG signals

The purpose of ECG signal emotion recognition is to use the collected ECG signals to accurately classify the corresponding emotional states. A large amount of research has been focused on this area.

An earlier study [34] combined Rescaled Range Statistics (RRS) and Finite Variance Scaling (FVS) methods with Higher Order Statistics (HOS) to extract new nonlinear features for ECG signals, which have also shown that the combination of nonlinear analysis and HOS can capture more fine-grained changes in emotion. Sepúlveda et al. [28] used wavelet transforms to extract features on different time scales of ECG signals and then used classifiers to identify emotional states, reporting an accuracy of 95.3%. Wang et al. [35] investigated the psychological responses of drivers to changes in the traffic environment and extracted the time–frequency domain features, waveform features, and nonlinear features of ECG signals, which the identification results were 91.34% and 92.89% for the calm and anxious states, respectively. Goshvarpour et al. [36] collected ECG signals from 11 students under emotional music stimulation and extracted MP coefficients from the ECG signals for emotion recognition. Harper et al. [30] proposed an end-to-end deep learning model for classifying emotions from single-peaked heartbeat time series. In addition, in order to model the uncertainty of emotion prediction, they further proposed a Bayesian framework and achieved a classification accuracy of 90%. Hwang et al. [16] proposed a deep ECGNet, which was validated on two datasets using ultra-short-term raw ECG signals, achieving 87.39% and 73.96% recognition accuracy.

A number of previous works such as [9,13,32] have utilized one or more of the datasets used in this paper, i.e., WESAD, DREAMER, and ASCERTAIN, for emotional recognition. To address the problems of small ECG sample data size and category imbalance, Nita et al. [13] proposed to enrich the ECG dataset using data augmentation and achieved 75.1% recognition accuracy on the DREAMER dataset. A

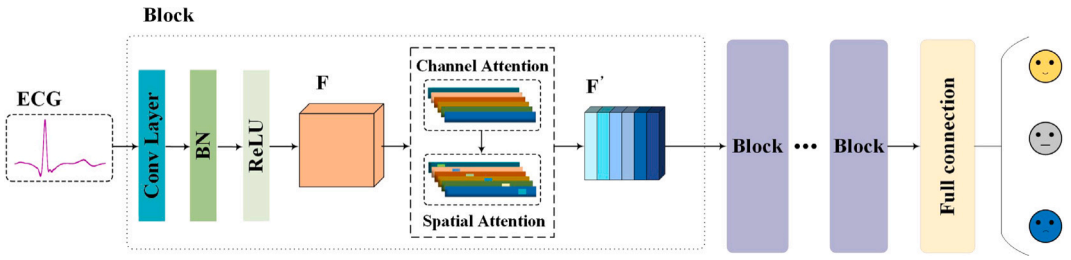


Fig. 1. The whole process of electrocardiographic emotion recognition. Subjects' ECG signals stimulated by multimedia materials were fed into the proposed model. ECG feature representations are obtained by CNN and assigned weights to them by channel space attention mechanism. Finally, classification is performed through a fully connected layer.

recent work [9] utilizes a self-supervised approach for ECG-based emotion recognition, which the proposed scheme consists of two stages: learning ECG representation and emotion classification. The first stage implements six different transformations of the ECG signal, which are then fed into the network for ECG representation learning. In the s stage, the convolutional layer from the first stage is frozen and the fully connected layer is fine-tuned for emotional state classification, recognition accuracies of 96.9% and 85.9% were achieved on the publicly available datasets WESAD and DREAMER. Behinaein et al. [32] designed a deep neural network based on convolutional layers and fused attention mechanisms for detecting stress states, achieving 91.1% accuracy on the publicly available dataset WESAD.

2.3. Attention mechanisms in ECG signals

Attention mechanisms have received extensive attention in various tasks based on ECG signals. Wang et al. [37] proposed a convolutional neural network with a nonlocal convolutional attention module (NCBAM) that feeds raw ECG signals into a CNN architecture to extract spatial and channel features. In the end, non-local attention is used to capture long-term dependent features in the spatial and channel dimensions. Ge et al. [38] introduced the Squeeze-and-Excitation (SE) module in a convolutional neural network for the ECG signal classification task. They used the residual module to achieve ECG feature extraction and the SE attention module to achieve effective feature enhancement by explicit modeling of channel dimensions. Song et al. [39] constructed a multimodal physiological emotion database and proposed an attentional long and short-term memory network (A-LSTM) that enhances the validity of discriminative features using attentional mechanisms. Recent work is to perform ECG classification tasks using the self-attention mechanism. Hu et al. [40] proposed a hybrid transformer model based on the self-attention mechanism to achieve better classification results by using the self-attention mechanism for weighted representation of important information in ECG signals.

Since the attention mechanism performs well in the above work, we propose to apply channel space attention to ECG emotion recognition. In this work, we first design a deep convolutional neural network for extracting ECG features, after that apply channel space attention to the extracted features, and finally obtain the ECG key features for emotion expression.

3. Method

The ECG emotion recognition task is a sequence-to-sequence task that takes a segment of the captured ECG signal $X = [x_1, \ldots, x_k]$ as input and the sequence of emotion labels $L = [l_1, \ldots, l_n]$ as output. Each label corresponds to a segment of ECG signals. Our goal in this paper is to distinguish different emotional states by ECG signals. Fig. 2 shows the deep convolutional neural network with a fused channel space attention mechanism. It consists of a convolutional module and a CBAM attention module. The CNN layer is responsible for capturing the ECG signal features and the CBAM module is responsible for assigning weights to the channel and spatial dimensions of the features, enabling the model to focus on the important features.

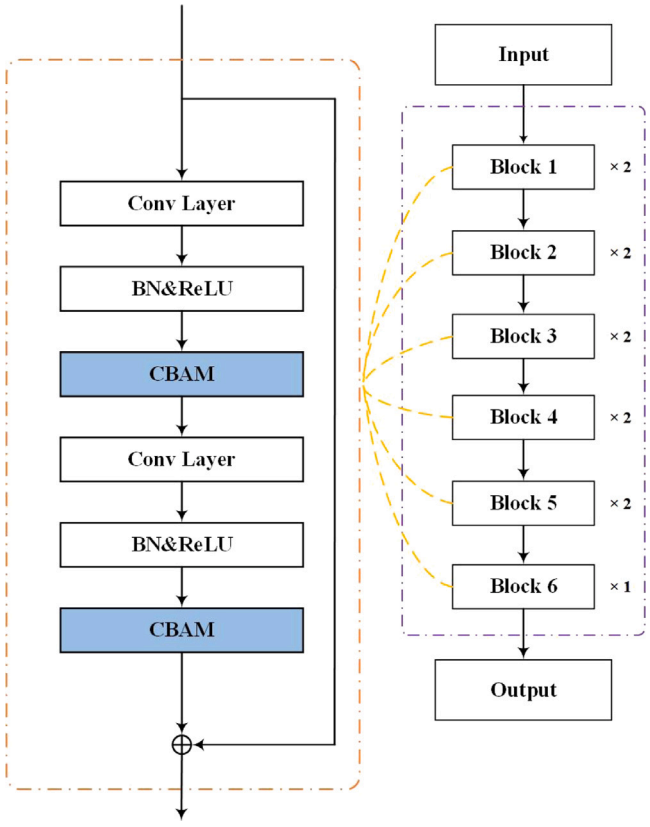


Fig. 2. The overall construction of the proposed model backbone network. We set kernel_size=[17, 15, 13, 11, 9, 7] (from Block 1 to Block 6).

