Neural Networks 144 (2021) 129–137

Contents lists available at ScienceDirect

Neural Networks

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

An end-to-end 3D convolutional neural network for decoding attentive mental state

Yangsong Zhang [a,d,1], Huan Cai [a,1], Li Nie [a], Peng Xu [b,*], Sirui Zhao [a], Cuntai Guan [c,*]

[a] School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China
[b] MOE Key Lab for Neuroinformation, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China
[c] School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
[d] Key Laboratory of Cognition and Personality, Ministry of Education, Chongqing, China

ARTICLE INFO

Article history:
Received 24 April 2021
Received in revised form 1 August 2021
Accepted 12 August 2021
Available online 20 August 2021

Keywords:
EEG
Attention
BCI
Deep learning
3D convolutional neural network

ABSTRACT

The detection of attentive mental state plays an essential role in the neurofeedback process and the treatment of Attention Deficit and Hyperactivity Disorder (ADHD). However, the performance of the detection methods is still not satisfactory. One of the challenges is to find a proper representation for the electroencephalogram (EEG) data, which could preserve the temporal information and maintain the spatial topological characteristics. Inspired by the deep learning (DL) methods in the research of brain–computer interface (BCI) field, a 3D representation of EEG signal was introduced into attention detection task, and a 3D convolutional neural network model with cascade and parallel convolution operations was proposed. The model utilized three cascade blocks, each consisting of two parallel 3D convolution branches, to simultaneously extract the multi-scale features. Evaluated on a public dataset containing twenty-six subjects, the proposed model achieved better performance compared with the baseline methods under the intra-subject, inter-subject and subject-adaptive classification scenarios. This study demonstrated the promising potential of the 3D CNN model for detecting attentive mental state.

1. Introduction

Attention refers to the capacity to concentrate on certain things. Human attention has made a tremendous impact (positive correlation) on our memory, learning, and cognitive ability, affecting our daily life (Qian et al., 2018). Attention Deficit and Hyperactivity Disorder (ADHD) is a developmental disorder of childhood with inattention symptoms. Children with ADHD usually exhibit a deficit of sustained attention, or lethargic problems.

Previous studies demonstrated that quantitative EEG (QEEG) could be a promising tool for ADHD diagnosis and treatment (Qian et al., 2019; Yeo et al., 2018). The attention level (attention or non-attention) of a subject can be measured and quantified from electroencephalogram (EEG), and serve as a control signal in the brain–computer interface (BCI) system (Wai, Dou, & Guan, 2020). Then, a potential therapy to effectively treat ADHD is adopting BCI-based neural feedback system, e.g., BCI-based attention training game system, to improve inattention symptoms (Fuchs, Birbaumer, Lutzenberger, Gruzelier, & Kaiser, 2003; Lim et al., 2012; Qian et al., 2018). Besides, this kind of BCI system is also valuable for improving memory and attention in healthy elderly, preventing Alzheimer's disease (Jiang, Abiri, & Zhao, 2017), etc. Therefore, in recent years, developing BCI-based attention training system has attracted increasing interest in the research community.

The performance of these BCI systems depends on the effectiveness of attention detection methods. In the early researches, some conventional methods were adopted (Fahimi et al., 2017; Liu, Chiang, & Chu, 2013). These methods first extracted handcrafted features from EEG data, and then adopted certain classifiers to classify these features. For instance, Hamadicharef et al. adopted the filter-bank and common spatial pattern filters to extract spectral–spatial features, and used a fisher linear discriminant classifier to classify these features (Hamadicharef et al., 2009). The obtained maximal and minimal accuracies were 89.4% and 60.1%, respectively. Liu et al. conducted research to monitor students' attention status during their learning process (Liu et al., 2013). Various common features, such as the power spectral density and energy value of each frequency band, were calculated and

* Corresponding authors.
E-mail addresses: zhangysacademy@gmail.com (Y. Zhang), caihuan_academy@163.com (H. Cai), nieliemily@163.com (L. Nie), xupeng@uestc.edu.cn (P. Xu), sirui@swust.edu.cn (S. Zhao), ctguan@ntu.edu.sg (C. Guan).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neunet.2021.08.019
0893-6080/

combined together. A polynomial kernel support vector machine (SVM) classifier was used to classify the combined features, and yielded an accuracy of 75.87%. In another study, Fahimi et al. explored the most discriminative features to classify attention and non-attention states based on a large dataset with 120 ADHD children (Fahimi et al., 2017). The common features, frequency band powers and their ratios, were firstly extracted from the EEG, and then subject-specific feature selection was performed with mutual information. An SVM classifier was used, and obtained accuracies of 65.42% and 65.55% on the conditions of within-session and cross-session tasks, respectively. In the methods described above, the procedures for feature extraction and classification were separated.

Different from conventional methods, deep learning (DL) has received increasing attention because of its end-to-end manner and simultaneous training of feature extractor and classifier. Because the DL methods have made remarkable achievements in many fields such as computer vision and speech recognition in recent years, they have been introduced for EEG studies, such as seizure detection (Raghu, Sriraam, Temel, Rao, & Kubben, 2020; Tsiouris et al., 2018; Wei, Zhou, Chen, Zhang, & Zhou, 2018), motor imagery (Kwon, Lee, Guan, & Lee, 2019; Ma, Qiu, Du, Xing, & He, 2018; Robinson, Lee, & Guan, 2019; Sakhavi, Guan, & Yan, 2018; Stieger, Engel, Suma, & He, 2021), mental workload classification (Zhang & Li, 2017), emotion recognition (Cui et al., 2020; Ding, Robinson, Zeng, & Guan, 2021), etc. DL method was also introduced to perform attention detection tasks. For instance, Fahimi et al. developed an end-to-end deep convolutional neural network (CNN) for attentive detection tasks using single bi-polar channel EEG data (Fahimi et al., 2019). The proposed method obtained an accuracy of 76.20% on the leave-one-subject-out(LOSO) approach, i.e., inter-subject classification. In that study, an adaptive technique was also adopted to evaluate the performance of the proposed method, i.e., subject-adaptive classification (Fahimi et al., 2019). For this classification approach, a pre-trained model was first obtained by training the DL model with training data from other subjects, and further fine-tuned by some training data from the test subject. This fine-tuning method could overcome the problem of data distribution shift when transferring the knowledge to the target subject (Zhang, Robinson, Lee, & Guan, 2021). With this subject-adaptive classification scenario, their model yielded a higher average recognition accuracy of 79.26%. The results indicated the potential of DL method for attention detection tasks using EEG data.

