Towards high-accuracy classifying attention-deficit/hyperactivity disorders using CNN-LSTM model

Cheng Wang¹,²,³, Xin Wang¹,²,₇, Xiaobei Jing¹, Hiroshi Yokoi⁴, Weimin Huang⁵, Mingxing Zhu⁶,∗, and Guanglin Li¹,³

¹ The CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, People’s Republic of China
² Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, People’s Republic of China
³ Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Chinese Academy of Sciences, Shenzhen 518055, People’s Republic of China
⁴ Department of Mechanical and Intelligent Systems Engineering and Center for Neuroscience and Biomedical Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan
⁵ Department of Neonatology, Shenzhen Children’s Hospital, Shenzhen 518038, People’s Republic of China
⁶ School of Electronics and Information Engineering, Harbin Institute of Technology, Shenzhen 518055, People’s Republic of China
⁷ The first two authors contributed equally to this work.

* Authors to whom any correspondence should be addressed.
E-mail: 458260329@qq.com and sx.chen@siat.ac.cn

Keywords: attention-deficit/hyperactivity disorder, deep learning, EEG, event-related potential

Abstract

Objective. The neurocognitive attention functions involve the cooperation of multiple brain regions, and the defects in the cooperation will lead to attention-deficit/hyperactivity disorder (ADHD), which is one of the most common neuropsychiatric disorders for children. The current ADHD diagnosis is mainly based on a subjective evaluation that is easily biased by the experience of the clinicians and lacks the support of objective indicators. The purpose of this study is to propose a method that can effectively identify children with ADHD.

Approach. In this study, we proposed a CNN-LSTM model to solve the three-class problems of classifying ADHD, attention deficit disorder (ADD) and healthy children, based on a public electroencephalogram (EEG) dataset that includes event-related potential (ERP) EEG signals of 144 children. The convolution visualization and saliency map methods were used to observe the features automatically extracted by the proposed model, which could intuitively explain how the model distinguished different groups.

Main results. The results showed that our CNN-LSTM model could achieve an accuracy as high as 98.23% in a five-fold cross-validation method, which was significantly better than the current state-of-the-art CNN models. The features extracted by the proposed model were mainly located in the frontal and central areas, with significant differences in the time period mappings among the three different groups. The P300 and contingent negative variation (CNV) in the frontal lobe had the largest decrease in the healthy control (HC) group, and the ADD group had the smallest decrease. In the central area, only the HC group had a significant negative oscillation of CNV waves.

Significance. The results of this study suggest that the CNN-LSTM model can effectively identify children with ADHD and its subtypes. The visualized features automatically extracted by this model could better explain the differences in the ERP response among different groups, which is more convincing than previous studies, and it could be used as more reliable neural biomarkers to help with more accurate diagnosis in the clinics.

1. Introduction

Attention is the cognitive process of selecting and focusing on related stimuli, and it is a very important cognitive ability in people’s daily life. According to the clinical practice guideline of the American Academy of Pediatrics, attention-deficit/hyperactivity disorder (ADHD) is the most common neurobehavioral disorder of childhood and can profoundly affect the academic achievement, well-being, and social
interactions of children [1]. Attention deficit disorder (ADD) is a subtype of ADHD, which is characterized by only inattention, not involving impulsivity and hyperactivity. The existing study indicates that ADHD is a highly complex and heterogeneous disorder in terms of its multi-factorial etiological risk factors, diverse neurocognitive impairments, and comorbid problems [2]. The underlying causes in terms of neurophysiology are related to changes in the activity of the frontal cortex [3], dopaminergic processing [4] and decreased functional brain network connectivity [5]. The most widely used ADHD diagnostic standards in the world are the Diagnostic and Statistical Manual-5 and the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). These methods are based on the subjective evaluation of behavioral psychology and clinicians which lack the support of objective measurement indicators.

Traditional machine learning methods have been used to evaluate the usefulness of functional imaging data such as electroencephalogram (EEG) [6]. It typically needs to perform feature extraction to obtain effective features and build prediction or clustering models to find neural markers [7]. Tenev [8] used SVM method combining data from multiple experimental conditions to classify ADHD and the results lacked further comparison and verification. Altnukaynak [9] used k-nearest neighbors, SVM, Random Forest, logistic regression, and other machine learning methods to classify ADHD, whereas the feature data of multiple dimensions needed complicated calculations. When the data of more electrodes and more subjects were involved, the calculation process became more complicated, and the feature representation was also difficult to explain. Although this method is helpful for the diagnosis of ADHD, their processing process is very different, and the research results lack a certain connection. The deep learning method can calculate all the information of the dataset and automatically extract the features, without the need for complex feature screening processes. There have been some studies using deep learning methods on ADHD classification and achieved good accuracy [10–14]. However, these studies used the same publicly available dataset ADHD-200 which is MRI data, while other EEG studies used data collected separately and lack an open dataset as a [15–17]. Due to characteristics such as real-time, ease of use, non-invasiveness, and low cost, EEG has more advantages than MRI in the field of clinical diagnosis.

In this study, we selected an existing open-source dataset of EEG research as the research object after careful investigation [18]. It was used to solve a three-class problem and the results of the original study were not satisfactory. The classification accuracy of healthy children and ADHD subtypes is only 83% after removing the data with low model accuracy. In our study, we proposed a CNN-LSTM model based on the existing CNN model to test this dataset and decode the brain activities of different groups. Chen proposed a combined CNN-LSTM model to automatically identify six types of ECG signals [19]. Xu et al proposed a 1D CNN-LSTM model for automatic recognition of epileptic seizures and achieved 99.39% and 82.00% on the binary and five-class task respectively [20]. Xu et al also proposed a network that combined LSTM units with CNN to explore temporal features in face anti-spoofing tasks [21]. Tasdelen proposed a CNN-LSTM network to solve the classification problem of microRNA, and its processing data was a complex gene sequence [22]. Compared with similar studies, our research tried to employ and modify the popular CNN-LSTM model specially for the ADHD recognition application, so as to understand the link between the attention deficits in children and their electrical brain activities.