3.1. ECG signal feature extraction

We propose to use CNNs to extract features from ECG signals. CNNs are able to extract spatial features adequately and their exploitation of translation invariance and localization allows them to learn useful representations with fewer parameters. Suppose the input is a one-dimensional data X and its mathematical hidden representation is H (X and H have the same shape). We use X_i and H_i to denote the input data and the data at position i in the hidden representation, respectively. Then the fully connected layer can be represented as follows, where V is the weight, U contains the bias information, and a is the translation.

$$H_i = U_i + \sum_a V_{i,a} X_{i+a} \tag{1}$$

The spatial invariance of the CNN implies that the translation of the detected object in the input X can only lead to a translation in the hidden representation H. That is, V does not actually depend on the position i, i.e., $V_{i,a} = V_a$, and U is a constant. At this point, H can be

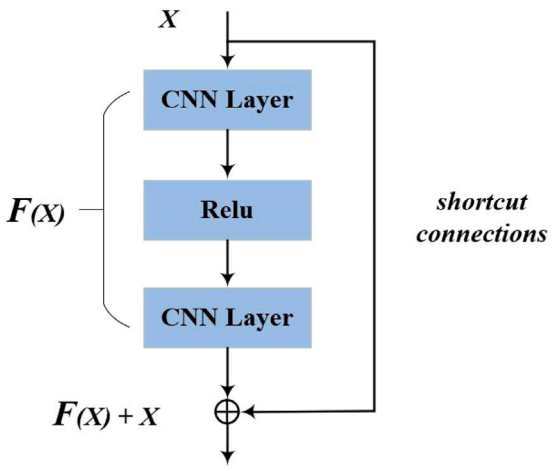


Fig. 3. Typical residual structure.

defined as follows.

$$H_i = u + \sum_a V_a X_{i+a} \tag{2}$$

We obtain H_i using the coefficients V_a weighted by the data near the position $(i + a)$, where the coefficients of V_a are much less than $V_{i,a}$, because the former no longer depends on the position in the data. Localization of CNN means that in order to collect information related to the parameter H_i used for training, distances from location i should not be considered, i.e., there exists a critical value δ, if the offset $a > \delta$, then $V_a = 0$. At this point, H can be defined as follows. Usually, V is called a convolution kernel or filter.

$$H_i = u + \sum_{a=-\delta}^{\delta} V_a X_{i+a} \tag{3}$$

We introduced a residual structure [18] in the proposed model in order to address the degradation caused by the deep network. The residual structure ensures that the results of the current layer are not worse than the previous ones, Fig. 3 shows its typical structure. We use $H(x)$ to denote the underlying expectation mapping, i.e., $H(x) = F(x) + x$, $F(x)$ is the residual mapping function and x is often referred to as the identity function, which is a shortcut connection. If there is a degeneracy problem at one level, $F(x) + x$ is still guaranteed to be the current optimal value by simply letting $F(x)$ converge to 0. In our model, we set up six layers of convolution blocks, each consisting of two layers of convolution with a residual structure, which is shown in detail in Fig. 2.

3.2. Attentional mechanisms

We introduced CBAM [21] in the proposed model, whose structure is shown in Fig. 4. Since convolutional operations extract information features by mixing cross-channel and spatial information, it may be important to add attentional representations in both channel dimension and spatial dimension, because not every channel and spatial location is equally important. We add the CBAM in the middle of the convolutional layer, which takes the convolutional layer output feature map $F \in R^{C\times L}$ as input. First, CBAM gets channel attention mapping $M_c \in R^{C\times 1}$ and spatial attention mapping $M_s \in R^{1\times L}$. The whole process is represented as follows, where $\otimes$ represents element-by-element multiplication.

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s\left(F'\right) \otimes F' \end{aligned} \tag{4}$$

In channel attention, we use average pooling and maximum pooling for spatial information aggregation. The average pooling calculates the average value of elements within the pooling window and the maximum pooling calculates the maximum value of elements within the pooling window. Maximum pooling and average pooling are used to reduce the dimensionality, they can be defined as:

$$\begin{aligned} out_{max} &= max[x_1, x_2, \cdots, x_n] \\ out_{avg} &= mean[x_1, x_2, \cdots, x_n] \end{aligned} \tag{5}$$

where x is the input data. First, average pooling feature F^c_{avg} and maximum pooling feature F^c_{max} are generated by average pooling and maximum pooling, respectively. After that, these two features are passed through a shared network, which in turn generates the channel attention mapping $M_c \in R^{C\times 1}$. We improve the shared MLP layer in the original CBAM to a convolutional layer, which allows the model to reduce the number of parameters by a certain amount. In our model, this shared network consists of a convolution containing a hidden layer. In order to reduce the parameter overhead, the output of the hidden layer is $R^{C/r\times 1}$, where r is the scaling rate. Finally, the element-by-element summation is used to output the channel attention weights, It can be expressed as follows, where σ is the $sigmoid$ function, $W_0 \in R^{C/r\times C}$, and $W_1 \in R^{C\times C/r}$.

$$\begin{aligned} M_c(F) &= \sigma(\mathrm{Conv}(\mathrm{AvgPool}(F)) + \mathrm{Conv}(\mathrm{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{max}\right)\right)\right) \end{aligned} \tag{6}$$

We use the spatial relationship between features to generate a spatial attention graph as a complement to channel attention, which focuses more on which position of the data is more effective. Spatial attention first aggregates the channel information of a feature map by two pooling operations to generate two mappings representing the average pooling feature $F^{avg}_s \in R^{1\times L}$ and the maximum pooling feature $F^{max}_s \in R^{1\times L}$ across channels, respectively. After that, the obtained mappings are connected and passed through a convolutional layer to produce a spatial attention map. Finally, the final spatial attention is obtained by a $sigmoid$ function. The calculation process of spatial attention can be expressed as follows. k is the convolution kernel size, and $k = 17$ in the proposed model.

$$\begin{aligned} M_s(F) &= \sigma\left(f^k([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])\right) \\ &= \sigma\left(f^k\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right) \end{aligned} \tag{7}$$