Inspired by previous studies that adopted the 3D representation for EEG signal (Chao & Dong, 2019; Zhao et al., 2019), we proposed an end-to-end 3D convolutional neural network (3D CNN) model for attention classification tasks using multiple channel EEG data. The proposed model could explore the spatial and temporal information simultaneously, and extract the features in a multi-scale manner by using three cascade blocks, each of which has two parallel branches of 3D convolution layers with different sizes of convolutional kernels. Inter-subject and subject-adaptive classification scenarios were used to evaluate the effectiveness of the proposed framework and baseline methods in this study. To verify whether the classification accuracy obtained with inter-subject classification strategy was higher than that obtained by training and testing based on single-subject data, an intra-subject classification was also used. Evaluated on a public dataset of twenty-six subjects, the experimental results show that the proposed method yields better performance compared with the baseline methods on the three classification strategies.

2. Materials and methods

2.1. Datasets and preprocessing

The experiment was evaluated on a public dataset, which was acquired from twenty-six healthy participants (Shin et al., 2016, 2018). This dataset provided both EEG and fNIRS recordings. Only the EEG data of the discrimination/selection response (DSR) task were adopted in the current study.

The DSR task included three sessions, and each session contained three series of twenty trials. Each subject performed 180 trials in the DSR task experiment. Each series contained an instruction period(2 s), a task period(40 s), and a rest period(20 s). During the instruction period, 'O: press a button' was presented on the monitor. During the task period, the experiment started with a 250 ms short beep, and ended with another 250 ms short beep followed a 'STOP' display lasting for 1 s on the monitor. For the remaining time in the task period, a symbol 'O' or 'X' was randomly selected to present for 0.5 s, and then a fixation cross for 1.5 s. When the symbol 'O' or the symbol 'X' was presented, the participants need to press the 'target' button (number 7) with their right index finger or 'non-target' button (number 8) with their right middle finger. In each series, the symbols 'O' and 'X' appeared at a 30% chance and 70% chance respectively. In the rest period, the subject was required to relax and gaze on the fixation cross on the monitor, and avoid excessive eye movements.

EEG data were recorded at a sampling rate of 1,000 Hz by a BrainAmp EEG amplifier, and subsequently resampled to 200 Hz. Thirty EEG active electrodes were placed on a stretchy fabric cap according to the international 10–5 system (Oostenveld & Praamstra, 2001). More descriptions of the experiment paradigm were provided in the Shin et al. (2018).

During the preprocessing procedure, the EEG data were band-pass filtered between 0.5 and 40 Hz, and then referenced to the average reference. EEG data in the DSR task period were served as attentive data, and the data in the rest period were served as the non-attentive data. 2 s sliding window with 1 s overlap was applied to segment the EEG data under two conditions. Because the periods of task and rest were not matching, the first half of the DSR task data in each trial were used in our attention classification task.

2.2. The proposed 3D CNN model

2.2.1. 3D representations for EEG signals

CNN is a kind of feed-forward neural network, and its representational learning ability has attracted much attention. Although the DL models using 2D convolution have achieved good results in EEG processing and analysis, the input EEG signal with the shape of electrodes $\times$ sample points in these models ignored the topological information between EEG data from different electrodes on the scalp. Owing to the exciting performance achieved in the image and video fields, 3D models were gradually applied to some brain signal classification tasks, such as motor imaginary (Lee, Jeong, Shim, & Kim, 2020; Zhao et al., 2019), emotion recognition (Chao & Dong, 2019), brain MRI segmentation (Chen, Dou, Yu, Qin, & Heng, 2018; Coupé et al., 2020), speech synthesis (Angrick et al., 2019) and brain tumor predicting (Elazab et al., 2020), etc. Inspired by these tasks, we introduced a 3D model for detecting attentive mental state in the current study.

As shown in Fig. 1, 28 EEG electrodes were converted into a 9×9 spatial matrix according to the spatial distribution on the cap. The relative locations of electrodes are described in Fig. 1(a). Black dots denote the electrodes used in this work, and red dots (TP9 and TP10) represent the reference and ground electrodes discarded in the following analysis. The middle part

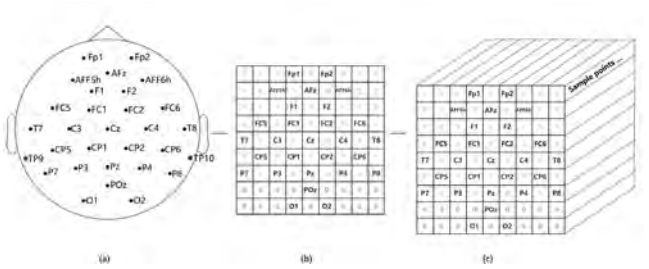

Fig. 1. The process of constructing a 3D representation for EEG data. (a) Spatial distribution of electrodes on the cap, (b) the corresponding 2D-matrix of the electrode locations, and (c) the 3D representation of the EEG data.