In general, using the deep learning method to classify ADHD and the healthy group is a new way to understand the mechanism of attention. The previous study tended to choose some features from a subjective perspective and then analyze the changes between them, which was always not comprehensive enough. Human cognitive activities and the corresponding EEG information are very complicated. A more reasonable research idea should be to first learn all the information in the way of artificial intelligence, draw certain conclusions, then analyze the characteristics of the learning process, and finally verify it through more experiments and data.

Although deep learning methods can effectively classify different groups of children, the principle of its classification is still unknown. Since all network layers only have input and output, its learning process is still a black box for us, which is not conducive to subsequent cognitive function research. The existing research had proposed some methods to explain the decision of deep learning nonlinear classifiers [23–25]. These methods are usually based on the calculation of network gradient data. Considering the complex calculation differences involved in various methods and their unknown effectiveness, we used a saliency map [26] and focused on the visualized comparison of brain electrical activities.

It should be noted that the visualization of the network model is very important because it can not only help us understand the intermediate learning process of the network but also in-depth explore the cognitive function information contained in the automatically extracted features. The content of this study is to carry out sufficient visualized analysis based on the effective identification of ADD and ADHD children, which can find potential neural biomarkers and provide a basis for clinical diagnosis.
2. Materials and methods

2.1. Data description
The EEG data used in this article comes from the Department of Child and Adolescent Psychiatry, the Technical University of Dresden, and are shared by professor Beste on https://osf.io/6594x/. According to the previous related paper [18], this data was approved by the local ethics committee, and the informed written assent/consent was obtained from all participants. The subjects included 144 children (44 healthy controls (HCs), 52 ADD, 48 ADHD), of which 100 children with AD(H)D had undergone professional clinical diagnosis such as ICD-10. Hyperactivity (\(p < 0.001\)) and impulsivity (\(p < 0.001\)) were significantly more pronounced in patients with ADHD than in those with ADD. There were no differences in age, IQ, or gender distribution among the groups (all \(F < 1.5\), all \(p > 0.3\)). And more information about this dataset can be found in his paper.

The EEG data input by the neural network is generally categorized into 3 classes: time-domain waveform data, frequency-domain oscillatory data and connection matrix of the functional brain network. This dataset was collected in a time estimation task, which belongs to event-related potential (ERP) data [27]. During the task, participants were asked to press a button when they thought 1200 milliseconds had passed after the visual stimulus (a white square on a black background). After the data undergoes 0.5–20 Hz filtering, independent component analysis to remove eye artifact, myoelectricity and other preprocessing methods, it will be input into the neural network model as a time-domain waveform type. ERP data has the characteristics of low-frequency, time-locked external stimulus, specific clear waveforms, etc., and is considered to contain rich cognitive function information. A brief description of preprocessed ERP data is given below. 144 participants contain 33 902 trials (10 129 trials for HC group, 13 031 trials for ADD group, 10 742 trials for ADHD group), each with 56 EEG channels and 385 time points. The data were downsampled to 256 Hz and the time length of each trial is 1.5 s. All data were used for a three-class classification. In general, the ERP data input to the neural network model was only preprocessed such as denoising and was still a time-domain waveform without calculating any features.

2.2. Classification methods
2.2.1. CNN-LSTM model
In this study, a model in which convolution neural network (CNN) and long short-term memory (LSTM) network were connected in series was designed to learn EEG architecture. This CNN-LSTM model was improved from the previous CNN model that processed a large amount of EEG datasets. Therefore, the CNN-LSTM model can be applied to the recognition and classification of multiple different EEG signals which represent different groups. Moreover, the CNN-LSTM model can produce interpretable features that have similar patterns or specific areas of interest relative to the original signal. The visualization and full description of the CNN-LSTM model were shown in figure 1, respectively. The input data of the model was channels and time points, which are 56 and 385 in this dataset. In the training stage, the input batch size was 256, the learning rate was 0.001, and training iterations were 300 epochs. The Adam optimizer and categorical cross-entropy loss function were used in the model compile. The early stop and learning rate reduction methods were also used to optimize the training process. In the above two callback methods, the loss parameters of the validation dataset were supervised. If it did not continue to decrease in 5 epochs, the learning rate would be multiplied by 0.5. And the training terminated if the loss was no longer decreasing in 20 epochs. All relevant codes of the CNN-LSTM model were running on 4 NVIDIA TITAN Xp GPU, with CUDA 11 and cuDNN v8.1, in Tensorflow [28], using Keras API [29]. The related code used in this research was attached to Github (https://github.com/zhemuzzz/EEG-classification-code).

The CNN-LSTM model architecture is divided into three blocks. In block 1, two convolutional steps were applied to compress the spatial and temporal characteristics of EEG. The convolutional part is similar to the EEGNet model [30] and some parameters and structures are adjusted appropriately according to the characteristics of this dataset. First, the 2D convolutional filters of size (1, 50) were used to extract the temporal features containing 200 ms window which output feature maps that capture frequency information at 5 Hz (sample rate is 256 Hz). Then a depthwise convolution [31] of size (C, 1) was used to learn spatial information. Depthwise convolution is a type of convolution where we apply a single spatial filter for each input feature map so that it can keep each channel separate and reduce the number of trainable parameters. We set the depth parameter \(D = 2\), so each temporal kernel produces two spatial filters (as shown in figure 1). After passing each layer of convolution, the batch normalization layer was used to make convolutional neural networks faster and more stable [32]. The exponential linear unit (ELU) was used as the activation layer to solve the problem of vanishing gradients and exploding gradients [33]. After that, average pooling operation with pool size (1, 40) and stride size (1, 20) was used to downsample the convolution output along its temporal dimensions. At the end of block 1, the dropout layer was used to prevent training overfitting [34].