3.3. Combination of CNN and CBAM

Attentional mechanisms can help alleviate the problem of individuality in ECG data. In ECG signals, there may be differences in ECG waveforms and characteristics between individuals. Such differences may become more pronounced in response to emotional stimuli. Attentional mechanisms can help the model learn important features in ECG signals and commonalities between individuals. We add a CBAM module after the convolution layer to pass the feature map output after convolution through channel attention and spatial attention, then apply weight information to the extracted features. We use convolutional networks to extract ECG signal features and the different weight parameters on the convolutional kernel are with fixed regions, i.e., the convolution can only form “limited attention” within its receptive field. Although the size of the receptive field increases as the number of layers increases, the perception of shallow and mid-level features is still indirect and weak. The contribution made by the attention mechanism is to focus on the global, e.g., all channels or the entire space, which can adaptively update the weights of the information for different channels or different spatial locations. CNN extracts features from the ECG signal and generates a feature map representing the input data. The CBAM mechanism is then applied to the feature map to achieve dynamic attention to the most important parts of the ECG signal. Specifically, the channel attention of CBAM focuses on the most important channels in the feature map, while the spatial attention focuses on the most important regions. The combination of these two attention mechanisms

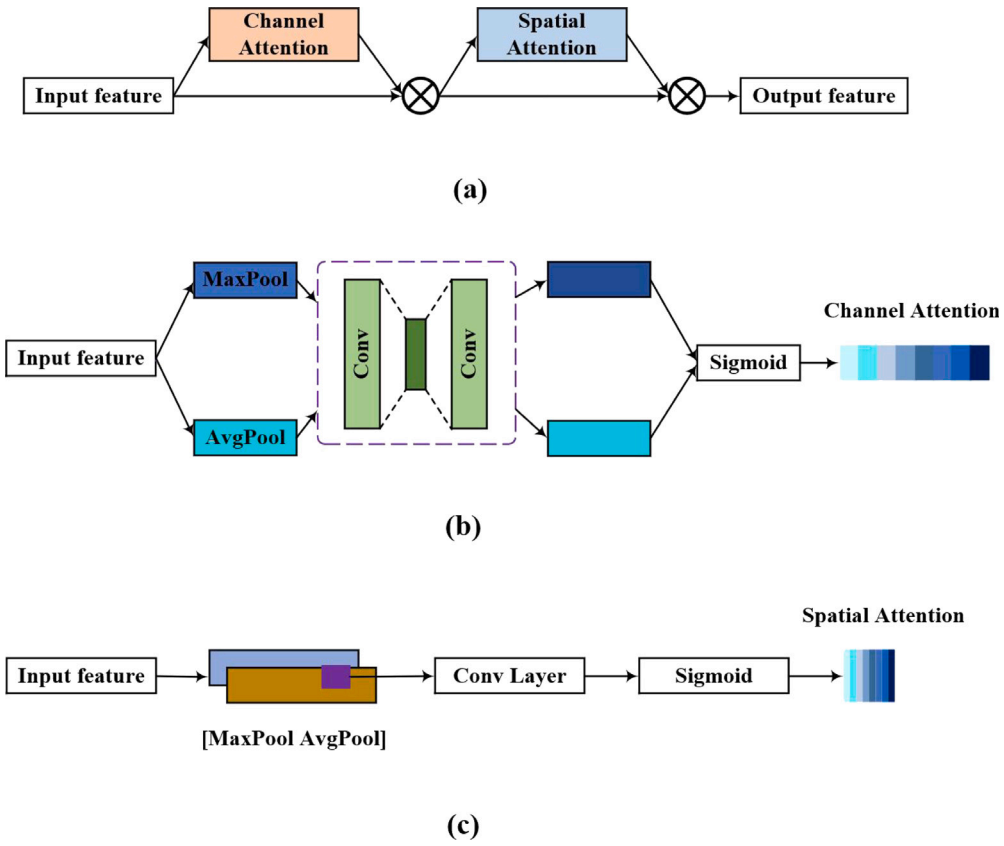


Fig. 4. CBAM structure. (a) is the overall CBAM structure, (b) is channel attention, and (c) is spatial attention.

Table 1
A description of the dataset used.

Dataset	Subjects	Attributes	Classes
WESAD	15	Affect State	4
DREAMER	23	Arousal, Valence	5
ASCERTAIN	58	Arousal, Valence	7

provides a more comprehensive representation of the ECG signal and helps to resolve individual differences in the ECG signal.

The original ECG signal is preprocessed and fed into the CNN layer to obtain the initial feature R_{ECG}. Subsequently, R_{ECG} are passed through the CBAM to obtain $A_{ECG} = \alpha R_{ECG}$ (α is the attention weight). The local information of the ECG signal is obtained by CNN and the global information is concerned by the CBAM module. In this way, we extract the weighted feature representation of the ECG signal from shallow to deep layers.

Finally, we use the fully connected layer to classify the sentiment of ECG signals. We use multicategorical cross-entropy as the loss function of the proposed model, which can be expressed as follows, where C is the sentiment category, y_i is the true label, and $\hat{y}_i$ is the predicted label.

$$Loss = -\sum_{i=1}^{C} y_i \cdot \log \hat{y}_i \tag{8}$$

4. Experiments

4.1. Datasets

We use three publicly available datasets and Table 1 provides brief information about them. We validated and deeply analyzed the proposed method on these three datasets.

4.1.1. WESAD

The WESAD [2] is a dataset for stress and impact detection. Seventeen subjects were recruited (final data valid for 15 subjects) and exposed to four different affective stimuli, namely neutral, stress, recreation, and meditation. ECG signals were acquired using a wearable device at a sampling rate of 700 Hz. Neutral affective ECG signals were first acquired for 20 min, during which time participants sat or stood at a table and viewed neutral material. In the entertainment scenario, subjects watched 11 entertainment video clips with a total length of 392 s. In the stress scenario, subjects performed a 10 min public speaking and mental arithmetic task. Finally, subjects performed a 7 min meditation task. After each trial, subjects were asked to fill out the affective scale (PANAS) [41].

4.1.2. DREAMER

The DREAMER dataset [14] recruited 25 subjects with film clip-style visual and auditory stimuli and simultaneous recording of ECG signals. Eighteen movie clips were selected to induce nine different emotions: amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, and surprise. To avoid data contamination, only the last 60 s of the movie clips were collected. During the experiment, ECG data were collected at 256 Hz using a wearable device. After each clip, arousal (ranging from uninterested/bored to excited/alert) and valence (ranging from unpleasant/stressed to happy/elated) were scored using the Self-Assessment Manikins (SAM) [42] on a scale of 1 to 5.

4.1.3. ASCERTAIN

The ASCERTAIN dataset [23] recruited 58 subjects (21 females, mean age = 30) for the study using 36 movie clips, each ranging from 51–127 s in length. While the subject was watching the video, the ECG signal was collected simultaneously in the left and right arms using a 256 Hz wearable device. Emotional ratings were used valence and arousal, ranging from −3 (very negative) to 3 (very positive) and 0

(very boring) to 6 (very exciting), respectively. During the experimental acquisition, the subject will produce body movements, which will affect the quality of the collected physiological signals. Therefore, each segment of the physiological signal was manually marked on a scale of 1 (good data) to 5 (missing data).