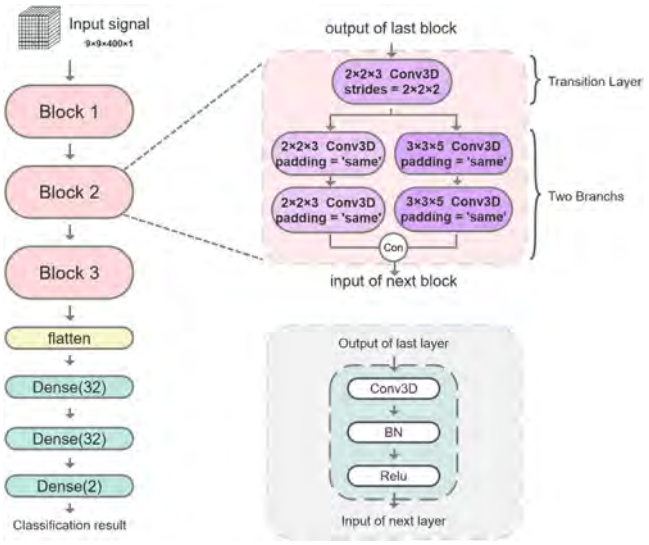

Fig. 2. The diagram of the proposed 3D CNN model.

of Fig. 1 describes the mapping matrix. Zero padding was used at the positions without electrodes. There are two reasons for this padding operation. One is to obtain rectangle-shaped data, which is more suitable for the inputs of the CNN model, the other is to better simulate the relative position of electrodes on the brain cap without introducing extra noises. Combining the two-dimensional spatial matrix with the temporal dimension, the three-dimensional representation of EEG data ($9 \times 9 \times 400$) was obtained as shown in Fig. 1(c). With this kind of representation, the spatial topological relationship between electrodes can be better extracted while thoroughly maintaining the temporal features. After expanding dimension in the final axis, a matrix with a size of $9 \times 9 \times 400 \times 1$ was input into the proposed 3D CNN model.

2.2.2. The structure of the proposed 3D CNN model

Inspired by previous studies (Zeng, Huang, Xu, Shen, & Chen, 2021; Zhao et al., 2019), an end-to-end 3D CNN model was proposed for attention classification task, and its diagram is shown in Fig. 2. The details of structure are summarized and demonstrated in Table 1, including the number of filters, kernel size, strides, and the activation functions.

The 3D CNN model consists of an input layer, three cascade blocks, a flatten layer, and three dense layers. The three cascade blocks are designed to extract multi-scale features. Each block contains a transitional layer with 3D convolution operation, and two parallel branches of 3D convolution with different sizes of kernels. For the transitional layer, it is used before the two parallel branches, which could extract some concrete features from the raw EEG signal or merge information from the previous block. The sizes of kernel and stride of this layer in Block 1 are set to $3 \times 3 \times 5$ and $2 \times 2 \times 4$, and those in Blocks 2 and 3 are set to $2 \times 2 \times 3$ and $2 \times 2 \times 2$, respectively. With this strategy, the network has a larger receptive field for improving attentive detection performance. For the two parallel branches as shown in the top right corner of Fig. 2, each branch consists of two cascade 3D convolution layers with the same number of kernels. The sizes of kernels in the two branches are different, which ensures the model to simultaneously detect spatial and temporal features with large and small receptive fields. The padding strategy is applied to keep the same size for the input and output of each layer. To some extent, this operation can prevent the rapid loss of the spatial information as the network became deeper. At the end of each branch, the results are combined together using a concatenation operation and input to the next block (denoted by the symbol 'con' in the gray circle). As described in the lower right corner of Fig. 2, each 3D convolution layer is followed by a bath-normalization layer and an activation function layer. The former can reduce the internal covariance shift and accelerate the training process, and the latter can enhance the nonlinear representation ability of the neural network.

After three cascade blocks, the obtained features are flattened into a vector, followed by three fully connected layers. Except for the activation function of the last dense layer being set as *Softmax*, the activation functions of other layers are set as *ReLU*.

2.3. Baseline methods

In order to evaluate the proposed 3D method, both conventional algorithm and DL models were adopted as baseline methods. Owing to its superior performance, FBCSP was used as the conventional baseline method in the current study (Ang, Chin, Zhang, & Guan, 2008). For DL models, DeepCNN proposed by Fahimi et al. was selected as the benchmark method of attention classification task (Fahimi et al., 2019). As the DL methods for detecting attentive mental state were limited, some popular DL models for EEG data classification were also adopted as baseline methods, i.e., ShallowNet, DeepNet, and EEGNet (Lawhern et al., 2018; Schirrmeister et al., 2017). These models have been applied to several BCI paradigms such as P300, movement-related cortical potentials, and sensory-motor rhythms, etc. Brief descriptions of these baseline models are presented as follows.

FBCSP (Ang et al., 2008). Filter Bank Common Spatial Pattern (FBCSP) is a state-of-art method for BCI classification (Ang, Chin, Wang, Guan, & Zhang, 2012). It can effectively optimize the subject-specific CSP features from multiple filter bands based on mutual information. FBCSP has been used for attention detection (Fahimi et al., 2019). Since the most suitable frequency range for attention detection is 0.5–40 Hz, nine frequency ranges of sub-bands were used, i.e., 0.5–8 Hz, 4–12 Hz, 8–16 Hz, …, 32–40 Hz, in the current study. Six-order Butterworth filter was used for bandpass filtering. We implemented the FBCSP method with the open source FBCSP toolbox which is available at https://fbcsptoolbox.github.io/.

ShallowNet (Schirrmeister et al., 2017). ShallowNet is a model with a shallow architecture inspired by the pipeline of the popular FBCSP algorithm (Ang et al., 2012). The first two layers are a temporal convolution layer and a spatial convolution layer, which perform the bandpass filtering and spatial filtering operations. After these layers, ShallowNet uses a squaring nonlinearity, a mean pooling operation, and a logarithmic activation function in sequence.