In block 2, two LSTM layers formed the main part of the recurrent neural network architecture. The outputs of the convolution part in block 1 will be reshaped and input to the LSTM layer. The input to the next LSTM layer is the output time sequence of the
Figure 1. Overall visualization of the CNN-LSTM model architecture. Lines represent the connections between network layers, and shape graphics represent the input and output of the layers. The network starts with a temporal convolution (Conv2d part) to learn time-domain filters, then uses a depthwise convolution (DepthwiseConv2d part), connected to each feature map of Conv2d outputs individually, to learn spatial filters. In the recurrent neural network part, two LSTM layers will process the sequence of a temporal summary feature map, which learn how to classify the abstract features of time series.

Each LSTM layer has U units and a common unit is composed of a cell, an input gate, an output gate and a forget gate \[35\]. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. An LSTM unit includes hidden states and cell states, and the input of the first LSTM layer is a time series of hidden states. After updating the states of the three gates in turn, the cell state and hidden state of the next time step can be obtained. LSTM networks are well-suited to classifying time series data since there can be lags of important temporal information between events in EEG signals.

In block 3, one dense layer constitutes the final classification part, and the final classification results will be output through the Softmax activation function.

The CNN-LSTM model proposed in this paper selects the most suitable convolution part for EEG features after learning from existing CNN models and it reduces the depth of convolutional layers. Additionally, visualization operations are performed throughout the whole training cycle of the model to extract the time-frequency features required for traditional EEG analysis. Compared with other CNN-LSTM models, this model has more advantages in EEG data processing, visual analysis of EEG time-domain waveforms and spectral energy changes.

2.2.2. Comparison with existing CNN approaches

This research will compare the classification performance of the proposed CNN-LSTM model with the existing CNN models. ShallowConvNet and DeepConvNet models were proposed to decode imagined or executed tasks from raw EEG [36]. They had achieved 84.0% decoding accuracy on brain-computer interface (BCI) competition IV 2a and 2b datasets. EEGNet model [30] has been compared with these two models on multiple datasets such as P300 ERP, feedback error-related negativity, movement-related cortical potential, and sensory motor rhythm. In short, these three network models have shown good classification results on multiple EEG datasets. It is necessary to compare the CNN-LSTM proposed in this article with them in various aspects.

The EEGNet architecture consists of three convolution layers with a dense layer for classification and the previous Conv2D layer and DepthwiseConv2D layer are consistent with the block 1 part of CNN-LSTM. In the original parameter setting for 128 Hz signal, the convolution size of the Conv2D layer is (1, 64), the depth parameter D is 1 and the pool size of averagePool2D layer is (1, 4). EEGNet then uses a separable convolution to learn how to summarize individual feature maps in time (the depthwise convolution) with how to optimally combine the feature maps (the pointwise convolution).

The ShallowConvNet architecture consists of a temporal convolution (25 time points), a spatial convolution (44 channels), a squaring nonlinearity \(f(x) = x^2\), a mean pooling layer and a log nonlinearity \(f(x) = \log(x)\). It was designed to learn the temporal structure of the band power changes within the trial. The DeepConvNet architecture consists of four convolution-max-pooling blocks and its first block has a temporal convolution and spatial filters that could better handle a large number of input channels.

As shown in the table 2 architecture section, the number of filters (convolution kernel) for the four models was 100, 400, 40 and 150 in order.
DeepConvNet is a variant of ShallowConvNet that increases the depth of the network. The first two convolutional layers of CNN-LSTM learn the architecture of EEGNET, which does not have a third layer of convolution but adds a small number of LSTM units (20 units). The number of training parameters of the final four network models is 23883, 179528, 2080, 13713 in sequence.

In general, the conventional CNN directly compresses the local spatiotemporal features obtained by shallow convolution and extracted more abstract features in the depth layer. Signal data related to EEG involves the influence of the number of channels and event-related time periods, and the compression of its spatiotemporal features does not require the use of excessively deep networks. In neuroscience, the rhythmic activity behind EEG signals is related to different frequency bands and brain regions, and the waveform of the previous time period may have a propagation relationship with the subsequent waveform. The network structure of CNN is limited to extracting local features in the data matrix, while LSTM can perform associated learning on the entire task-related time period data. For example, the EEG signal data is usually 32 or 64 channels, the time of one trial is usually about 2 s, and the time range of the ERP related waveform is about 100 ms to 300 ms. That is to say, the model needs to learn features with a certain length of time and spatial connection, and too many convolutions will extract too short and meaningless features. The more graph convolutional layers are used, the local features extracted by their nodes tend to converge to the same vector so that the information of the signal is over-squeezed into regions of fixed size. Therefore, excessive convolution of EEG data will lead to the depression of network learning ability. Shallow convolution has been proved to be effective for EEG spatiotemporal feature extraction, and the serial connection of LSTM can more appropriately learn the underlying meaningful EEG activity features associated with multiple ERP waveforms.

2.2.3. Data analysis

In this study, five-fold cross-validation and leave-one-out-cross-validation (LOOCV) were used to measure model performance. In five-fold cross-validation, 6000 trials were randomly selected from 33902 trials as the test set in each fold. Then the training set and the validation set were selected at a ratio of 4:1 in the remaining trials. In the LOOCV method, each subject was used as the test set, and 20 other subjects were randomly selected as the validation set, and the rest of the data was used for training. Both verification methods are cross-subject, while five-fold cross-validation is based on trials, and LOOCV is based on subjects.

In addition to accuracy indicators, there are also sensitivity, specificity, AUC, and F1 scores. Sensitivity is the probability of testing positive in the gold standard judged by the disease (positive) population. Specificity is the probability of testing negative in people whose real label is negative. AUC is the area under the receiver operating curve which means the probability that the score of the positive sample is greater than the score of the negative sample in the case of randomly selecting a single sample of different labels. F1 score is the harmonic average of accuracy and recall rate, which can comprehensively indicate model classification performance. In addition, a one-way analysis of variance (ANOVA) method will be used for the brain electrical activity of different groups of children.

2.2.4. Neural network feature visualization

Although deep learning methods can effectively classify children with different attention functions, the learning process and the basis of classification judgment are still a black box. Therefore, the existing research usually looks for methods that can explain the classification performance by automatically extracting features, and these features have certain physiological significance. Here presented two different approaches explaining the classification process of CNN-LSTM.

