applied sciences

MDPI

Article

Psychological Stress Detection According to ECG Using a Deep Learning Model with Attention Mechanism

Pengfei Zhang [1,2], Fenghua Li [3], Lidong Du [1,4], Rongjian Zhao [1,2], Xianxiang Chen [1,4], Ting Yang [5] and Zhen Fang [1,2,4,*]

1 Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100000, China; 17611113966@163.com (P.Z.); lddu@mail.ie.ac.cn (L.D.); rongjian8780@163.com (R.Z.); cxxvac@163.com (X.C.)
2 School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100000, China
3 Institute of Psychology, Chinese Academy of Sciences, Beijing 100000, China; lifh@psych.ac.cn
4 Personalized Management of Chronic Respiratory Disease, Chinese Academy of Medical Sciences, Beijing 100000, China
5 China-Japan Friendship Hospital, Beijing 100000, China; zryyyangting@163.com
* Correspondence: zfang@mail.ie.ac.cn

Abstract: To satisfy the need to accurately monitor emotional stress, this paper explores the effectiveness of the attention mechanism based on the deep learning model CNN (Convolutional Neural Networks)-BiLSTM (Bi-directional Long Short-Term Memory) As different attention mechanisms can cause the framework to focus on different positions of the feature map, this discussion adds attention mechanisms to the CNN layer and the BiLSTM layer separately, and to both the CNN layer and BiLSTM layer simultaneously to generate different CNN–BiLSTM networks with attention mechanisms. ECG (electrocardiogram) data from 34 subjects were collected on the server platform created by the Institute of Psychology of the Chinese Academy of Science and the researches. It verifies that the average accuracy of CNN–BiLSTM is up to 0.865 without any attention mechanism, while the highest average accuracy of 0.868 is achieved using the CNN–attention–based BiLSTM.

Keywords: ECG; psychological stress; deep learning; attention; CNN; BiLSTM

Citation: Zhang, P.; Li, F.; Du, L.; Zhao, R.; Chen, X.; Yang, T.; Fang, Z. Psychological Stress Detection According to ECG Using a Deep Learning Model with Attention Mechanism. *Appl. Sci.* **2021**, *11*, 2848. https://doi.org/10.3390/app11062848

Academic Editor: Fabio La Foresta

Received: 30 January 2021
Accepted: 11 March 2021
Published: 23 March 2021

1. Introduction

In today's society, as greater psychological stress is experienced, people are more prone to suffer harm caused by that high stress. High psychological stress not only reduces the efficiency of study and work; it can also endanger one's physical health. High long-term stress can even induce depression and addiction [1,2]. Thus, people have started to pay more attention to mental health. Many methods can be used to deal with mental health. Among them, Ecological Momentary Assessment (MAE) [3] and Just-in-Time Adaptive Interventions (JITA) [4] are effective. Both require real-time monitoring of stress. With the development of wearable technology and computer technology, real-time monitoring of emotional stress through physiological parameters becomes possible. Physiologists have learned that psychological stress is related to the human autonomic nervous system, and each system can affect the other [5]. Therefore, physiological parameter monitoring methods provide the possibility of real-time monitoring of emotional stress. The autonomic nervous system includes the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). When humans suffer from stress, the sympathetic nervous system become excited and generates adrenal hormones, which then cause the heart to beat faster and breathing to shorten. This change strengthens the human body's functions, making it easier to meet challenges [6]. After the stress is over, the parasympathetic nervous system become excited to help the body calm down, which causes the heart to beat slower and breathing to lengthen. Therefore, psychological pressure in humans can be inferred by

monitoring changes in the physiological parameters, such as heart rate, breathing rate, and others. In recent years, the advent of deep learning has promoted the development of artificial intelligence, some teams have started to use deep learning techniques to speculate on psychological stress. Jin built deep neural network (DNN) models and used conditional random field to classify mental stress based on information about adolescents' social behavior [7]. Lin used a deep sparse neural network to monitor stress levels based on cross-media microblog data [8]. Although these results showed that the modules constructed by Jin and Lin were effective, the real-time monitoring still cannot be guaranteed as that model needs to input linguistic information or video information for a period of time. Hwang [9] categorized stress into two classes using 10 s ECG data with CNN (Convolutional Neural Networks) and LSTM (Long Short-Term Memory), but they did not realize the three categories of stress. The attention mechanism [10], which is among the latest deep learning technologies, could help the model to recognize important features and determine the global relationship between its features by adding adaptive weight coefficients to the hidden variables in the network. Identifying how to use attention mechanism has been a hot topic. Guo used a visual attention mechanism to recognize the traffic signs [11]. Huang proposed a convolutional attention mechanism to learn the utterance structure relevant to a task for speech emotion recognition [12]. Choi used a time attention model for healthcare data analysis and was able to achieve high accuracy [13]. These research efforts definitely showed the promise of attention mechanism in deep learning. However, there is a lack of research on how to applicate attention mechanism in the subject of monitoring emotional stress based on physiological parameters. On the basis of proposing a network using CNN [14] and BiLSTM(Bi-directional Long Short-Term Memory), we added different attention mechanisms to the CNN and BiLSTM layers separately and also to the whole network to explore the effectiveness of the attention mechanism for the task of identifying emotional stress based on ECG signals. To the best of our knowledge, this effort is the first where deep learning models were combined with attention mechanisms to detect psychological stress using ECG signals.

2. Materials and Methods

This section is divided into 5 parts, among them, Section 2.1 mainly introduces the design of stress induction experiment and the construction of data set. Section 2.2 mainly describes the structure of the CNN–BiLSTM which is used as the basis or reference of the network with mechanism. Section 2.3 mainly introduce the three attention mechanisms used in this article. Section 2.4 proposed the fusion of the three attention mechanisms and CNN–BiLSTM.