Table 1
The settings of the proposed 3D CNN model. *k* denotes the size of convolution kernel, *s* denotes the strides, and *BN* denotes the batch-normalization layer. *ReLU* and *Softmax* represent the corresponding activation function layer, and the number in brackets stands for the number of convolution kernels or units of fully connected layers.

Model Structure				Input shape	Output shape
block1	Transition	Conv3D(16), BN, Relu $k = 3 \times 3 \times 5$, $s = 2 \times 2 \times 4$		$9 \times 9 \times 400 \times 1$	$4 \times 4 \times 99 \times 16$
	Two-branch	conv3D(16), BN, Relu $k = 2 \times 2 \times 3$, pad = 'same'	conv3D(16), BN, Relu $k = 3 \times 3 \times 5$, pad = 'same'	$4 \times 4 \times 99 \times 16$	$4 \times 4 \times 99 \times 16$
		conv3D(16), BN, Relu $k = 2 \times 2 \times 3$, pad = 'same'	conv3D(16), BN, Relu $k = 3 \times 3 \times 5$, pad = 'same'		
		Concatenate		$4 \times 4 \times 99 \times 16$ and $4 \times 4 \times 99 \times 16$	$4 \times 4 \times 99 \times 32$
block2	Transition	Conv3D(16), BN, Relu $k = 2 \times 2 \times 3$, $s = 2 \times 2 \times 2$		$4 \times 4 \times 99 \times 32$	$2 \times 2 \times 49 \times 16$
	Two-branch	conv3D(16), BN, Relu $k = 2 \times 2 \times 3$, pad = 'same'	conv3D(16), BN, Relu $k = 3 \times 3 \times 5$, pad = 'same'	$2 \times 2 \times 49 \times 16$	$2 \times 2 \times 49 \times 16$
		conv3D(16), BN, Relu $k = 2 \times 2 \times 3$, pad = 'same'	conv3D(16), BN, Relu $k = 3 \times 3 \times 5$, pad = 'same'		
		Concatenate		$2 \times 2 \times 49 \times 16$ and $2 \times 2 \times 49 \times 16$	$2 \times 2 \times 49 \times 32$
block3	Transition	Conv3D(32), BN, Relu $k = 2 \times 2 \times 3$, $s = 2 \times 2 \times 2$		$2 \times 2 \times 49 \times 32$	$1 \times 1 \times 24 \times 32$
	Two-branch	conv3D(32), BN, Relu $k = 2 \times 2 \times 3$, pad = 'same'	conv3D(32), BN, Relu $k = 3 \times 3 \times 5$, pad = 'same'	$1 \times 1 \times 24 \times 32$	$1 \times 1 \times 24 \times 32$
		conv3D(32), BN, Relu $k = 2 \times 2 \times 3$, pad = 'same'	conv3D(32), BN, Relu $k = 3 \times 3 \times 5$, pad = 'same'		
		Concatenate		$1 \times 1 \times 24 \times 32$ and $1 \times 1 \times 24 \times 32$	$1 \times 1 \times 24 \times 64$
Full-Connected Layers		Dense(32), BN, Relu Dense(32), BN, Relu Dense(2), Softmax		1536 32 32	32 32 2

DeepNet (Schirrmeister et al., 2017). This model was simultaneously proposed with ShallowNet (Schirrmeister et al., 2017). The DeepNet was inspired by those successful architectures in computer vision. In contrast to ShallowNet, DeepNet has a deeper architecture that composed of four blocks. The first block includes a temporal convolution layer performing a convolution over time dimension and a spatial layer performing spatial filtering across the dimensions of all the electrodes and all the filters after temporal convolution. No activation function is used in these two layers. After the first block, three standard blocks with convolution and max-pooling are adopted. The final layer is a fully connected layer with a softmax activation function. A detailed description of DeepNet and ShallowNet could be found in the Ref. Schirrmeister et al. (2017).

EEGNet (Lawhern et al., 2018). This model has a compact CNN architecture designed for EEG analysis. It begins with a temporal convolution to capture frequency information, then a depthwise convolution is used to learn frequency-specific spatial filters. After these two layers, a separable convolution that is useful for EEG feature extraction is used to optimally learn the combined feature maps. Finally, a flatten layer is added, and a fully connected layer with softmax activation function is adopted for classification. EEGNet has become a benchmark model for different EEG classification tasks. A detailed description of the EEGNet can be found in Ref. Lawhern et al. (2018).

DeepCNN (Fahimi et al., 2019). DeepCNN is the first DL model based on CNN for detecting attentive mental state using raw EEG data. It is composed of three convolution layers which performed 1D convolution, a max-pooling layer, a flatten layer and two fully connected layers. The max-pooling layer is used after the first convolution layer. Besides, the dropout was implemented to avoid over-fitting problem. With inter-subject strategy, DeepCNN yields better performance than the baseline methods, such as ShallowNet and FBCSP, etc (Fahimi et al., 2019). DeepCNN demonstrated the great potentials of the end-to-end network structure for EEG data analysis, and the effectiveness to learn features from raw EEG data.

2.4. Model implementation and experimental evaluation

The three-dimensional representation of EEG signals, as shown in Fig. 1(c), was fed into the proposed 3D deep model as input, the two-dimensional EEG representation (electrods×timesteps) was fed into the other five baseline methods. To provide a fair comparison, the implementation of other baseline DL models was set the same as the original Refs. Fahimi et al. (2019), Lawhern et al. (2018) and Schirrmeister et al. (2017) except for DeepNet. To make the model fit for current data (400 sample points), the kernel size (1×10) and pooling size (1×3) of DeepNet were reduced to 1×5 and 1×2 respectively, which were initially designed for the EEG data with 522 sample points.