4.2. Data pre-processing and training

Since the above datasets were collected by different devices and under different scenarios, the data may have minor differences and pre-processing of the data is necessary. We first downsample the WESAD dataset to 256 Hz to ensure the same sampling rate for all three datasets. Then we set up a 3–45 Hz FIR bandpass filter to filter the raw ECG signals. Finally, we normalized the data. We split the filtered ECG signals into fixed windows of 1 s and the windows did not overlap. It should be noted that the setting of the window size is empirical, we think it is appropriate to set 1 s as the window size because a heartbeat cycle is about 0.8 s. We implement the proposed model architecture using Pytorch and use NVIDIA A5000 GPU 24G for training. The experimental program is written in Python 3.8. We evaluate the model using 10-fold cross-validation, we use 90% of the data for training and 10% of the data for testing (this process is repeated 10 times), the model uses the Adam optimizer [43] with a learning rate of 0.001 and we apply the early stop method in our experiments.

4.3. Evaluation metrics

We evaluated the performance of the algorithm model using the average accuracy and $F1-Score$, which were calculated as follows, where TP indicates true positive, TN indicates true negative, FP indicates false positive and FN indicates false negative. C is the number of categories.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

$$F1-Score = \frac{1}{|C|} * \sum_{i=1}^{C} \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \tag{12}$$

4.4. Results

We used the proposed model to classify the different sentiments of WESAD, DREAMER, and ASCERTAIN (multi-class classification) and the results are shown in Table 2. Table 2(a) shows the results of the model with the convolutional structure only (Deep CNN) and Table 2(b) shows the results of the model with the addition of CBAM (Deep CNN-CBAM). We compared the multi-class classification results (the WESAD dataset has four categories of emotions and the DREAMER dataset has five categories of emotions) of the proposed model with the current state-of-the-art results, as shown in Table 3. In addition, to the best of our knowledge, no previous work has utilized ASCERTAIN for multi-class classification and our work is the first implementation of multi-class classification for the ASCERTAIN dataset in the arousal and valence dimensions. In ASCERTAIN, 7-class classification is performed, achieving accuracies of 68.0 and 64.5 percent for arousal and valence respectively. For the WESAD dataset, our base convolutional model achieves an accuracy of 96.1% for the recognition of the four emotional states, and a 0.5% improvement with the addition of the CBAM block. Compared to the previous state-of-the-art results, our proposed model improves by 1.1% and 1.5% in recognition accuracy, respectively. For the DREAMER dataset, our base convolutional model achieves multi-class recognition accuracy of 83.1% and 83.8% for arousal and valence, respectively, after adding the CBAM block, it increased to 83.6% and 84.2%. Compared with the previous state-of-the-art results, our proposed model improves 6.0% and 6.5% in arousal recognition accuracy, improving 8.9% and 9.3% in valence recognition accuracy, respectively.

Table 2
Multi-class sentiment recognition results are given for three datasets using the proposed two models.

(a) **Deep CNN**

Dataset	Attributes	Classes	Acc.	F1
WESAD	Affect State	4	0.961	0.953
DREAMER	Arousal, Valence	5	0.831, 0.838	0.758, 0.853
ASCERTAIN	Arousal, Valence	7	0.674, 0.612	0.612, 0.568

(b) **Deep CNN-CBAM**

Dataset	Attributes	Classes	Acc.	F1
WESAD	Affect State	4	0.965	0.967
DREAMER	Arousal, Valence	5	0.836, 0.842	0.806, 0.844
ASCERTAIN	Arousal, Valence	7	0.680, 0.645	0.580, 0.673

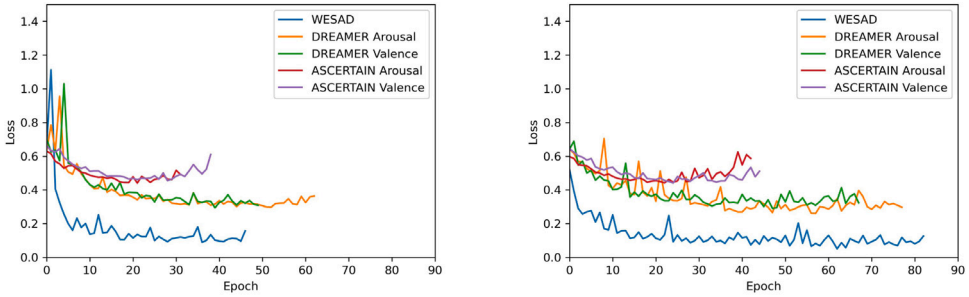
In order to better compare with previous work, we bifurcate the DREAMER and ASCERTAIN datasets. We set the arousal and valence thresholds to three for the DREAMER dataset. For the ASCERTAIN dataset, we set the arousal threshold to three and the valence threshold to zero. For the WESAD dataset, benefiting from the inspiration of [9], we remove the meditation class and implement a three-class division of the WESAD dataset. Table 4 shows the comparison of the proposed model with other works. In [9], a self-supervised CNN was implemented to classify WESAD emotions and the reported three classification accuracy was 96.9%. The results in Table 4(a) show that the proposed deep convolutional network achieved 97.2% recognition accuracy, while the addition of CBAM blocks achieved 97.5% recognition accuracy, which is better than previous work. Table 4(b) shows the results of the DREAMER dataset. For DREAMER, the self-supervised CNN achieves 85.9% and 85.0% recognition accuracy on arousal and valence, respectively, while our proposed deep convolutional model achieves 87.1% and 87.4% recognition accuracy in the two dimensions, respectively. The addition of CBAM blocks improve the recognition accuracy of the arousal dimension by 0.6%. Table 4(c) shows the results of the ASCERTAIN dataset. In [23], emotion classification of ASCERTAIN using Support Vector Machine (SVM) and Naive Bayes (NB). SVM reported F1 scores of 0.570 and 0.560 in the arousal and valence dimensions, respectively. NB reported F1 scores of 0.590 and 0.600, respectively. Our proposed deep convolutional model achieves F1 scores of 0.742 and 0.762 in two dimensions, respectively, and the F1 score in the arousal dimension improves to 0.767 after adding CBAM blocks.

The early stop is used during the training of our model, which can be considered as a regularization tool, mainly to moderate overfitting and reduce the training time. Fig. 5 shows the variation of the loss of our model with the number of epochs. Fig. 5(a) shows that the proposed deep convolutional model has a high loss fluctuation at the beginning of training. This may be due to the fact that the model does not adequately capture the key features of the ECG signal in the early stage of training. Fig. 5(b) shows that the fluctuations at the beginning of training are significantly reduced after the inclusion of the attention mechanism. This indicates that the model starts to be able to focus on the important ECG features after the attention mechanism is added. It is important to note that the number of epochs is not the same for different tasks because of the early stop. We observe that for the three datasets used, the model stops training between the 60th and 80th epochs in most cases.