2.1. Experiments and Data Acquisition

To facilitate the collection, management, and analysis of ECG data, we built a data processing system that consists of a frontend and a server. The server, based on Alibaba Cloud, was designed with flask as its framework, as depicted in Figure 1. We selected MongoDB as the database. Compared to the other relational databases, such as MySQL, MongoDB is more convenient for storing time series data. Whenever the frontend collects 10 s of ECG data, it sends that data to the platform through the 4G network. After the server receives the data, if necessary, three processes generated by Gunicorn are executed at the same time. To avoid data loss caused by the server analyzing prior ECG data when the new data arrive, the ECG data need to be stored. Each process thus starts with two tasks. Task1 is responsible for storing the data in the database and Task2 is responsible for analyzing the ECG and then storing the result in the same database. Thus, the server can store the data without interruption and also avoid data loss as much as possible. The frontend consists of an application for cell phones and sticker-type acquisition devices, the device is as shown in Figure 2. By using a cellphone to temporarily store and transfer data, we made the attachable ECG acquisition device as compact and lightweight as possible. The sticky

acquisition device samples the ECG signals at a frequency of 250 Hz and continuously transmits the ECG signal via Bluetooth 4.0.

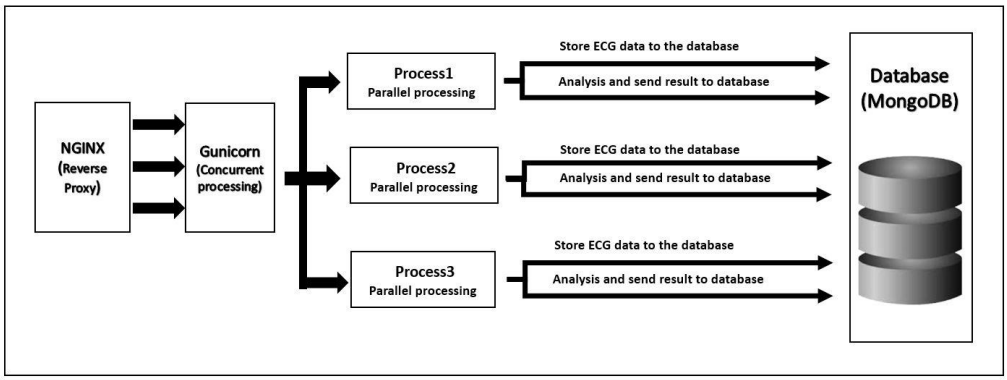

Figure 1. The cloud server designed with parallel processing for data collection, management and analysis.

Figure 2. The sticker-type ECG (electrocardiogram) acquisition device.

Common pressure-inducing methods include the ice water test, speech, video and audio. As we aimed to identify three levels of psychological stress, considering the controllability of pressure-induced intensity and the difficulty of operation of the experiment, we used three different computational tasks to induce three different levels of psychological stress according to the Montreal Stress Paradigm [15]. For the light-stress phase, there was no computing task for the subjects and soothing music was played to make the subject feel as relaxed as possible throughout the process. For the medium-stress phase, the subject was asked to perform simple two-digit addition and subtraction tasks without any time limitation. During the high-stress phase, to induce as much stress as possible, we added a time limitation and reward measures to the task. The process for this stress-inducing experiment was developed together with the Chinese Academy of Psychology and us, as shown in Figure 3. The total duration of each experiment was 15 min. The first 5 min was the rest phase, then the medium-stress phase and the high-stress phase were alternately arranged. The visual analogue scale (VAS) report of psychological stress was used to gather data from the subjects every twenty seconds, which was used as the ground truth to label the data. The ECG data and self-reports, collected by the sticky acquisition equipment and its application, respectively, were transferred to the server for further processing. On the server, the ECG data were filtered by a bandpass filter ranging from 5 Hz to 11 Hz to remove baseline drift and noise, the self-reports were processed by one-hot encoding as data labels. The data and its corresponding labels constituted the data set. All models in this paper were trained and tested using a five-fold validation.

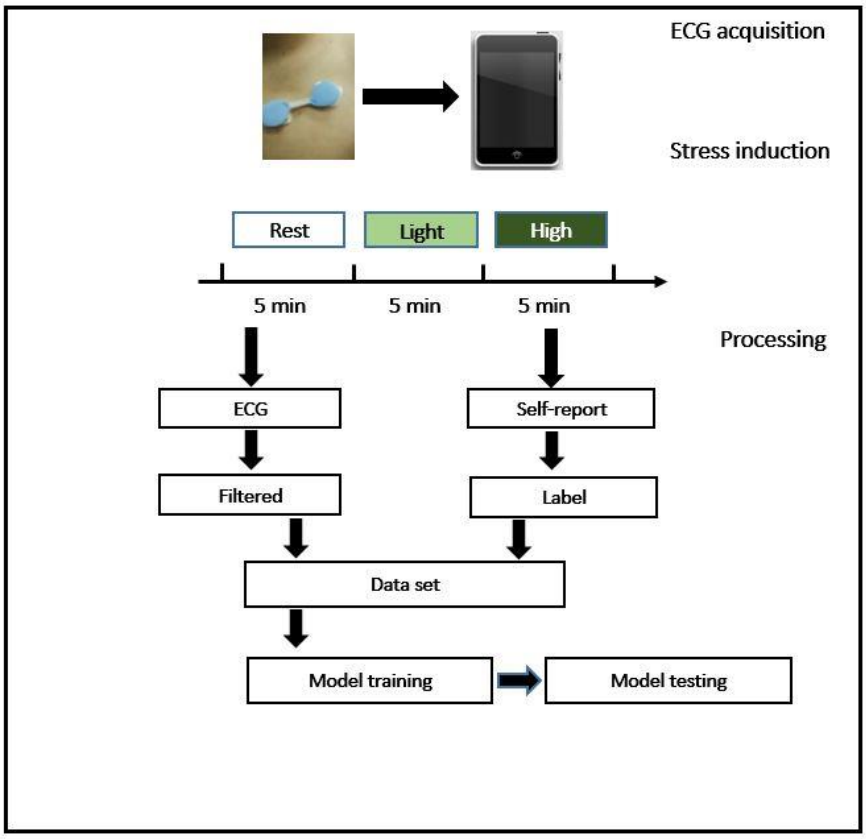

Figure 3. The experimental program.

We convened 34 subjects at the Chinese Academy of Sciences. Their ages ranged from 21 to 35 years. There were 20 men and 14 women. The subjects were informed about the procedure, and we ensured that there was no history of drinking for the last three days prior to the experiments.

Each data point of the data set was the 10 s of ECG, and there were 3097 total data points. For all the experiments, we randomly chose 80% of the data set for training and 20% of the dataset for testing. The accuracy and specificity of both were calculated using the average of the five results based on the five-fold cross-validation.