In the stage of the training process of DL models, the binary cross-entropy was used as the loss function. Adaptive moment estimation (ADAM) optimizer was adopted as the optimization method (Kingma & Ba, 2014), and the batch size was set to 32. The learning rate of intra-subject and inter-subject strategy was initially set to 0.001, and then multiplied by 0.4 every ten epochs. Note that the learning rate of the adaptive process of the subject-adaption method was much smaller, and set as 1e−4. The total number of training epochs of intra-subject, inter-subject and subject-adaptive methods are set to 30, 30 and 100 respectively.

Table 2
The Average accuracies and $f1$ scores of each method in intra-subject and inter-subject classification scenarios. The symbol '±' denotes the standard deviation.

Models	Intra-Subject		Inter-subject	
	Average accuracy(%)	$f1$ Score(%)	Average accuracy(%)	$f1$ Score(%)
FBCSP	66.94 ± 10.16	63.76 ± 15.86	63.36 ± 9.87	65.97 ± 11.06
ShallowNet	68.49 ± 9.04	64.52 ± 15.60	69.36 ± 8.65	65.28 ± 19.66
DeepNet	57.47 ± 6.61	67.85 ± 5.45	68.48 ± 9.21	71.24 ± 16.00
EEGNet	67.57 ± 9.45	**71.43 ± 10.93**	70.26 ± 8.38	69.73 ± 16.28
DeepCNN	66.37 ± 11.13	68.80 ± 11.20	70.75 ± 7.85	69.20 ± 12.27
3D(Ours)	**70.15 ± 9.70**	70.18 ± 10.95	**77.07 ± 7.23**	**75.48 ± 10.96**

All the DL models were implemented based on Keras framework with the TensorFlow backend (Abadi et al., 2016; Chollet, 2018).

The accuracy and $f1$-score were used as the metrics to evaluate the performance of all the methods used in current study. $f1$-score was calculated from the confusion matrix and reflected the comprehensive performance of each model. There was no randomness for the conventional FBCSP, so we only need to train the model once. For DL models, the models obtained at the last training epoch were taken as the final models to compute the metrics for test data. In order to avoid excessive randomness in DL models, intra-subject, inter-subject and subject-adaptive experiments were conducted ten times with different model initializations, and the average results were taken as the final results reported in this paper. The $f1$-score was calculated by the following formulas:

$$p = \frac{TP}{TP + FP} \tag{1}$$

$$r = \frac{TP}{TP + FN} \tag{2}$$

$$f1 = \frac{2 \times p \times r}{p + r} \tag{3}$$

where TP, FP and FN represent the number of True-Positive, False-Positive and False-Negative samples, respectively. Besides, the paired t-test was used to compute the level of statistical significance between the proposed model and each baseline model, respectively.

In the current study, we evaluated the performance of all the methods with three classification scenarios, i.e., intra-subject, inter-subject, and subject-adaptive classifications. For the intra-subject classification which is subject-dependent, the training and test data were both from a single subject. In current study, the EEG data were split into training data and test data in chronological order, namely, the first 80% data for training and the remaining 20% data for testing.

In practical application, the subject-independent scenario that performs inter-subject classification with transfer learning techniques could avoid the time-consuming calibration procedures which were required for intra-subject classification scenario. For the inter-subject classification, the LOSO approach was used, in which one subject was regarded as a target subject while other subjects were treated as the source subjects. The data of the target subject were used as the test set, and all the remaining data from other subjects were used as the training set.

For the subject-adaptive classification, a part of data from the target subject was used to fine-tune the pre-trained model with a much smaller learning rate (1e−4). In our experiment, the data of the target subject were divided into two parts in chronological order under task and resting conditions. We used two-fold cross validation to evaluate the performance of the methods. Specifically, one part of the data was used as the fine-tuning data (adaptive data) and another part as test data. This procedure was repeated twice. The average values of the evaluation metrics of the two-fold process were used as the final experimental results.

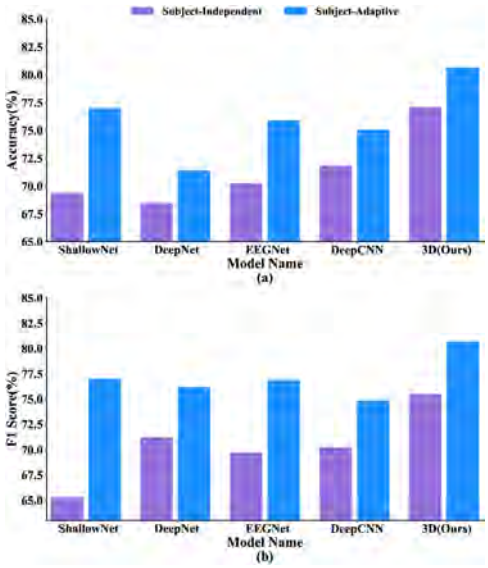

Fig. 3. The average accuracies (a) and $f1$ scores (b) of five DL models with the inter-subject and subject-adaptive classification.

Table 3
The accuracies(%) of the 3D model using different sizes of the matrices to represent the EEG data.

The size of 2D-matrix	6 × 7	8 × 7	9 × 9	10 × 11
Intra-subject accuracy(%)	67.96	67.80	70.15	**70.68**
Inter-subject accuracy(%)	75.12	73.74	**77.07**	74.26

3. Results and discussion

The experimental results of all methods are summarized in Table 2. For the FBCSP, the result of intra-subject classification was better than that of inter-subject classification. This may attribute to the inter-subject variability of EEG signals (Fahimi et al., 2019). For all DL models, the classification accuracies obtained with inter-subject strategy were higher than those with intra-subject strategy. This may be due to larger training set for DL methods. For the proposed 3D model, it achieved the best average accuracies of 70.15% and 77.07% in intra-subject and inter-subject classification, respectively. Except for the result on the intra-subject classification of EEGNet, the proposed method achieved higher $f1$ scores compared with other methods on both intra-subject and inter-subject classification. The paired t-test indicated the proposed model significantly outperformed other models ($p < 0.05$). Meanwhile, the standard deviations of different subjects on the proposed 3D model were small, especially for inter-subject classification.