The first method is to visualize the output of each convolution kernel. This method focuses on the process of how the model learns and understands this dataset. Since the input is ERP data, it has rich time-domain information and 56 channels of spatial information. The output of the convolution kernel corresponds to a temporal or spatial filtering process. In other words, after getting the output of each convolutional layer, and drawing it into a time-domain waveform or spectrogram, we can clearly understand how the network model extracts features.

The second method is to depict a saliency map related to the time-frequency characteristics of EEG. This method aims to identify which input features contribute the most to the classification process and make the classifier decide final outputs [26]. It was calculated by taking the gradient data of the classification score with respect to the input data. Actually, the gradient data comes from the dense layer outputs before the softmax-activation function. That is, we reversely deduce its gradient data from the processed data of all network layers of the entire CNN-LSTM. After obtaining the gradient data, we will draw the corresponding time-frequency feature map and compare it with the original signal. The above two methods are to extract features that have a significant impact on classification, rather than other brain electrical activities, motion artifacts and other noises. In this study, the creation of feature visualization involves using tools such as EEGLAB [37], fieldtrip [38].
3. Results

3.1. Cross-subject classification

In this study, all the data classification is based on cross-subjects and the difference lies in whether it is based on a single subject or a single trial. In the LOOCV method, we tested each subject individually and ran a total of 144 folds in 4 models. In figure 2, The classification results showed large differences in individual brain electrical activity. The average accuracy of the CNN-LSTM model was 39.59%, and the standard deviation was 28.11%. Among the 144 test subjects, some had an accuracy rate of over 90%, while others did not more than 10%. Large individual differences were observed for the EEG classification results. The test results of the other three models were 33.9% (ShallowConvNet, std = 30.4%), 37.78% (DeepConvNet, std = 25.3%), 37.79% (EEGNet, std = 32.7%) respectively. When compared with the three models, the classification performance of the proposed CNN-LSTM model was slightly higher. In the previous study [18], the three-classification accuracy obtained on this dataset using the EEGNet model reached only a chance level of 33%.

In five-fold classification method, the model parameters and architecture were the same as the LOOCV method. The loss changes during training epochs were shown in figure 3. It could be seen that the CNN-LSTM model was relatively stable, and the final loss of its training was relatively smaller. There was much smaller between the loss curve of the training process and the loss curve of the validation process. It indicated that models that converge quickly and steadily had the ability to capture useful features for classification.

The confusion matrix of four models in five-fold cross-validation method was shown in figure 2. It showed that only the CNN-LSTM model exceeded 98% in these three groups (HC, 98.22%; ADD, 98.15%; ADHD, 98.33%). In the ShallowConvNet model, the classification accuracy of the three groups was lower than 96%, and the accuracy of the other two models was much lower than that of CNN-LSTM. In figure 4, we could find that the mean accuracy of CNN-LSTM was 98.23% which was significantly higher than the other CNN models. Furthermore, the CNN-LSTM model has the lowest standard deviation (3.3%) of the classification accuracy.

In table 1, the detailed architecture and classification performance indicator showed that the CNN-LSTM model was better than others. The CNN-LSTM model reached 98.23% of accuracy, 99.11% of specificity, 98.24% of sensitivity, 99.64% of AUC score and 98.22% of F1 score. When ShallowConvNet had the same kernels as CNN-LSTM, the network parameters had reached 23833, and its accuracy was 95.49%, which was lower than CNN-LSTM. The accuracy of DeepConvNet was only 86.06%, even though it had 400 kernels and a large number
of network parameters (179528). Therefore, blindly
increasing the number of convolution kernels and
network parameters could not achieve the ideal clas-
sification results. Given that EEGNet only had 40 ker-
nels, CNN-LSTM appropriately increases the number
of convolution kernels according to the characterist-
cs of this dataset. After adding 20 LSTM units and
modifying the deep convolution layer into a recurrent
neural network, the accuracy was further improved in
this dataset.

Based on the design of the CNN-LSTM model
of this study, we have removed or added convo-
lution layers to compare the performance of our
model with 3 different cases: (1) LSTM only; (2)
less Convolutions + LSTM; (3) excessive Convo-
lutions + LSTM, as shown in table 2. As can be
seen from the table, the accuracy of the model only
reached 45.76% when the convolutional part of the
network was removed, leaving only two layers of
LSTM (Case 1). Less CNN-LSTM was a model that
removed one layer of time series convolution, and its
accuracy was 69.25% (Case 2). Deep CNN-LSTM was
a model that increased to four layers of convolution
and then connect them to the LSTM layer, resulting
in an accuracy rate of 83.46% (Case 3). By comparing
the results of all the four different cases, it was found
our proposed CNN-LSTM model achieved the best
performance with an accuracy as high as 98.23%.

3.2. CNN-LSTM feature visualization
3.2.1. Visualizing the convolutional kernel outputs
In this study, we visualized the feature extraction pro-
cess of each convolution kernel based on the output
of the middle layer of the CNN-LSTM network. As
previously introduced, the first two layers of convo-
lution of CNN-LSTM were similar to EEGNet, which
consisted of a layer of temporal convolution and spa-
tial depth convolution. Then the LSTM part further
learned the convolutionally compressed features. The
data points involved in this process were the hidden
states of multiple time steps data that would be intro-
duced by gradient data in the saliency map.