2.2. CNN-BiLSTM

The excellent performance of CNN in the field of image processing is one of the most representative examples of deep learning. Due to the characteristics of weight sharing and local connection, CNN is good at extracting spatial and structural features. The number of nodes and parameters needed by a CNN network are exponentially lower. A typical convolutional neural network contains a convolutional layer and a pooling layer. The convolutional layer performs autocorrelation operations on the local data as intercepted by the window function and the convolution kernel, so that the features related to the convolution kernel is extracted. The number of features can be increased by increasing the number of convolution kernels in a certain range. To highlight the effective feature extraction and the decrease in time required for training, the pooling layer mainly downsamples the features generated by the convolution layer through either maximum sampling or average sampling. Considering the ability of the convolutional networks to extract spatial features, we added CNN to the first layer of the network to extract the structural characteristics of the ECG signals.

In general, neural networks do not consider the back-and-forth connection of signals in real time, so they are not effective at processing time series signals. The advent of RNN [16] (Recurrent Neural Network) overcame this problem. RNN divides the signal

into units that contain complete local information and iterate over each unit according to the time order. During that iteration, the input variables not only include the input of the current, but also the stated value of the previous unit, so that the current unit can also consider information from prior units. This process enables RNNs to extract features over time and process time series signals such as natural language.

LSTM is a variant of the RNN. Compared to other RNNs, LSTM adds a memory gate and a forget gate when judging the importance of historical information. With the memory gate and the forget gate, LSTM is more capable of capturing useful historical information in long series data. The time series is back-and-forth connected, and in addition to historical information, we can infer current status according to future trends. BiLSTM iterates from the beginning and the end simultaneously with two LSTM chains so that both historical features and future features can be extracted. Based on the ability of BiLSTM to extract features from a global perspective, we used two BiLSTM layers after the CNN to extract more abstract effective features, and the constructed CNN-BiLSTM is shown in Figure 4.

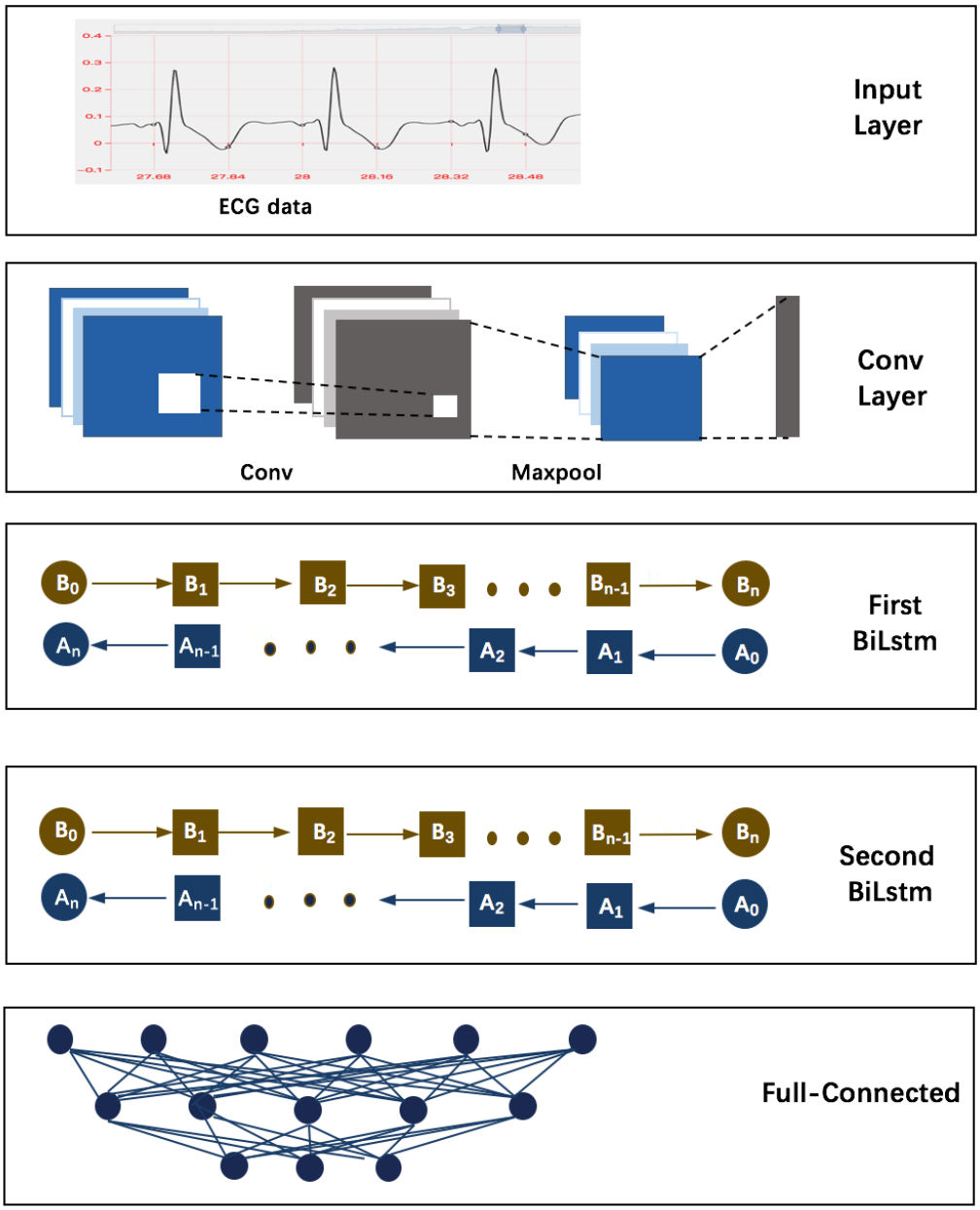

Figure 4. The CNN (Convolutional Neural Networks)–BiLSTM (Bi–directional Long Short–Term Memory) network.

The structures of CNN and BiLSTM are shown in Table 1. The optimal features of the convolutional network and the BiLSTM network were determined through multiple experiments. In the CNN, the rectified linear unit (ReLU) after the convolutional (conv) layer was used as an activating function to highlight scarcity and non-linear mapping. To speed up the training of the network, batch normalization was used to normalize the parameters of the conv layer. Cross-entropy was selected as the loss function. To reduce overfitting and increase the generalization ability of the model, we added the second-order normal form of the parameters to the loss and added a drop out layer so as to randomly discard some of the nodes during the training process. Adam was selected as the trainer.