4.5. Discussion

4.5.1. Effectiveness of attentional mechanisms

In terms of multi-class classification results, Table 2(a) and Table 2(b) show the results of emotional state recognition without and



(a) Deep CNN (b) Deep CNN-CBAM

Fig. 5. Loss curve of the proposed model on the WESAD, DREAMER and ASCERTAIN datasets (two or three classification). (a) is the loss curve of the proposed convolutional model and (b) is the loss curve of the model after adding the CBAM attention mechanism.

Table 3
Comparison of multi-class classification results of the proposed model with state-of-the-art results.

(a) **WESAD**			
Ref.	Method	Affect State	
		Acc.	F1
[9]	Self-Supervised CNN	0.950	0.940
Ours	Deep CNN	**0.961**	**0.953**
	Deep CNN-CBAM	**0.965**	**0.967**

(b) **DREAMER**			
Ref.	Method	Arousal, Valence	
		Acc.	F1
[9]	Self-Supervised CNN	0.771, 0.749	0.740, 0.747
Ours	Deep CNN	**0.831, 0.838**	**0.758, 0.853**
	Deep CNN-CBAM	**0.836, 0.842**	**0.806, 0.844**

Table 4
The proposed model has two/three class classification results in three datasets and is compared with other works.

(a) **WESAD**			
Ref.	Method	Affect State	
		Acc.	F1
[2]	DT	0.578	0.517
	RF	0.604	0.522
	AB	0.617	0.525
	LDA	0.663	0.560
	KNN	0.548	0.478
[9]	Self-Supervised CNN	0.969	0.963
Ours	Deep CNN	**0.972**	**0.978**
	Deep CNN-CBAM	**0.975**	**0.971**

(b) **DREAMER**			
Ref.	Method	Arousal, Valence	
		Acc.	F1
[13]	SVM	0.738, 0.751	–
[14]	Data Enhancement CNN	0.624, 0.624	0.580, 0.531
[9]	Self-Supervised CNN	0.859, 0.850	**0.859**, 0.845
Ours	Deep CNN	**0.871, 0.874**	0.843, **0.876**
	Deep CNN-CBAM	**0.877, 0.874**	0.853, **0.868**

(c) **ASCERTAIN**			
Ref.	Method	Arousal, Valence	
		Acc.	F1
[23]	SVM	–	0.570, 0.560
	NB	–	0.590, 0.600
Ours	Deep CNN	**0.787, 0.756**	**0.742, 0.762**
	Deep CNN-CBAM	**0.787, 0.763**	**0.767, 0.737**

with the addition of the attention mechanism, respectively. On the WESAD dataset, the model with the addition of CBAM attention improved the accuracy by 0.4% and the F1 score by 0.014 compared to before the addition. For the DREAMER dataset, CBAM brings an accuracy improvement of 0.5% and 0.4% for arousal and valence dimensions, respectively. On the ASCERTAIN multi-class classification, CBAM helped the model to improve the accuracy by 0.6% and 3.3% on arousal and valence, respectively. The results in Table 4 show the performance improvement brought by CBAM on two or three classifications. From the above analysis, it can be concluded that adding CBAM can bring some improvement to the recognition performance of the model. In ECG emotion recognition, attention mechanisms can focus on the important parts of ECG signals related to emotional states and suppress redundant information and noise. Emotion as a highly abstract physiological state tends to have a large individual variability, which may affect the model performance. The attention mechanism allows the model to assign different weights to different parts of the ECG signal for each individual to accommodate individual differences, thus reducing the effect of individual differences and improving accuracy. CBAM includes channel attention and spatial attention. Channel attention is responsible for capturing the importance of feature channels and then enhancing or inhibiting different feature channels for different emotional states. Spatial attention is responsible for focusing on which part of the information of different feature channels is important and enhancing the feature weights of the important parts within the channels. We obtain the ECG feature maps before the fully connected layer of the model and visualize their weight information. Fig. 6 shows the ECG feature maps of the three data sets used before and after the addition of the CBAM mechanism. It can be seen that after adding the attention mechanism, the model suppresses or enhances some features and the ECG feature weights of different channels and regions are updated. In Fig. 6, the boxed area shows the ECG features that the model focuses on. Before the addition of the CBAM attention mechanism, the model was able to attend to the effective features of the ECG signal, but this attention was weak because the redundant ECG features were not effectively suppressed. With the addition of CBAM, the model's focus on effective ECG features is enhanced, or this enhancement is brought about by the effective suppression of redundant ECG features. In other words, Fig. 6 shows that the CBAM makes the effective features in the ECG signal get enhanced representation by suppressing the redundant features. We explored the role of each of these components separately. In the following experiments, we divided the dataset into a training set, a validation set, and a test set in the ratio of 8:1:1. We set the early stop in the model training and stop training when the validation set loss does not drop for 10 consecutive times. Fig. 7 shows the results of our experiments on three datasets for spatial attention, channel attention, and channel space attention, respectively. The results in Fig. 7 show that channel attention has a high level of contribution in the CBAM block. In contrast, spatial attention has a lower contribution level. A plausible explanation for this phenomenon may be that when conducting emotion-evoking experiments, subjects' emotions tend to vary continuously, i.e., from superficial to deeper degrees, rather than transiently. It can also be seen from the results in Fig. 6 that there is a

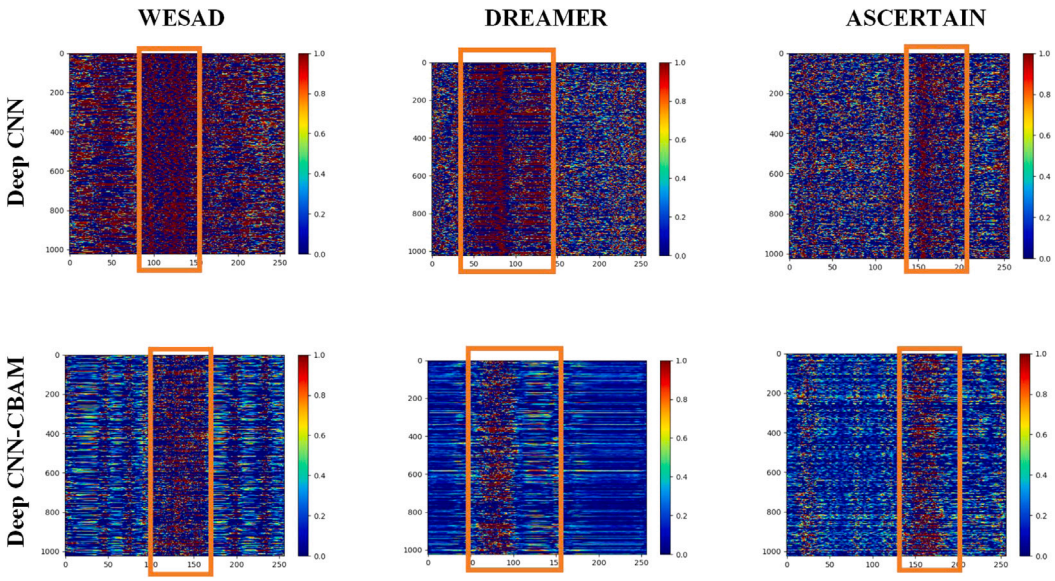


Fig. 6. ECG feature maps for the WESAD, DREAMER, and ASCERTAIN datasets before and after the addition of the CBAM attention mechanism. The boxed section indicates the feature areas that the model focuses on.