Since the input was 56 channels, 385 time
points of ERP waveform data and the first layer

Figure 3. The loss changes during training epochs of four models in 5-fold cross-validation. The blue lines indicated the loss of the training dataset, and the orange lines indicated the loss of the validation dataset. The horizontal axis represented the epochs, and the vertical axis represented the size of the loss parameter.
of convolution performed edge padding during the extraction process, we could get waveforms of the same length. There were 50 filters with the length of 50 time points, and it meant that it could capture 5 Hz frequency information. So, we processed the waveforms of 56 channels by time-frequency method, and each filter (convolution kernel) output had a corresponding spectrogram with 1.5 s length and 1–20 Hz range. As shown in figure 5, the time-frequency energy map showed that the first convolutional layer could extract multiple time periods or fixed frequency range signals. We could observe that most filters extracted the energy of two periods which focused on 200–300 ms and 1200–1300 ms. In addition, some filters extracted the energy of a fixed frequency range of the entire 1.5 s time window, some were in the 5 Hz range, and some were in the 10 Hz range. It should be noted that the data displayed here is the signal of

![Figure 4. Confusion matrix of four models in 5-fold cross-validation. The above numbers indicate the accuracy rate, and the following numbers indicate the corresponding trials. The horizontal axis represents the predicted label, and the vertical axis represents the ground truth label.](image-url)

Table 1. Architecture and performance comparison of CNN models and CNN-LSTM model in 5-fold cross-validation.

| Architecture     | ShallowConvNet | DeepConvNet | EEG-NET | CNN-LSTM |
|------------------|----------------|-------------|---------|----------|
| Filters          | 100            | 400         | 40      | 150      |
| Units            | /              | /           | /       | /        |
| Parameters       | 23883          | 179258      | 2080    | 13713    |
| Accuracy         | 95.49%         | 86.06%      | 92.63%  | 98.23%   |
| Specificity      | 97.72%         | 93.27%      | 92.70%  | 99.11%   |
| Sensitivity      | 95.46%         | 86.94%      | 96.32%  | 98.24%   |
| AUC              | 99.49%         | 97.90%      | 98.95%  | 99.64%   |
| F1 score         | 95.48%         | 86.09%      | 92.56%  | 98.22%   |

Table 2. Performance comparison of ablation study in 5-fold cross-validation.

| Parameters   | CNN layers | LSTM layers | Accuracy |
|--------------|------------|-------------|----------|
| LSTM only    | 16713      | 0           | 2        | 45.76%   |
| LessCNN-LSTM | 19613      | 1           | 2        | 69.25%   |
| DeepCNN-LSTM | 247880     | 4           | 2        | 83.46%   |
| CNN-LSTM (this study) | 13713 | 2 | 2 | 98.23% |
Figure 5. Visualization of the first temporal convolutional layer outputs for subject 1 of the HC group. There are 50 time-frequency EEG graphs (from 50 temporal filters), and each subgraph learned the temporal features for a 0.3 s window (50 time points) from the 1.5 s time-length data. The title of each subgraph indicates which filter it belongs to.

Figure 6. Visualization of the second spatial convolutional layer outputs for subject 1 of the HC group. There were 100 time-domain EEG waves (from 2 depthwise kernels), and each subgraph learned the spatial features from 56 channels of 50 temporal convolutions.

In the second layer of convolution, we set the depth parameter to 2 and compressed all channels, so 100 single-channel waveform outputs were obtained. In figure 6, we could observe that quite a few filters extracted low-frequency signals, and a small part of filters extracted signals in the alpha band and beta band. In fact, the low-frequency signals had the characteristics of ERP such as positive or negative waveforms in a specific time period. And the frequency information learned by spatial filters was corresponding to the energy map in the previous convolution layer.

3.2.2. Comparison of gradient data and the original signal

The intermediate learning process of feature extraction had been visualized before, but the final feature extracted by the network model and their connections with the input signals needed to be further explained. Here, the saliency map, which was also called the relevance map, was drawn by calculating the gradient data of each subject. By comparing the original signals and the saliency map's
ERP waves, time-frequency characteristics, and brain function network connections, we got different brain electrical activities of this dataset in detail and further understood how the model judged the possible group distribution.

When the data of all subjects had been drawn, a small part of the subjects in each group had different brain electrical activities due to individual differences (almost 10% ratio), so we selected five representative subjects for each group. The selection principle was that most of the other subjects in the same group had similar brain electrical activities. As shown in figure 7(A), subjects in the same group had a small difference in the original signal ($p < 0.05$), and the brain electrical activities of average signals in three groups were concentrated in the time period before 500 ms. It could be found that the graph after averaging all the trials of subjects from the same group could not visually observe the difference in the time-frequency feature of the three groups. As for gradient data in figure 7(B), subjects corresponding to Part A showed that EEG activities of HC and ADD group focused on time period before 500 ms, while the ADHD group also paid attention to the time period after 1200 ms in addition to this similar activity ($p < 0.05$). It could be found that the original signal energy of a single subject was generally distributed in the low frequency band of the entire time period, and the gradient data obtained through the network model tended to focus on a certain time period. The first time period in the overlap maps showed that the energy of the HC group was stronger than ADD and ADHD groups.

The previously extracted saliency map of gradient data mainly observed the specific time period between different groups. Next, the spatial topographic map was used to observe which channels the network model focused on. After knowing which part of the brain area and which time period the model had extracted, we then looked for the corresponding activity in the original signals to further observe the potential differences between different groups of children in this dataset.

**Figure 7.** Time-frequency EEG map of original signals (A) and gradient data (B). The time length is 1.5 s and the frequency range is 0–20 Hz. All data were averaged from 56 channels and divided into three groups (HC, ADD, ADHD). The title of each subgraph on the left indicates the number of the subject. On the right is the average spectrogram of the original signal (A) and the overlapped maps of gradient data (B). For example, HC-overlap was superimposed from 44 maps of each subject in the HC group.
In figure 8(A), EEG activities of three groups were all distributed on the opposite sides of the parietal lobe, with slightly different intensities. The results showed that the topographic maps between individual subjects were different, even if they belonged to the same group, with some subjects having parietal lobe activity on the left side, some on the right side, and some on both sides ($p < 0.05$). In figure 8(B), EEG activities of gradient data focused on the frontal lobe and there were also some activities near the central area. The distribution of the gradient data of each group on the spatial domain was similar.