Table 1. The structure of the CNN (Convolutional Neural Networks) and BiLSTM (Bi-directional Long Short-Term Memory).

CNN	
Convolutional layer	Filter = 32, kernel size = 200, stride = 8
Rectified Linear Unit	
Batch normalization + dropout (0.70)	
Max pooling	Pool size = 8, stride = 8
First BiLSTM	
LSTM layer + dropout (0.70)	Step size = 25, LSTM size = 40
The stitching method of LSTM	Adding
Second BiLSTM	
LSTM layer + dropout (0.70)	Step size = 25, LSTM-size = 40
The stitching method of LSTM	Adding

Compared with stochastic gradient descent (SGD) and momentum, Adam can adaptively adjust the learning rate for each parameter, so that the various parameters in the network can be trained better. We set the initial training rate at 0.2×10^{-4}, the number of epochs to 1500, and the batch size to 30.

2.3. The Attention Mechanism with CNN-BiLSTM

Attention mechanisms can tell the network where to focus and highlight the effective features by adding adaptive weights to hidden parameters of the network. As different attention mechanisms can cause the framework to focus on different positions of the feature map, we tried to use CBAM (The Convolutional Block Attention Module), Non-Local Neural Networks and DA-NET (Dual Attention Network) adding attention mechanisms to the CNN layer, use the Attention-Based Bidirectional Long Short-Term Memory Networks instead of BiLSTM to generate different CNN-BiLSTM network with attention mechanisms.

2.3.1. The Convolutional Block Attention Module (CBAM)

The convolutional block attention module (CBAM) [17] is an effective attention module to use for convolutional neural networks. Instead of directly computing the attention map, CBAM infers the attention map along the channel and the spatial dimensions with the channel attention module and the spatial module as shown in Figure 5. In the channel attention module as Figure 6 showed, for the input feature in each channel, the average-pooled features and the max-pooled features are calculated simultaneously. Then, the MLP (Multi-Layer Perceptron) and sigmoid function are used to fuse the two features and generate channel attention map.

In the spatial attention module as shown in Figure 7, in order to generate the spatial attention map, the spatial attention module uses average pooling and max-pooling operations on the channel attention feature to generate the average-pooled features and the max-pooled features across the channel. Then, the concatenation and convolution are used to produce the spatial attention map.

Finally, the attention feature $F \in \mathbb{R}^{L*F}$ after CNN can be calculated with channel attention features and spatial attention features, wherein L is the length of the ECG and F is the number of convolutional filters on the last convolutional layer.

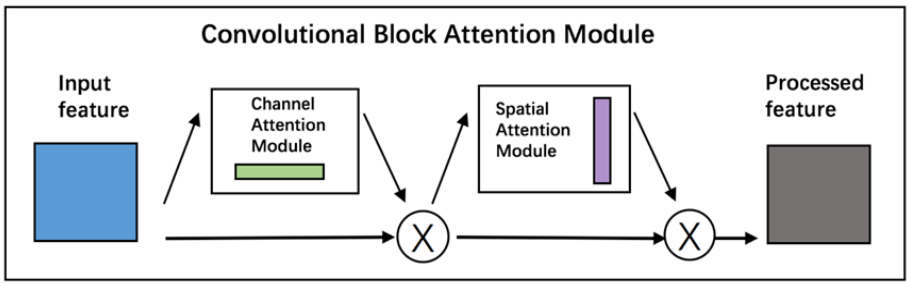

Figure 5. Overview of the convolutional block attention module (CBAM) including two models: the channel attention module and the spatial attention module.

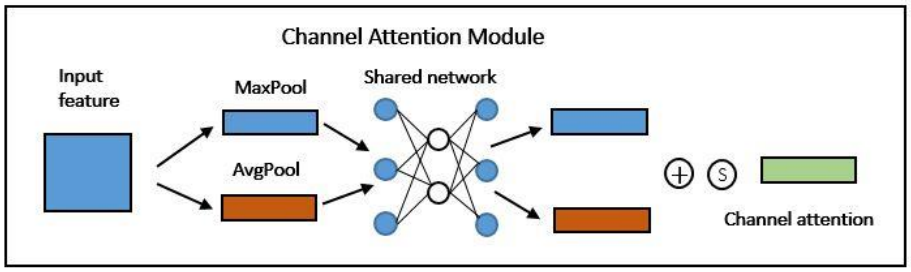

Figure 6. The channel attention module.

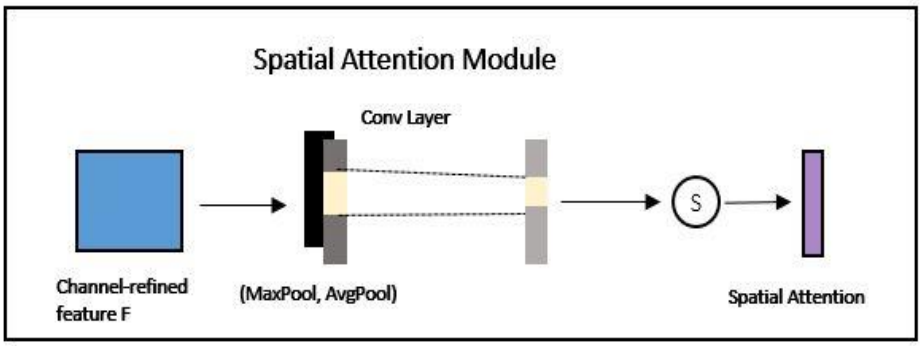

Figure 7. Spatial attention module.