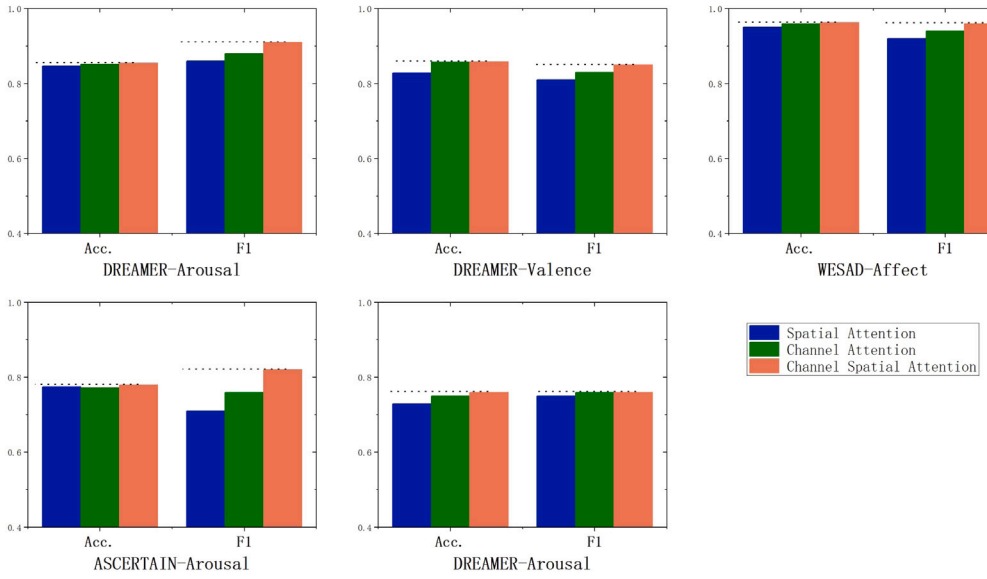


Fig. 7. The results of the DREAMER dataset and the ASCERTAIN dataset on the two evaluation dimensions Arousal and Valence are shown. In the end, the results of the WESAD dataset on the four-category emotion recognition are shown.

range in the effective part of the ECG features. In this case, it may be appropriate to focus on continuous changes in a segment of the ECG, whereas spatial attention focuses on ECG features in a particular region, which may produce biased results.

4.5.2. Impact of large-size convolution kernels

The convolutional model we propose in this study consists of six convolutional modules, each containing two CNN layers, and we set the kernel size to $[17, 15, 13, 11, 9, 7]$. Recent work [44] indicates that larger convolutional kernels can improve task performance and a convolutional kernel size of up to 32 is proposed in [9]. We explore the effect of convolutional kernel size on the performance of the final model, with a caveat: in order to simplify the experiments, we performed validation on the WESAD dataset only. Instead of performing a full 10-fold cross-validation, the dataset was divided into training, validation, and test sets based on an 8:1:1 approach. In addition, we only validate the proposed deep convolutional model (no attention mechanism is involved). Without adjusting other super parameters, we arbitrarily set the convolution kernel size (in this case, the six convolution block kernel sizes of the proposed convolution model are consistent, such as $[3, 3, 3, 3, 3, 3]$). The experimental results on the WESAD dataset are shown in Table 5. We also set up several sets of convolution kernels of different sizes, and the comparison results are shown in Table 6. Fig. 8 shows the effect of the change in convolution kernel size on the accuracy in Tables 5 and 6. As can be seen in Table 5 and Fig. 8(a), the performance of small convolution (e.g., convolution kernel sizes of 3 and 5) is slightly lower than that of large convolution (convolution kernel size greater than 5), when the kernel size of the six convolutional blocks of the proposed convolutional model is set to 17, the recognition accuracy of the model reaches the highest. However, continuing to increase the convolutional kernel size, the accuracy starts to decrease. A reasonable explanation is that when using small convolutional kernels, the convolutional window cannot completely contain

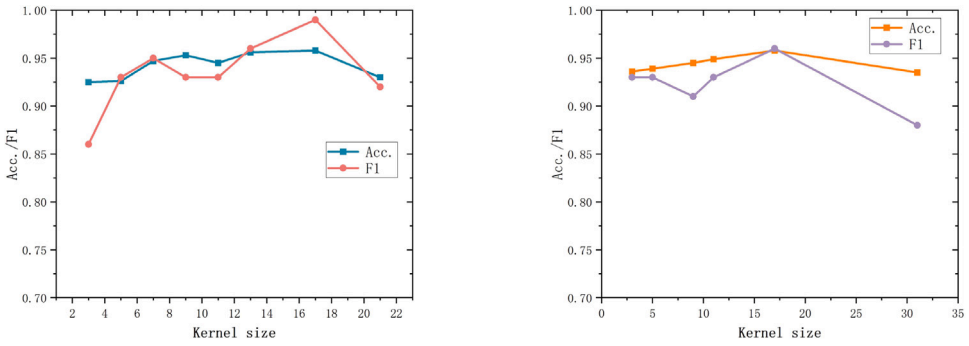


Fig. 8. (a) and (b) show the trend of the effect of the change of convolutional kernel size on the model accuracy in Tables 6 and 7, respectively.

Table 5
Set different convolution kernel sizes to classify the WESAD dataset. The parameters and FLOPs of the model with different kernel sizes are given.

Kernel size	Acc.	F1	Params (M)	FLOPs (G)
3	0.925	0.86	26	211
5	0.926	0.93	42	348
7	0.947	0.95	59	486
9	0.953	0.93	76	623
11	0.945	0.93	93	760
13	0.956	0.96	109	897
17	**0.958**	**0.99**	143	1200
21	0.930	0.92	176	1480

Table 6
Combination of convolutional kernels of different sizes for emotional state classification of the WESAD dataset. The parameters and FLOPs of the model with different kernel sizes are given.