For the subject-adaptive classification, the results are presented in Fig. 3. We found that DL models with adaptive method could yield better performance than inter-subject classification. The average accuracies got an improvement of 7.62%, 2.94%,

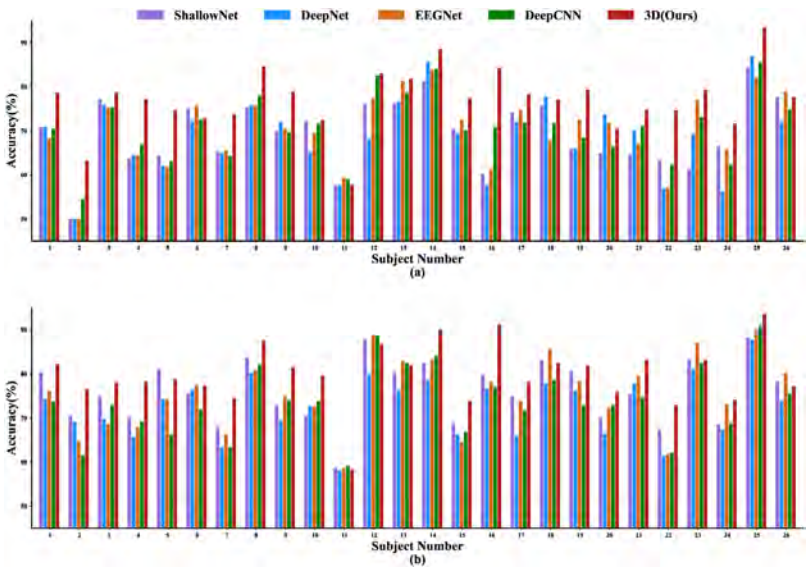

Fig. 4. The average accuracies of each subject with inter-subject (a) and subject-adaptive (b) classification strategies based on the proposed 3D model and other baseline models.

5.64%, 3.3% and 3.58% for ShallowNet, DeepNet, EEGNet and DeepCNN and the proposed 3D model, respectively. Besides, the ShallowNet achieved the best results among the four baseline methods in the adaptive approach, while it was relatively low in the inter-subject case. Structurally, both ShallowNet and EEGNet are shallower and have fewer parameters than the other models (see Table 4 for details). These two models are more likely to learn the general features of EEG data across different subjects. Therefore, ShallowNet and EEGNet are more conducive to transfer learning between different subjects, and have larger improvement with the adaptive approach. On the contrary, the models with more parameters, such as DeepNet, DeepCNN and the 3D model, learned more specific characteristics from the training data of source subjects, which results in the limited adaptive ability. With limited fine-tuning data from the target subject, these models showed smaller improvement on the target subject.

To further compare different models, the accuracies of each subject under inter-subject classification and subject-adaptive classification are depicted in Fig. 4. We found that the proposed model could yield better classification performance on most subjects when compared with other methods. Besides, t-distributed stochastic neighbor embedding (t-SNE) was used to project and visualize the learned embedding features (at the final fully connected layer of each model). The two-dimensional embedding features of different models obtained by t-SNE on four representative subjects were shown in Fig. 5. Only the inter-subject classification was considered. It can be found that the separability between attention and non-attention samples was relatively larger for the proposed 3D model, especially for subject 25, which can lead to better classification performance.

In addition, the confusion matrices of each DL model for inter-subject and subject-adaptive classification are outlined in Figs. 6 and 7. Note that the results shown in the two figures are the average results from ten times classification of each model. The larger the values on the main diagonal are, the better the classification performance of the model is. It can be found that the numbers of misidentified non-attention samples and misclassified attention samples from the 3D model were least among those from all the DL models under the strategies of inter-subject and subject-adaptive classification. These results further verify the effectiveness of the proposed 3D model.

In order to check the influence of matrix size to represent electrode spatial locations for the EEG data, we further investigated various sizes of the 2D matrices, i.e., 6×7, 8×7, 9×9, 10×11 as shown in Fig. 8. The results under different settings were listed in Table 3. We found that the size of 9×9 yielded the best accuracy in inter-subject classification, and 10×11 yielded the best accuracy in the intra-subject classification. These results indicate the optimal size of the representation matrix may vary in different conditions. It should be mentioned that the optimal size of the representation matrix may be different when EEG data with different numbers of channels are used.

Although the above results show that our proposed model yields promising results on both inter-subject and subject-adaptive classification scenarios, some limitations should be considered. For DL models, the total number of parameters, average training time across subjects and average test time across trials and subjects are listed in Table 4. The computation complexity of the proposed 3D model was higher than the baseline methods. In the future, the pruning algorithms on 3D convolutional networks can be used to optimize the proposed 3D model (Chen et al., 2020). Besides, although the proposed model performed better on most subjects, it failed on subjects 6 and 11. The huge differences among subjects may attribute to the subject heterogeneity and the experimental states during EEG recording.

Table 4
The number of parameters, average training time across subjects (minutes) and average test time across trials and subjects(ms) of each DL model in the inter-subject classification scenario.