2.3.2. Non-Local Neural Networks

A non-local neural network [18] is an attention mechanism used on CNN layers. It is a block module that performs non-local operations to capture long-range dependencies as a time dimension, length dimension, or width dimension. The generic non-local operation can be defined as follows:

$$y_i = 1/C(x) \sum_{\forall j} f(x_i, x_j) g(x_j) \quad (1)$$

In the current article, i is the index of the length dimension generated by the convolutional layer, j is the index of the position in the length dimension related to i, x is the output of the last convolutional network, and y is the output that is the same size as x. A pairwise function f computes the response between i and all values of j. The unary function, g, computes an approximate local response of x_j. $C(x)$ is used for normalization. Compared with a fully connected layer for computing the relationship between x_i and x_j, the f function used in non-local neural networks is better able to reflect the non-linear relationship between x_i and x_j. We selected Embedded Gaussian as the f function and Softmax as the C function in Equation (1) as Equation (2) showed, where θ, ϕ and g represent a one-dimensional convolutional with different convolutional kernels. The final output Z is the sum of x and y as shown in Figure 8. The Value B in Figure 8 represents the size of the batch in this article,

W represents the length of the ECG signal, and F represents the number of filters in the last convolutional network.

$$y = softmax\left(x^T W_\theta^T W_\phi x\right) g(x) \quad (2)$$

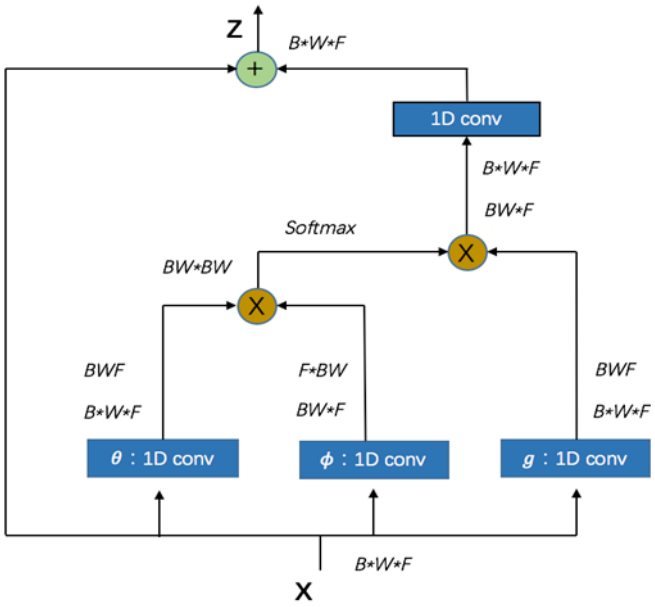

Figure 8. The non-local-neural network with embedded Gaussian as function *F*.

2.3.3. The Dual Attention Network (DA-NET)

The dual attention network (DA-NET) [19] used after CNN as an attention mechanism that can capture feature dependencies found in the spatial and channel dimensions. As shown in Figure 9, DA-NET contains two parallel attention modules. For the position attention module, a self-attention mechanism is introduced to capture the spatial dependencies between any two positions of the feature maps. For the channel attention module, a similar self-attention mechanism is used to capture the channel dependencies between any two channel maps. Then, a certain channel dependence is calculated by a weighted sum of all the channel maps. The weight is decided by the similarity between the corresponding two channels. Finally, the outputs of these two modules are fused by a summation and processed via a convolutional layer to generate the output of DA-NET.

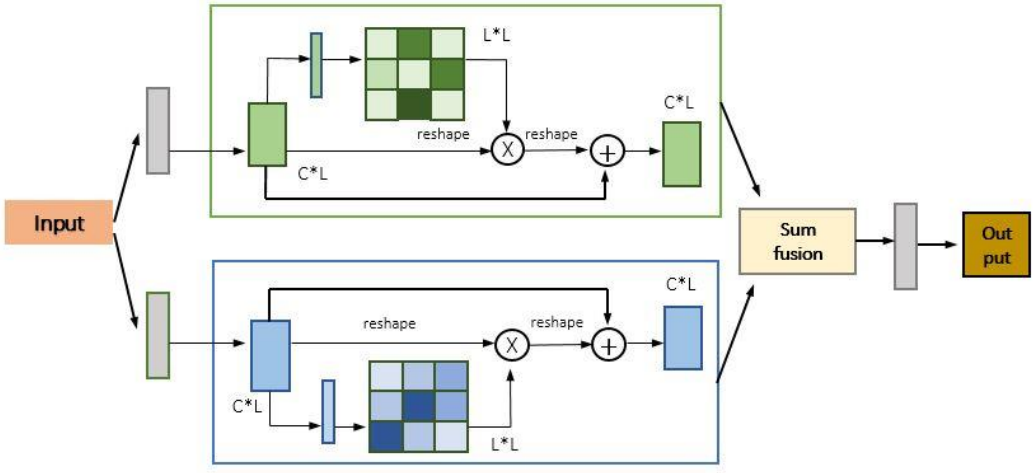

Figure 9. Overview of the Dual Attention Network (DA-NET). The grey boxes represent the convolutional layers.

2.3.4. Attention-Based Bidirectional Long Short-Term Memory Networks

The attention-based BiLSTM [20] is used for relationship classification, which can capture the most important semantic information in a sentence without using any of the features derived from the NLP (Natural Language Processing) system. As ECG signals are time series, like natural language, signal segments are temporally correlated between the forward and the afterward chains. We used an attention-based BiLSTM in the CNN-BiLSTM to identify the important features. As the attention-based BiLSTM can automatically focus on words that have a decisive effect on the classification in a sentence, in this effort, each sample that corresponds to a 10 s long ECG data point is seen as a sentence, and every 25 points in the sample is seen as a word. That length of 25 points was determined experimentally.

The attention-based BiLSTM was altered and used as shown in Figure 10. In the LSTM layer, the BiLSTM was used to obtain high level features from the input. The attention layer, which is on the LSTM layer, can produce a weight vector and merge word-level features from each time step into a sentence-level feature vector by multiplying the weight vector. In the attention layer, $\overrightarrow{H}$ represents the output of the forward LSTM chain and $\overleftarrow{H}$ represents the output of the afterward LSTM chain. T is the length of each ECG data point, which corresponds to 10 s. The representation y of the ECG sample is calculated as the sum of the weighted output vectors:

$$M = \tanh(H), \alpha = softmax\left(w^T M\right), y = H\alpha^T \quad (3)$$

where $H \in \mathbb{R}^{d*T}$ (d is the size of the LSTM and w is a trained parameter vector). Finally, h^*, which is used for classification, is calculated by:

$$h^* = tanh(r) \quad (4)$$

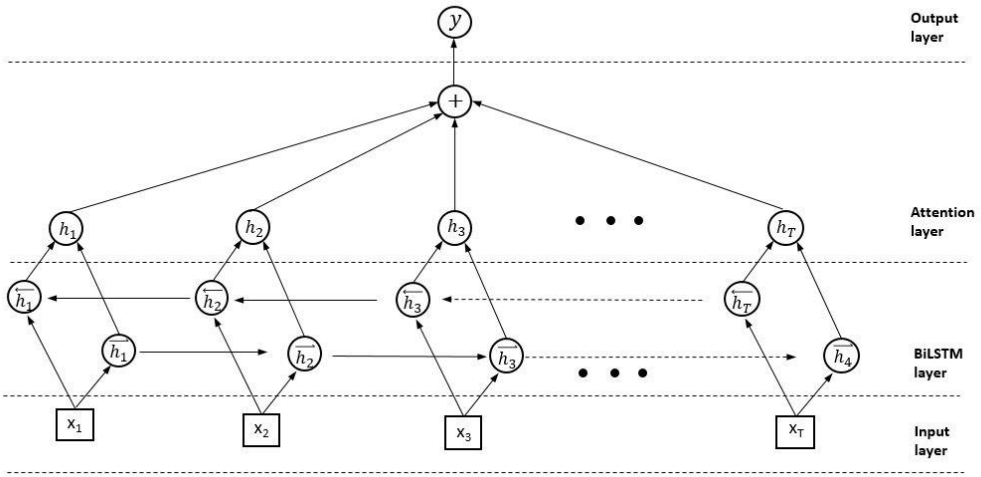

Figure 10. Attention-based bidirectional long short-term memory networks.

2.4. The Structure and Parameters of the Models with Attention Mechanism

To satisfy the need to accurately monitor emotional stress, which is based on building the framework (CNN-BiLSTM) with deep learning tools, we attempted to improve CNN-BiLSTM using the attention mechanism. For the CNN layer in the CNN-BiLSTM, we used CBAM, a non-local neural network, and DA-NET separately to make the CNN-BiLSTM focus on efficient features and filter out any invalid features. For the BiLSTM layer in the CNN-BiLSTM, attention-based BiLSTM was used instead of normal BiLSTM to increase the attention of the CNN-BiLSTM. As different attention mechanisms make the framework focus on different positions of the feature map, to find the most effective

attention mechanism to use with CNN-BiLSTM for this particular issue, we separately constructed the CNN-BiLSTM with the CBAM framework, the CNN-BiLSTM with the non-local neural network, the CNN-BiLSTM with DA-NET, the CNN-attention-based BiLSTM, the CNN-attention-based BiLSTM with CBAM, and the CNN-attention-based BiLSTM with the non-local neural network for verification. The specific parameters of the network were obtained through repeated experiments. For the CNN-BiLSTM with CBAM as shown in Table 2, we added the CBAM module after the CNN. In the CBAM, we set the drop out coefficient of the fully connected layer (K) to 0.3 and the kernel size of the convolution layer to 28. For the CNN-BiLSTM with the non-local neural network shown in Table 3, we added the non-local neural network after the CNN. In the non-local neural network, the convolutional layer was used to generate θ, ϕ, and g. By considering the training time and performance of the framework, we set the number of filters in the convolutional layer (F) to 16. For the CNN-BiLSTM with DA-NET in Table 4, the convolutional layer in the DA-NET was used to generate the feature maps B and C. Considering the training time and performance of the framework, we set the number of filters in the convolutional layer (F) to 8.

Table 2. The CNN (Convolutional Neural Networks)–BiLSTM (Bi-directional Long Short-Term Memory) with CBAM (convolutional block attention module).

CNN	
Convolutional layer	Filter = 32, kernel size = 200, stride = 8
Rectified Linear Unit	
Batch normalization + dropout (0.70)	
Max-pooling	Pool size = 8, stride = 8
CBAM	K = 0.3, kernel size = 200
First BiLSTM	
Second BiLSTM	

Table 3. The CNN (Convolutional Neural Networks)-BiLSTM (Bi-directional Long Short-Term Memory) with non-local neural network.

CNN	
Convolutional layer	Filter = 32, kernel size = 200, stride = 8
Rectified Linear Unit	
Batch normalization + dropout (0.70)	
Max-pooling	Pool size = 8, stride = 8
Non-local neural network	F = 16 (1/2*Filter)
First BiLSTM	
Second BiLSTM	

Table 4. The CNN (Convolutional Neural Networks)-BiLSTM (Bi-directional Long Short-Term Memory) with DA-NET (Dual Attention Network).

CNN	
Convolutional layer	Filter = 32, kernel size = 200, stride = 8
Rectified Linear Unit	
Batch normalization + dropout (0.70)	
Max-pooling	Pool size = 8, stride = 8
Dual Attention Network	The number of filters in the conv layer to generate B and C in the position module (F) = 8
First BiLSTM	
Second BiLSTM	

For the CNN-attention-based BiLSTM shown in Table 5, we used attention-BiLSTM instead of the first and second layers of the convolutional BiLSTM. For the CNN-attention-based BiLSTM with CBAM as shown in Table 6, the CBAM and attention-based BiLSTM were used simultaneously to obtain the full attention network with the CNN layer and the BiLSTM layer both using the attention mechanism. For the CNN-attention-based BiLSTM with a non-local neural network in Table 7, the non-local neural network and attention-based BiLSTM were both used on the CNN and the BiLSTM layer.

Table 5. The CNN (Convolutional Neural Networks)-attention-based BiLSTM (Bi-directional Long Short-Term Memory).

CNN	
Convolutional layer	Filter = 32, kernel size = 200, stride = 8
Rectified Linear Unit	
Batch normalization + dropout (0.70)	
Max-pooling	Pool size = 8, stride = 8
First Attention-Based BiLSTM	
Second Attention-Based BiLSTM	

Table 6. The CNN (Convolutional Neural Networks)-attention-based BiLSTM (Bi-directional Long Short-Term Memory) with CBAM (Channel attention module).

CNN	
Convolutional layer	Filter = 32, kernel size = 200, stride = 8
Rectified Linear Unit	
Batch normalization + dropout (0.70)	
Max-pooling	Pool size = 8, stride = 8
Non-local-neural network	F = 100 (1/2*Kernel size)
First Attention-Based BiLSTM	
Second Attention-Based BiLSTM	

Table 7. The CNN (Convolutional Neural Networks)-attention-based BiLSTM (Bi-directional Long Short-Term Memory) with non-local neural network.