Kernel size	Acc.	F1	Params (M)	FLOPs (G)
[3, 1, 3, 1, 3, 1]	0.936	0.93	19	157
[5, 3, 5, 3, 5, 3]	0.939	0.93	36	294
[9, 7, 9, 7, 9, 7]	0.946	0.95	69	568
[11, 9, 7, 5, 3, 1]	0.949	0.93	26	210
[17, 15, 13, 11, 9, 7]	**0.958**	**0.96**	76	621
[31, 27, 19, 17, 15, 13]	0.935	0.88	127	1037

the key features of the ECG signal, and as the convolutional kernel size increases, the global features of the ECG signal can be extracted by the convolutional network. In addition, Effective Receptive Field (ERF) theory shows that $ERF \propto O(K\sqrt{L})$ [45], where K represents the convolutional kernel size and L represents the network depth. This indicates that increasing the convolutional kernel or increasing the number of convolutional layers can expand the effective sensory field of the model. However, as the size of the convolutional kernel continues to increase, especially when the convolutional kernel in the deeper layers of the network becomes larger, the model loses some detailed information and is unable to learn fine features, leading to a drop in performance. Fig. 8(b) shows that as the size of the convolution kernel of the model gradually increases, the accuracy rate first increases and then starts to decrease. For example, after the convolution kernel size is increased from [3, 1, 3, 1, 3, 1] to [9, 7, 9, 7, 9, 7], the accuracy rate increases by 1.0%. After the convolutional kernel size was upgraded from [17, 15, 13, 11, 9, 7] to [31, 27, 19, 17, 15, 13], the accuracy decreased by 2.3%. In addition, Tables 5 and 6 contain the number of parameters of the model for different convolutional kernel sizes, and the number of parameters and the computational resources occupied by the model gradually increases with the increase of the convolutional kernel size. In fact, as the convolution layer deepens, the more complex the features to be extracted and the size of the convolution kernel should be reduced to ensure that the model can capture the features in sufficient detail.

Table 7
The performance of the dilated convolution is compared with the normal convolution for the same sensory field size. We selected the optimal parameters in Tables 5 and 6 for the comparison.

Convolution type	Kernel size	Acc.	F1	Dilated ratio
Normal convolution	[17, 17, 17, 17, 17, 17]	**0.958**	**0.99**	–
	[9, 9, 9, 9, 9, 9]	0.954	0.96	2
Dilated convolution	[5, 5, 5, 5, 5, 5]	0.947	0.95	4
	[3, 3, 3, 3, 3, 3]	0.936	0.94	8
Normal convolution	[17, 15, 13, 11, 9, 7]	**0.958**	**0.96**	–
Dilated convolution	[9, 8, 7, 6, 5, 4]	0.957	0.95	2

Dilated convolution is another way to increase the perceptual field, which performs cross-distance sampling by adding some voids (dilated ratio) in the middle of the convolution kernel. The actual convolution kernel size of the expanded convolution can be expressed by the following equation, where k represents the input convolution kernel size, r represents the dilated ratio, and k' represents the convolution kernel size of the dilated convolution equivalent.

$$k' = r \times (k-1) + 1 \tag{13}$$

We implement the dilated convolution according to the optimal parameters in Tables 5 and 6. Table 7 shows the performance comparison between the dilated convolution and the normal convolution with the same perceptual field size. Dilated convolution is a sparse sampling method, which makes it possible to have a smaller number of parameters. However, the superposition of multiple layers of dilated convolution will cause some features not to be sampled, resulting in the loss of information continuity. In the experiments, the accuracy showed a decreasing trend when we fixed the model perceptual field and set the dilated ratio to 2, 4, and 8. In addition, the accuracy of the dilated convolution with an dilated ratio of 2 in the experiments is reduced by 0.1% compared with that of the normal convolution.

In short, our experiments on the WESAD dataset indicate that increasing the convolutional kernel size can improve the model performance to some extent, even if this brings some increase in computational resources. In addition, gradually decreasing the convolutional kernel size as the number of convolutional layers deepens is a good strategy. In the proposed model, we set the kernel size to [17, 15, 13, 11, 9, 7], which may not be the optimal combination, but it is the result of the tradeoff between the model performance and the number of parameters.

5. Conclusion

In this work, we propose a deep convolutional neural network incorporating attentional mechanisms for ECG signal emotion recognition. The proposed model uses convolutional neural networks to extract ECG features and a weighted representation of ECG features in both channel

and space dimensions through a channel space attention mechanism. Three publicly available datasets, WESAD, DREAMER, and ASCERTAIN are used in this study to perform emotion recognition. We set new state-of-the-art results for the multi-class classification of the three datasets. In the paper, the contribution levels of channel attention and spatial attention are analyzed. We find that channel attention has better performance in the ECG task. In addition, we used large convolutional kernels, which are often neglected, in the ECG emotion recognition task. We found that the increase in ERF may be critical for the ECG task. Using some large convolutional kernels instead of small convolutional kernels can improve the performance of CNN in the ECG task.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was jointly supported by the National Natural Science Foundation of China under Grant No. 62272081, No. 61803072, No. 61873044 and No. 62062034, Natural Science Foundation of Liaoning Province, China under Grant 2021-MS-111, in part by the Fundamental Research Funds for the Central Universities, China under Grant No. DUT22YG128. The authors would like to express their thanks to these funding bodies.

References

[1] P.J. Bota, C. Wang, A.L.N. Fred, H. Plácido Da Silva, A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals, IEEE Access 7 (2019) 140990–141020, http://dx.doi.org/10.1109/ACCESS.2019.2944001.

[2] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing WESAD, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 400–408, http://dx.doi.org/10.1145/3242969.3242985.

[3] M.A. Hasnul, N.A.A. Aziz, S. Alelyani, M. Mohana, A.A. Aziz, Electrocardiogram-based emotion recognition systems and their applications in healthcare—A review, Sensors 21 (15) (2021).

[4] A. Sepas-Moghaddam, A. Etemad, P.L. Correia, F. Pereira, A deep framework for facial emotion recognition using light field images, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII, 2019, pp. 1–7, http://dx.doi.org/10.1109/ACII.2019.8925445.

[5] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, B.W. Schuller, Survey of deep representation learning for speech emotion recognition, IEEE Trans. Affect. Comput. (2021) 1, http://dx.doi.org/10.1109/TAFFC.2021.3114365.

[6] M.A. Hashmi, Q. Riaz, M. Zeeshan, M. Shahzad, M.M. Fraz, Motion reveal emotions: Identifying emotions from human walk using chest mounted smartphone, IEEE Sens. J. 20 (22) (2020) 13511–13522, http://dx.doi.org/10.1109/JSEN.2020.3004399.

[7] F. Ahmed, A.S.M.H. Bari, M.L. Gavrilova, Emotion recognition from body movement, IEEE Access 8 (2020) 11761–11781, http://dx.doi.org/10.1109/ACCESS.2019.2963113.

[8] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, J. Wang, SST-EmotionNet: Spatial-Spectral-Temporal Based Attention 3D Dense Network for EEG Emotion Recognition, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2909–2917, http://dx.doi.org/10.1145/3394171.3413724.

[9] P. Sarkar, A. Etemad, Self-supervised ECG representation learning for emotion recognition, 2020.

[10] F. Al Machot, A. Elmachot, M. Ali, E. Al Machot, K. Kyamakya, A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors, Sensors 19 (7) (2019) http://dx.doi.org/10.3390/s19071659.

[11] F. Agrafioti, D. Hatzinakos, A.K. Anderson, ECG pattern analysis for emotion detection, IEEE Trans. Affect. Comput. 3 (1) (2012) 102–115, http://dx.doi.org/10.1109/T-AFFC.2011.28.