In figure 9, it was observed that the three groups of ERP waveforms had commonalities and differences by selecting the electrodes corresponding to the active brain regions. Here, some waveforms were identified according to the time frame. For example, P2 represents a positive wave of about 200 ms and CNV (contingent negative variation) represented continuously oscillating negative waves. In the Fz electrode, the waveforms that appeared were N1, P2, P3, and CNV in sequence. In the relevant frontal area, the waveform in the HC group had the largest drop change, followed by the ADHD group, and the ADD group had the smallest waveform change ($p < 0.05$). Among the three groups, the differences between the P3 wave around 300–400 ms and the CNV wave before 1200 ms were the most obvious. In the Cz electrode, only the HC group had a steady increase in CNV amplitude while the ADD and ADHD groups had relatively small amplitude change ($p < 0.05$). Besides, the P2 wave of the HC group was also the highest, while the ADD group was the lowest. The choice of FZ and Cz was derived from the visualization method of the saliency map. The P7 electrode was located on the left side of the parietal lobe, corresponding to the original signal topographic map activity. Moreover, the activities of P8 on the right were similar to those of P7. Here we marked P2, N2, P3 and CNV waves in order, and found no significant difference in brain electrical activity of the three groups ($F = 2.03, p = 0.1319$). The ERP waveform before 300 ms was related to early visual perception, such as N1, P2, and their differences were not obvious. Waves in the range of 300 ms–1200 ms like P3 wave and CNV wave had a certain difference.

4. Discussion

4.1. The classification results on three-category dataset

In the five-fold cross validation method, the CNN-LSTM model reached 98.23% accuracy which was the highest, and only the CNN-LSTM model exceeded 98% in all three groups. DeepConvNet had the deepest network structure and the most training parameters, but the accuracy was the lowest. This meant that increasing the number of convolution kernels and the depth of the convolution layer could not fully learn effective EEG features. After the part of the temporal filter and spatial filter, the compressed feature was single-channel time points, which was extracted from a specific frequency range. Compared with the CNN network commonly used in two-dimensional image processing applications, the LSTM network had more advantages for learning.
with these data. In general, the most critical part of learning EEG signals is how to extract the information between its time series and multiple channels. EEG information is a long sequence and there are lots of connections between multiple channels. This connection may be short or long distance, and the CNN network may not be suitable for extracting this information\(^3\). From the first part of the model visualization in this paper, it can be found that the features extracted by the CNN model cover activities in multiple frequency bands and complex ERP waveforms. However, this did not help the network to find the final key activities. Therefore, the CNN-LSTM model proposed in this paper can obtain a series of feature maps highly correlated with ERP activity after shallow convolution of complex EEG data. These feature maps will continue to learn through the LSTM layer, which not only avoids excessive convolutional compression, but also learns the relationship between various ERP activities on the timeline, and even learns the potential relationship between every single channel and multiple frequency bands.

When training based on trials, the currently popular models for learning EEG data have achieved good results, and when training based on subjects in the LOOCV method, these models are easy to overfit during the training process, with some subjects having a high accuracy rate, and some close to zero. The poor performance of the subject-based classification is due to the management of the dataset. When the dataset is divided based on subjects, the cardinality of its data distribution becomes only 144 (the total subject number), which is very small compared to the input of the deep learning model. Meanwhile, the data from one subject can be in either the training dataset or the test dataset. As a result, the data for model training is too small, and the individual EEG data for testing is not included in the training set. Given that EEG activities and ERP response of the experimental task show dramatical differences, the model trained by one group of subjects might not be suitable to test another group of subjects, since the EEG features of the two groups show large individual differences. In future research,

---

**Figure 9.** ERP waveforms of original signals at Fz, Cz, P7 electrodes. The vertical axis was the normalized voltage range of the original signals, and the horizontal axis was a time period of 1.5 s, of which 1.2 s was the expected button time response set by the ERP task. The legend showed different lines representing different groups (HC, green; ADD, blue; ADHD, red).
datasets of much larger subject numbers could be used to prevent the problem of overfitting caused by too small cardinality when the dataset is based on subject distribution. Meanwhile, an improved deep learning model that can extract more spatial feature information from multi-channel EEG signals could be employed to identify commonalities and differences between different individuals.

4.2. The CNN-LSTM feature visualization

In the first temporal convolution layer, the output of most convolution kernels was very similar to the time-frequency diagram of the original signal, and its energy was concentrated in 200–300 ms and 1200–1300 ms. These two periods were related to the ERP task of this dataset in that participants were asked to press a button when they thought 1200 milliseconds had passed after the visual stimulus [18]. Therefore, the first period of activity corresponded to the early visual cognitive response, and the latter period of activity corresponding to the process of preparing for movement. Besides, some convolution kernels outputted energy in a fixed frequency range, indicating that the convolutional network could extract frequency-domain features. In the second spatial convolution layer, most of the output low-frequency time-domain waveforms had ERP characteristics. In short, the convolution layer could extract time-domain information with specific waveform and frequency-domain information containing different frequency bands.

As mentioned above, we calculated the gradient data from the output to the input and combined the time-frequency decomposition method to draw the corresponding brain electrical activity map. Relative to the time-frequency diagram of the original signal of each subject, we could find that the model focused on the signal changes in a fixed time period. Then we further found through the EEG topographic map that the channels the model focused on were distributed in the frontal area and central area. In the traditional EEG signal analysis method, we analyzed the cognitive activities of subjects in different groups after obtaining the differences in time-frequency characteristics and tried to find patterns. The results now showed that the activities model focused on were inconsistent with the prominent activities depicted by the original method. In fact, the brain electrical activity collected during the execution of the specified task contained all the cognitive responses at the time, while only part of the cognitive activities determined the differences between different individuals, and other activities should have the same meaning. Simply drawing the energy distribution of the entire time series only got the strongest changes and the most active brain area. These activities might be common to all groups. Therefore, the deep learning method was introduced to solve this problem. It could find potential differences and neural connections without more cumbersome analysis.