Models	ShallowNet (Schirrmeister et al., 2017)	DeepNet (Schirrmeister et al., 2017)	EEGNet (Lawhern et al., 2018)	DeepCNN (Fahimi et al., 2019)	3D(ours)
Number of parameters	40,000	150,000	2,000	200,000	240,000
Training time	0.64	1.96	1.22	0.68	3.64
Test time	58.33	88.91	69.43	59.55	258.45

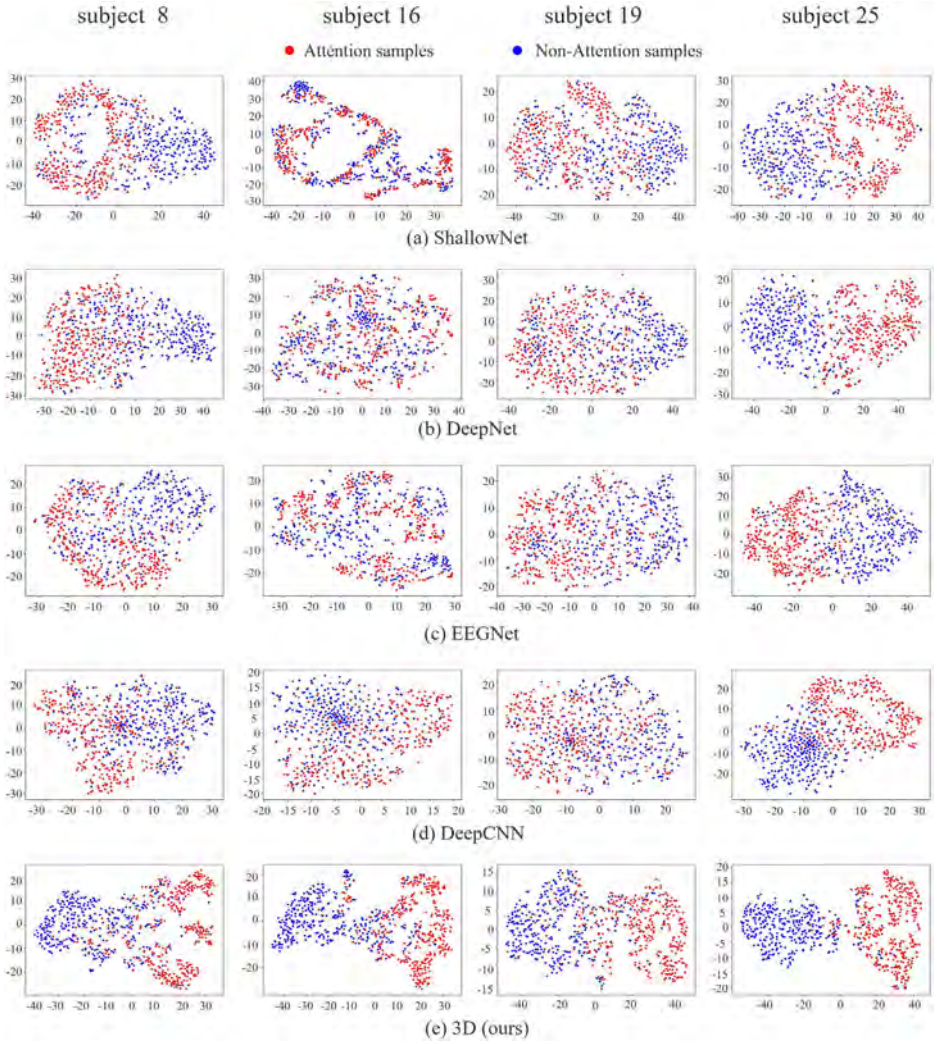

Fig. 5. The t-SNE visualization results of four representative subjects in inter-subject classification with the five models. The red color dots denote the features obtained with attention samples, and the blue color ones denote the features obtained with the non-attention samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

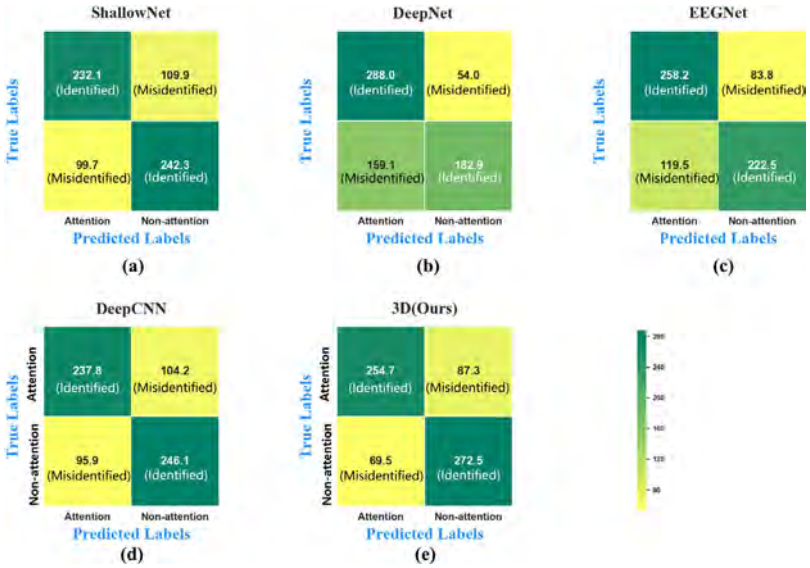

Fig. 6. The confusion matrix of each model with inter-subject classification. Numbers on the main diagonal represent the number of correctly identified samples, while numbers on the sub-diagonal represent the number of misidentified samples. The horizontal axis of each subfigure denotes the predicted tags, and the vertical axis represents the actual labels.