[12] M. Baghizadeh, K. Maghooli, F. Farokhi, N.J. Dabanloo, A new emotion detection algorithm using extracted features of the different time-series generated from ST intervals poincaré map, Biomed. Signal Process. Control 59 (2020) 101902, http://dx.doi.org/10.1016/j.bspc.2020.101902.

[13] S. Nita, S. Bitam, M. Heidet, A. Mellouk, A new data augmentation convolutional neural network for human emotion recognition based on ECG signals, Biomed. Signal Process. Control 75 (2022) 103580, http://dx.doi.org/10.1016/j.bspc.2022.103580.

[14] S. Katsigiannis, N. Ramzan, DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices, IEEE J. Biomed. Health Inf. 22 (1) (2018) 98–107, http://dx.doi.org/10.1109/JBHI.2017.2688239.

[15] S. Qiu, T. Fan, J. Jiang, Z. Wang, Y. Wang, J. Xu, T. Sun, N. Jiang, A novel two-level interactive action recognition model based on inertial data fusion, Inform. Sci. 633 (2023) 264–279, http://dx.doi.org/10.1016/j.ins.2023.03.058.

[16] B. Hwang, J. You, T. Vaessen, I. Myin-Germeys, C. Park, B.-T. Zhang, Deep ECGNet: An optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals, Telemed. E-Health 24 (10) (2018) 753–772, http://dx.doi.org/10.1089/tmj.2017.0250, PMID: 29420125.

[17] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, IEEE, 2016.

[19] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023, http://dx.doi.org/10.1109/TPAMI.2019.2913372.

[20] Z. Qin, P. Zhang, F. Wu, X. Li, FcaNet: Frequency channel attention networks, 2020, CoRR abs/2012.11879, arXiv:2012.11879.

[21] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: Convolutional block attention module, 2018, CoRR abs/1807.06521, arXiv:1807.06521.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, CoRR abs/1706.03762. arXiv:1706.03762.

[23] R. Subramanian, J. Wache, M.K. Abadi, R.L. Vieriu, S. Winkler, N. Sebe, ASCERTAIN: Emotion and personality recognition using commercial sensors, IEEE Trans. Affect. Comput. 9 (2) (2018) 147–160, http://dx.doi.org/10.1109/TAFFC.2016.2625250.

[24] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, C.-H. Hung, Automatic ECG-based emotion recognition in music listening, IEEE Trans. Affect. Comput. 11 (1) (2020) 85–99, http://dx.doi.org/10.1109/TAFFC.2017.2781732.

[25] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang, Z. Zhao, Detecting work-related stress with a wearable device, Comput. Ind. 90 (2017) 42–49, http://dx.doi.org/10.1016/j.compind.2017.05.004.

[26] M. Bahrami, M. Forouzanfar, Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms, IEEE Trans. Instrum. Meas. 71 (2022) 1–11, http://dx.doi.org/10.1109/TIM.2022.3151947.

[27] W. Robert, Levenson, The autonomic nervous system and emotion, Emot. Rev.: J. Int. Soc. Res. Emot. 6 (2) (2014) 100–112.

[28] A. Sepúlveda, F. Castillo, C. Palma, M. Rodriguez-Fernandez, Emotion recognition from ECG signals using wavelet scattering and machine learning, Appl. Sci. 11 (11) (2021).

[29] T. Dissanayake, Y. Rajapaksha, R. Ragel, I. Nawinne, An ensemble learning approach for electrocardiogram sensor based human emotion recognition, Sensors 19 (20) (2019) http://dx.doi.org/10.3390/s19204495.

[30] R. Harper, J. Southern, A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat, IEEE Trans. Affect. Comput. 13 (2) (2022) 985–991, http://dx.doi.org/10.1109/TAFFC.2020.2981610.

[31] J. Lin, S. Pan, C.S. Lee, S. Oviatt, An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2069–2072, http://dx.doi.org/10.1145/3357384.3358160.

[32] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, A. Etemad, A Transformer Architecture for Stress Detection from ECG, ISWC '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 132–134, http://dx.doi.org/10.1145/3460421.3480427.

[33] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, W. Zhang, A systematic review on affective computing: Emotion models, databases, and recent advances, Inf. Fusion 83–84 (2022) 19–52, http://dx.doi.org/10.1016/j.inffus.2022.03.009.

[34] J. Selvaraj, M. Murugappan, K. Wan, S. Yaacob, Classification of emotional states from electrocardiogram signals: A non-linear approach based on hurst, BioMed. Eng. OnLine 12 (1) (2013) 44.

[35] X. Wang, Y. Guo, J. Ban, Q. Xu, C. Bai, S. Liu, Driver emotion recognition of multiple-ECG feature fusion based on BP network and D–S evidence, IET Intell. Transp. Syst. 14 (8) (2020) 815–824, http://dx.doi.org/10.1049/iet-its.2019.0499.

[36] A. Goshvarpour, A. Abbasi, A. Goshvarpour, An accurate emotion recognition system using ECG and GSR signals and matching pursuit method, Biomed. J. 40 (6) (2017) 355–368, http://dx.doi.org/10.1016/j.bj.2017.11.001.

[37] J. Wang, X. Qiao, C. Liu, X. Wang, Y. Liu, L. Yao, H. Zhang, Automated ECG classification using a non-local convolutional block attention module, Comput. Methods Programs Biomed. 203 (2021) 106006, http://dx.doi.org/10.1016/j.cmpb.2021.106006.

[38] R. Ge, T. Shen, Y. Zhou, C. Liu, L. Zhang, B. Yang, Y. Yan, J.-L. Coatrieux, Y. Chen, Convolutional squeeze-and-excitation network for ECG arrhythmia detection, Artif. Intell. Med. 121 (2021) 102181, http://dx.doi.org/10.1016/j.artmed.2021.102181.

[39] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, Z. Cui, MPED: A multi-modal physiological emotion database for discrete emotion recognition, IEEE Access 7 (2019) 12177–12191, http://dx.doi.org/10.1109/ACCESS.2019.2891579.

[40] S. Hu, W. Cai, T. Gao, M. Wang, A hybrid transformer model for obstructive sleep apnea detection based on self-attention mechanism using single-lead ECG, IEEE Trans. Instrum. Meas. 71 (2022) 1–11, http://dx.doi.org/10.1109/TIM.2022.3193169.

[41] D. Watson, L.A. Clark, A. Tellegen, Development and validation of brief measures of positive and negative affect: The PANAS scales, J. Pers. Soc. Psychol. 54 (6) (1988) 1063–1070.

[42] M.M. Bradley, P.J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, J. Behav. Ther. Exp. Psychiatry 25 (1) (1994) 49–59.

[43] D. Kingma, J. Ba, Adam: A method for stochastic optimization, Comput. Sci. (2014).

[44] X. Ding, X. Zhang, Y. Zhou, J. Han, G. Ding, J. Sun, Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs, 2022, arXiv e-prints.

[45] W. Luo, Y. Li, R. Urtasun, R.S. Zemel, Understanding the effective receptive field in deep convolutional neural networks, 2017, CoRR abs/1701.04128, arXiv:1701.04128.