CNN	
Convolutional layer	Filter = 32, kernel size = 200, stride = 8
Rectified Linear Unit	
Batch normalization + dropout (0.70)	
Max-pooling	Pool size = 8, stride = 8
Non-local neural network	F = 100 (1/2*Kernel size)
First Attention-Based BiLSTM	
Second Attention-Based BiLSTM	

3. Result

The results gathered for these frameworks are shown in Table 8. In terms of accuracy, the accuracy of the original CNN-BiLSTM without any attention mechanism was 0.865. For the frameworks with attention, the CNN-BiLSTM with CBAM, the CNN-attention-based BiLSTM with CBAM, and the CNN-attention-based BiLSTM with a non-local neural network exhibited worse performance than CNN-BiLSTM. The CNN-attention-based BiLSTM obtained the highest accuracy, 0.868, which was an improvement over CNN-BiLSTM.

The reason why CNN-attention-based BiLSTM outperformed the other frameworks is that as the BiLSTM is performed after the CNN, the features of BiLSTM, as the further processing of CNN, can more accurately characterize the category of the sample. Thus focusing on the important feature of CNN-BiLSTM is more effective. We also noticed that the CNN-attention-based BiLSTM with a non-local neural network and the CNN-attention-based BiLSTM with CBAM, which focus on the CNN and BiLSTM simultaneously, performed worse than the CNN-attention-based BiLSTM did.

Table 8. The CNN (Convolutional Neural Networks)-attention-based BiLSTM (Bi-directional Long Short-Term Memory) with non-local neural network.

Model	Accuracy	Specificity	Accuracy Compared with CNN-BiLSTM
CNN-BiLSTM	0.865	0.928	
CNN-BiLSTM with CBAM	0.862	0.926	
CNN-BiLSTM with DA-Net	0.861	0.926	↓
CNN-BiLSTM with non-local neural network	0.857	0.923	↓
CNN-attention-based BiLSTM	0.868	0.930	↑
CNN-attention-based BiLSTM with CBAM	0.862	0.926	↓
CNN-attention-based BiLSTM with non-local neural network	0.860	0.924	↓

We speculate that the reason for this different is that the attention mechanism performed on the CNN filters highlights the features that are important to the BiLSTM, which leads to the degradation of the overall model. As the CNN-BiLSTM with DA-NET and the CNN-BiLSTM with non-local neural network perform worse than the CNN-BiLSTM, DA-NET and the non-local neural network make the CNN-BiLSTM focus on certain invalid features. For specificity, the results showed that the CNN-attention-based BiLSTM also performed better than CNN-BiLSTM. The confusion matrixes of these frameworks are shown in Figure 11. It can be seen that CNN-attention based BiLSTM performed better for the detection of three levels compared with CNN-BiLSTM.