After locking the frontal area, central area and two time periods, we only needed to draw the corresponding original signals to analyze the ERP characteristics of different groups. At the Fz electrode, there were indeed differences in the P3 wave and CNV wave. The P3 wave was also called P300 which was thought to index the operation of attention and memory processes engaged during stimulus processing [40]. P300 had differences in amplitude and latency in different groups of people in a variety of experimental tasks. Therefore, it was used as the detection of cognitive impairment diseases [41, 42] and to interpret the underlying neural mechanism of BCI in response to changes in events [43]. Indeed, many studies believed that P300 was related to visual attention and the amplitude was proportional to the number of attentional resources that were available for stimulus processing [44, 45]. It could be found that the HC group had the largest P300 amplitude, and the ADHD group was similar to the ADD group with smaller amplitudes. Although the positive value of the ADD group was the largest, we should pay attention to the magnitude of changes in EEG signals. The previous N1 and P2 parts were related to early visual response, and there was no significant difference [46]. The CNV wave of ADD group had the smallest decrease. The decrease in the CNV wave of the ADHD group and HC group was more obvious than that of ADD group. At the Cz electrode, the most obvious difference was also the CNV wave, only the HC group had a significantly negative oscillation wave. A previous study showed that these activities did exist at the Cz electrode in this dataset [27]. One related study had proposed an activation-executive model to help diagnose ADHD, where the activation cortexes were the frontal area and central area [3]. And the CNV wave was considered to reflect a common core preparatory process related to brain system optimization [47]. The earliest discovered active readiness potential for movement response was similar to this CNV activity [48]. In general, we could clearly distinguish three groups of children with different cognitive functions from the difference between CNV wave and P300 wave of frontal area and central area.

In this study, we specifically proposed the CNN-LSTM model for both performance improvement and automatic EEG feature extraction for the ADHD application. Figures 6 and 7 in the original manuscript were examples of visualizing the EEG features automatically extracted by the model. The EEG features extracted by the proposed model may be different from the patterns observed from individual EEG raw signals, but they played a more important role in separating different groups. It should be noted that the feature visualization was very useful because it
could not only help us understand the intermediate learning process of the network but also explore the cognitive-related information contained in the automatically extracted features. In this way, the proposed CNN-LSTM no longer acted as a black box for ADHD classification, and it instead became a powerful tool to explore the possible underlying inherent relation between the EEG features and the attention defects.

5. Conclusions

This paper proposed a CNN-LSTM model for the three-category classification of the ADHD children dataset. The results showed that the proposed model could achieve a much higher accuracy rate of 98.23% when compared with existing models of similar studies. Another advantage of the proposed model is that it could automatically extract and illustrate the key features that are important for accurate classification so that the classification model no longer remained a black box in similar deep learning studies. Among the extracted features, it is observed that the P300 and CNV in the frontal lobe had the largest decrease in the HC group, whereas the ADD group had the smallest decrease. In the central area, only the HC group had a significant negative oscillation of CNV waves. Through deep learning and visualization methods, combined with traditional time-frequency feature analysis, the proposed model can automatically extract the key features and use them to classify different ERP activities from different ADHD groups. The proposed CNN-LSTM model could also be useful for the extraction of important biomarkers and the accurate classification of other evoked potentials.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://osf.io/6594x/.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (#81927804, #62101538), Shenzhen Governmental Basic Research Grant (#JCYJ20180507182241622), Science and Technology Planning Project of Shenzhen (#JSGG20210713091808027, #JSGG20211029095801002), China Postdoctoral Science Foundation (2022M710968) SIAT Innovation Program for Excellent Young Researchers (E1G027), and CAS President’s International Fellowship Initiative Project (2022VEA0012).

Conflict of interest

None.

References

[1] Subcommittee on Attention-Deficit/Hyperactivity Disorder, Steering Committee on Quality Improvement and Management 2011 ADHD: clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents Pediatrics 128 1007–22
[2] Lao Y, Weibman D, Halperin J M and Li X 2019 A review of heterogeneity in attention deficit/hyperactivity disorder (ADHD) Front. Hum. Neurosci. 13 42
[3] Rodríguez C, Gonzalez-Castro P, Cuell M, Areces D and Gonzalez-Pienda J A 2016 Attention deficit/hyperactivity disorder (ADHD) diagnosis: an activation-executive model Front. Psychol. 7 1406
[4] Cubillo A, Halari R, Smith A, Taylor E and Rubia K 2012 A review of fronto-striatal and fronto-cortical brain abnormalities in children and adults with attention deficit hyperactivity disorder (ADHD) and new evidence for dysfunction in adults with ADHD during motivation and attention Cortex 48 194–215
[5] Cao M, Wang Z and He Y 2015 Connectomics in psychiatric research: advances and applications Neuropsychiatr Dis. Treat. 11 2801–10
[6] Wolters T, Buitema J K, Beckmann C F, Franke B and Marquand A F 2015 From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics Neurosci. Biobehav. Rev. 57 328–49
[7] Miotto R, Wang F, Wang S, Jiange X and Dudley JT 2018 Deep learning for healthcare: review, opportunities and challenges Brief. Bioinform. 19 1236–46
[8] Tenev A, Markovska-Simosa S, Kocarev L, Pop-Jordanov J, Müller A and Candrian G 2014 Machine learning approach for classification of ADHD adults Int. J. Psychophysiol. 93 162–6
[9] Atkmutnayak M, Dolo N, Güven A, Pektası F, Ozmen S, Demirci E and İzzetoglu M 2020 Diagnosis of attention deficit hyperactivity disorder with combined time and frequency features Biocybern. Biomed. Eng. 40 927–37
[10] Quaak M, van de Mortel L, Thomas R M and van Wilgen G 2021 Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis Neuroimage Clin. 30 102384
[11] Ruiz A, Asad M, Alonso E and Slabbaugh G 2020 DeepFMRI: end-to-end deep learning for functional connectivity and classification of ADHD using fMRI J. Neurosci. Methods 335 108506
[12] Khullar V, Salgotra K, Singh H P and Sharma D P 2021 Deep learning-based binary classification of ADHD using resting state MR images Augment. Hum. Res. 6 5
[13] Mao Z, Su Y, Xu G, Wang X, Huang Y, Yue W, Sun L and Xiong N 2019 Spatio-temporal deep learning method for ADHD fMRI classification Inf. Sci. 499 1–11
[14] Itanı S, Lecron F and Fortemps P 2018 A multi-level classification framework for multi-site medical data: application to the ADHD-200 collection Expert Syst. Appl. 91 36–45
[15] Chen H, Song Y and Li X 2019 A deep learning framework for identifying children with ADHD using an EEG-based brain network Neurocomputing 356 83–96
[16] Dubreuil-Vall L, Ruffini G and Camprodon J A 2020 Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG Front. Neurosci. 14 251