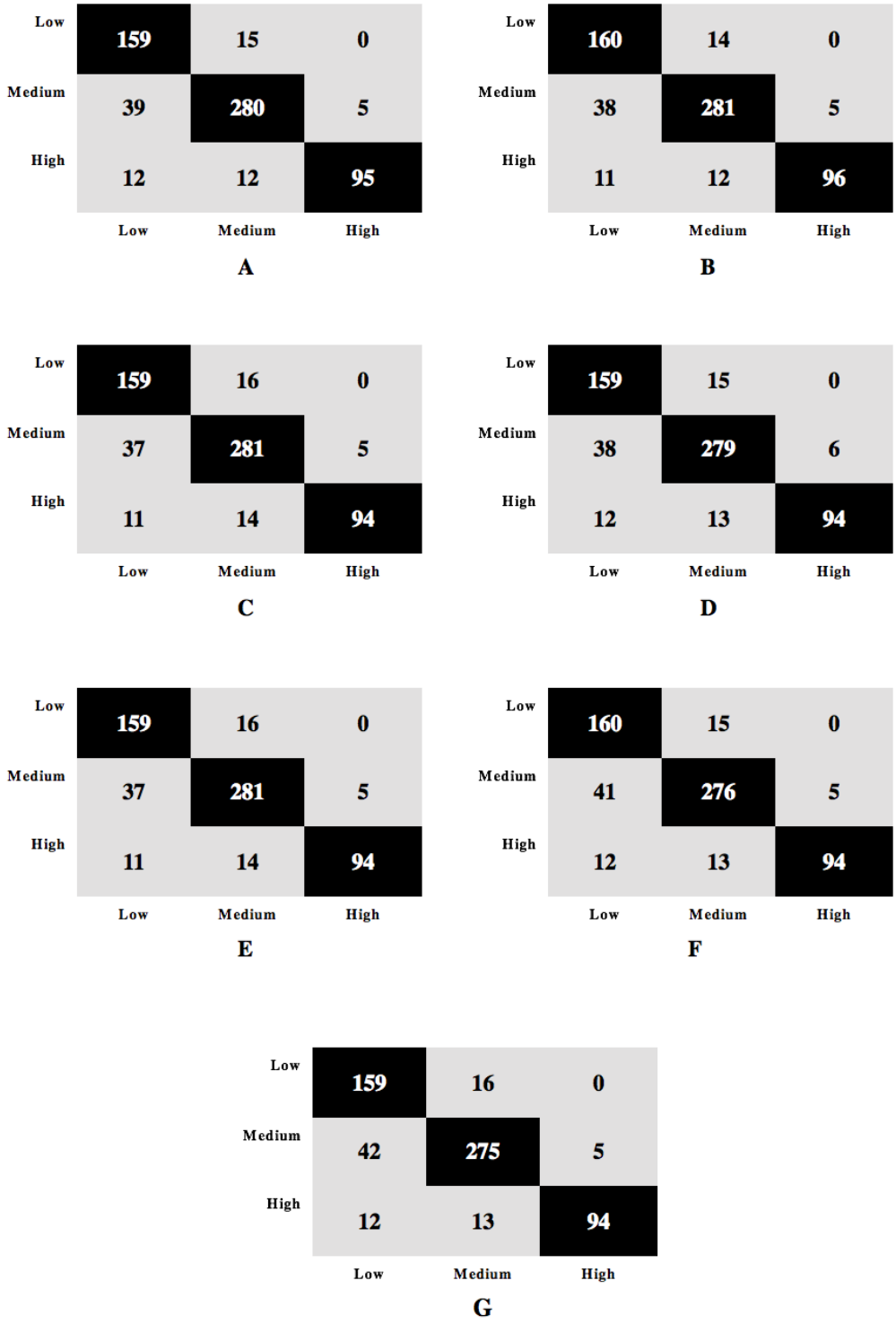

Figure 11. (**A**) The confusion matrix for the CNN (Convolutional Neural Networks)-BiLSTM (Bi-directional Long Short-Term Memory); (**B**) the confusion matrix of CNN (Convolutional Neural Networks)-attention-based BiLSTM (Bi-directional Long Short-Term Memory); (**C**) the confusion matrix of CNN-BiLSTM (Bi-directional Long Short-Term Memory) with CBAM (Convolutional Block Attention Module); (**D**) the confusion matrix of CNN (Convolutional Neural Networks)-BiLSTM (Bi-directional Long Short-Term Memory) with DA-NET (Dual Attention Network); (**E**) the confusion matrix of CNN (Convolutional Neural Networks)-attention-based BiLSTM (Bi-directional Long Short-Term Memory) with CBAM (Convolutional Block Attention Module); (**F**) the confusion matrix of CNN (Convolutional Neural Networks)-attention-based BiLSTM (Bi-directional Long Short-Term Memory) with non-local neural network; (**G**) the confusion matrix of CNN (Convolutional Neural Networks)-BiLSTM (Bi-directional Long Short-Term Memory) with non-local neural network.

4. Conclusions and Discussion

To satisfy the need to accurately monitor emotional stress based on the ECG signal, we tried to add different attention mechanisms to the CNN and BiLSTM layers of a CNN-BiLSTM network separately and to the whole network to explore the effectiveness of the attention mechanism. According to that result, we found the most effective attention model that pushed the performance by exploiting the attention mechanism. That result showed that the attention mechanism is effective for solving psychological stress recognition based on ECG signals. More importantly, the experimental results indicated that the combination of attention mechanism and RNN is more effective for this subject compared with CNN, thereby improving the combination of the attention mechanism. Thus RNN will become our future research direction. During the experiment, the impact of motion noise was not considered, so how to evaluate motion noise and eliminate its influence are also a main research focuses for the future. To the best of our knowledge, this is the first time that psychological stress has been recorded using ECG and analyzed using a deep learning framework with attention mechanism. We hope the application of an attention mechanism in psychological stress recognition using ECG can provide an important reference for other researchers.

Author Contributions: Conceptualization, P.Z. and F.L.; methodology, P.Z., L.D. and Z.F.; software, T.Y. and R.Z.; validation, X.C.; formal analysis, X.C. and L.D.; investigation, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Key Research and Development Project 2018YFC2001101, 2018YFC2001802, 2020YFC2003703, 2020YFC1512304, National Natural Science Foundation of China (Grant 62071451), and CAMS Innovation Fund for Medical Sciences (2019-I2M-5-019).

Institutional Review Board Statement: This research was reviewed and approved by the IRB of Beijing Tiantan Hospital, Capital Medical University (no. KYSQ 2019-013-01).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. As the data were generated during this study, we did not find an appropriate platform to share the data.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sauter, S.L.; Murphy, L.R.; Hurrell, J.J. Prevention of work-related psychological disorders: A national strategy proposed by the National Institute for Occupational Safety and Health (NIOSH). *Am. Psychol.* **1990**, *45*, 1146–1158. [CrossRef] [PubMed]
2. Hillebrandt, J. *Work-Related Stress and Organizational Level Interventions-Addressing the Problem at Source*; GRIN Verlag: London, UK, 2008.
3. Ebner-Priemer, U.W.; Trull, T.J. Ecological momentary assessment of mood disorders and mood dysregulation. *Psychol. Assess.* **2009**, *21*, 463–475. [CrossRef] [PubMed]
4. Spruijt-Metz, D.; Nilsen, W. Dynamic Models of Behavior for Just-in-Time Adaptive Interventions. *IEEE Pervasive Comput.* **2014**, *13*, 13–17. [CrossRef]
5. Carrasco, G.A.; Kar, L.D.V.D. Neuroendocrine pharmacology of stress. *Eur. J. Pharmacol.* **2003**, *463*, 235–272. [CrossRef]
6. Tsigos, C.; Chrousos, G.P. Hypothalamic-pituitary-adrenal axis, neuroendocrine factors and stress. *J. Psychosomat. Res.* **2002**, *53*, 865–871. [CrossRef]
7. Jin, L.; Xue, Y.; Li, Q.; Feng, L. Integrating Human Mobility and Social Media for Adolescent Psychological Stress Detection. In *International Conference on Database Systems for Advanced Applications*; Springer: Berlin/Heidelberg, Germany, 2016.
8. Lin, H.; Jia, J.; Guo, Q.; Xue, Y.; Huang, J.; Cai, L.; Feng, L. Psychological stress detection from cross-media microblog data using Deep Sparse Neural Network. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 14–18 July 2014.
9. Hwang, B.; You, J.; Vaessen, T.; Myin-Germeys, I.; Park, C.; Zhang, B.T. Deep ECGNet: An Optimal Deep Learning Framework for Monitoring Mental Stress Using Ultra Short-Term ECG Signals. *Telemed. e-Health* **2018**, *24*, 753–772. [CrossRef] [PubMed]
10. Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. *arXiv* **2015**, arXiv:1506.07503.

11. Guo, H.R.; Wang, X.J.; Zhong, Y.X.; Peng, L.U. Traffic signs recognition based on visual attention mechanism. *J. China Univ. Posts Telecommun.* **2011**, *18* (Suppl. S2), 12–16. [CrossRef]
12. Huang, C.W.; Narayanan, S.S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In Proceedings of the IEEE International Conference on Multimedia & Expo, Hong Kong, China, 10–14 July 2017.
13. Choi, E.; Bahadori, M.T.; Kulas, J.A.; Schuetz, A.; Stewart, W.F.; Sun, J. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *arXiv* **2016**, arXiv:1608.05745.
14. Mnih O'Shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
15. Dedovic, K.; Renwick, R.; Mahani, N.K.; Engert, V.; Lupien, S.J.; Pruessner, J.C. The Montreal Imaging Stress Task: Using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *J. Psychiatry Neurosci.* **2005**, *30*, 319. [PubMed]
16. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. *arXiv* **2014**, arXiv:1409.2329.
17. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
18. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
19. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
20. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 2, pp. 207–212.