ORCID iDs

Xin Wang https://orcid.org/0000-0003-3352-6829
Shixiong Chen https://orcid.org/0000-0002-5868-6952
[17] Moghaddari M, Lighvan M Z and Danihvar S 2020 Diagnostic ADHD disorder in children using convolutional neural network based on continuous mental task EEG Comput. Methods Prog. Biomed. 197 105738

[18] Vahid A, Bluschke A, Roessner V, Stober S and Beste C 2019 Deep learning based on event-related EEG differentiates children with ADHD from healthy controls J. Clin. Med. 8 10553

[19] Chen C, Hua Z, Zhang R, Liu G and Wen W 2020 Automated arrhythmia classification based on a combination network of CNN and LSTM Biomed. Signal Process. Control 57 101819

[20] Xu G, Ren T, Chen Y and Che W 2020 A one-dimensional cnn-lstm model for epileptic seizure recognition using eeg signal analysis Front. Neurosci. 14 1253

[21] Xu Z, Li S and Deng W 2015 Learning temporal features using LSTM-CNN architecture for face anti-spoofing 2015 3rd IAPR Asian Conf. on Pattern Recognition (ACPR) pp 141–5

[22] Tasdelen A and Sen B 2021 A hybrid CNN-LSTM model for pre-miRNA classification Sci. Rep. 11 1–9

[23] Bachrens D et al 2010 How to explain individual classification decisions J. Mach. Learn. Res. 11 1800–31

[24] Shrikumar A, Greenside P and Kundaje A 2017 Learning important features through propagating activation differences Int. Conf. on Machine Learning (PMLR) pp 3145–53

[25] Montavon G, Samek W and Müller K-R 2018 Methods for interpreting and understanding deep neural networks Digit. Signal Process. 73 1–15

[26] Simonyan K, Vedaldi A and Zisserman A 2013 Deep inside convolutional networks: visualising image classification models and saliency maps (arXiv:1312.6034)

[27] Bluschke A, Schuster J, Roessner V and Beste C 2018 Neuropsychological mechanisms of interval timing dissociate inattentive and combined ADHD subtypes Sci. Rep. 8 2033

[28] Abadi M et al 2016 Tensorflow: a system for large-scale machine learning 12th [USENIX] Symp. on Operating Systems Design and Implementation (OSDI) 16 pp 265–83

[29] Chollet F (ed) 2015 Keras GitHub (available at: https://github.com/chollet/keras)

[30] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGnet: a compact convolutional neural network for EEG-based brain-computer interfaces J. Neural. Eng. 15 056013

[31] Chollet F 2017 Xception: deep learning with depthwise separable convolutions Proc. IEEE Conf. on Computer Vision and Pattern Recognition pp 1251–8

[32] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift Int. Conf. on Machine Learning (PMLR) pp 448–56

[33] Clevert D-A, Unterthiner T and Hochreiter S 2015 Fast and accurate deep network learning by exponential linear units (elus) (arXiv:1511.07289)

[34] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting J. Mach. Learn. Res. 15 1929–58

[35] Greff K, Srivastava R K, Koutnìk J, Steunebrink B R and Schmidhuber J 2016 LSTM: a search space odyssey IEEE Trans. Neural Netw. Learn. Syst. 28 2222–32

[36] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization Hum. Brain Mapp. 38 5391–420

[37] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis J. Neurosci. Methods 134 9–21

[38] Oostenveld R, Fries P, Maris E and Schoffelen J M 2011 FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data Comput. Intell. Neurosci. 2011 156869

[39] Zhang D, Yao L, Chen K, Wang S, Chang X and Liu Y 2020 Making sense of spatio-temporal preserving representations for EEG-based human intention recognition IEEE Trans. Cybern. 50 3033–44

[40] Polich J J P M and Clinics R 2004 Clinical application of the P300 event-related brain potential Phys. Med. Rehabil. Clin. 15 133–61

[41] Tiolaki A C, Kosmidou V, Papadaniil C, Hadjileontiadis L, Adam A and Tiolaki M 2017 Brain source localization of MMN and P300 ERPs in mild cognitive impairment and Alzheimer’s disease: a high-density EEG approach Neurobiol. Aging 55 190–201

[42] Paul I, Gawrilow C, Zech F, Gollwitzer P, Rockstroh B, Odenthal G, Kratzer W and Wienbruch C 2007 If–then planning modulates the P300 in children with attention deficit hyperactivity disorder Neuroreport 18 653–7

[43] Haider A and Fazel-Rezai R 2017 Application of P300 event-related potential in brain-computer interface Event-Related Potentials and Evoked Potentials (Norderstedt: BoD–Books on Demand) (https://doi.org/10.5772/intechopen.69309)

[44] Krigolson O E, Williams C C and Colino F L 2017 Using portable EEG to assess human visual attention Int. Conf. on Augmented Cognition (Springer) pp 56–65

[45] Gray H M, Ambady N, Lowenthal W T and Deldin P 2004 P300 as an index of attention to self-relevant stimuli J. Exp. Soc. Psychol. 40 216–24

[46] Luck S J 2014 An Introduction to the Event-Related Potential Technique (Cambridge, MA: MIT Press)

[47] Kononowicz T W and Penney T B 2016 The contingent negative variation (CNV): timing isn’t everything Curr. Opin. Behav. Sci. 8 231–7

[48] Kornhuber H H and Deecke L 2016 Brain potential changes in voluntary and passive movements in humans: readiness potential and reafferent potentials Pflugers Arch. 468 1115–24