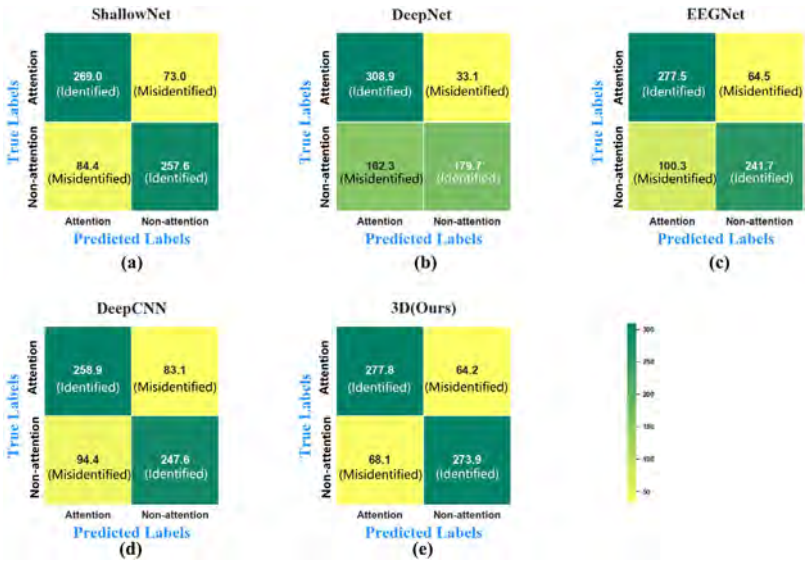

Fig. 7. The confusion matrix of each model with and subject-adaptive classification. Numbers on the main diagonal represent the number of correctly identified samples, while numbers on the sub-diagonal represent the number of misidentified samples. The horizontal axis of each subfigure denotes the predicted tags, and the vertical axis represents the actual labels.

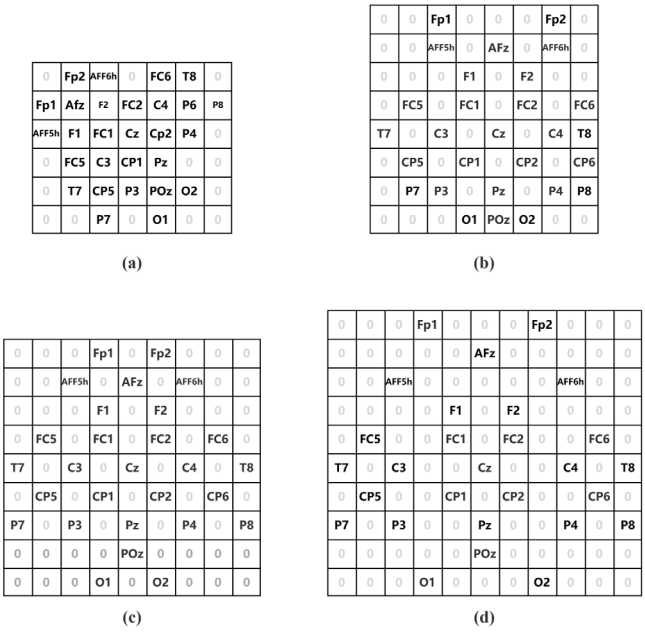

Fig. 8. Two dimensional matrix representation for the electrode spatial locations with different sizes of matrices. The sizes of matrices in the subfigures (a), (b), (c) and (d) are 6×7, 8×7, 9×9 and 10×11, respectively.

The open dataset has EEG and fNIRS recordings, but the fNIRS data were not considered in current study. Besides, only the EEG data with DSR task in the dataset were used to evaluate the DL models. The generalization of the proposed method should be investigated on different datasets with various cognitive task paradigms and recording modalities (Wai et al., 2020). We need to address these issues in our future study.

Owing to the nonlinearity and nonstationarity of EEG data, the distributions of EEG data vary across different subjects and even across different sessions on the same subject (Zhang et al., 2021). Based on the experimental results, we could find that the results of subject-adaptive were better than those of the inter-subject, which demonstrates that when some data from the target subject was available, it could help fine-tune and further enhance the model trained completely with data from other subjects. As acquiring calibration data is time-consuming and costly, more effective models need to explore in the future.

4. Conclusion

In current study, a 3D representation was utilized to maintain the spatial and temporal information for multichannel EEG data on attention detection task. Based on this representation, a novel 3D convolutional neural network model was proposed to simultaneously extract multi-scale features from EEG data with cascade and parallel 3D convolution operation, which could effectively detect attentive mental state. Evaluated on a public dataset, the proposed model yielded better performance compared with the baseline methods on the intra-subject, inter-subject and subject-adaptive classification scenarios, respectively. The extensive experiments demonstrated that the proposed model holds the promise to provide robust performance for BCI-based attention detection.

Declaration of competing interest

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

Acknowledgments

The authors sincerely thank all anonymous reviewers for their insightful suggestions and comments. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62076209 and No. 61871423.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467.

Ang, K. K., Chin, Z. Y., Wang, C., Guan, C., & Zhang, H. (2012). Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, *6*, 39.

Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2008). Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 2390–2397). IEEE.

Angrick, M., Herff, C., Mugler, E., Tate, M. C., Slutzky, M. W., Krusienski, D. J., et al. (2019). Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of Neural Engineering*, *16*(3), Article 036019.

Chao, H., & Dong, L. (2019). Emotion recognition using three-dimensional feature and convolutional neural network from multichannel EEG signals. *IEEE Sensors Journal*, *21*(2), 2024–2034.

Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P.-A. (2018). Voxresnet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage*, *170*, 446–455.

Chen, H., Wang, Y., Shu, H., Tang, Y., Xu, C., Shi, B., et al. (2020). Frequency domain compact 3D convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1641–1650).

Chollet, F. (2018). *Deep Learning with Python*. Manning Publications.

Coupé, P., Mansencal, B., Clément, M., Giraud, R., de Senneville, B. D., Ta, V.-T., et al. (2020). Assemblynet: A large ensemble of cnns for 3d whole brain mri segmentation. *Neuroimage*, Article 117026.

Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., & Chen, X. (2020). Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge-Based Systems*, *205*, Article 106243.

Ding, Y., Robinson, N., Zeng, Q., & Guan, C. (2021). Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition, arXiv preprint arXiv:2104.02935.

Elazab, A., Wang, C., Gardezi, S. J. S., Bai, H., Hu, Q., Wang, T., et al. (2020). Gp-gan: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal mr images. *Neural Networks*, *132*, 321–332.

Fahimi, F., Guan, C., Goh, W. B., Ang, K. K., Lim, C. G., & Lee, T. S. (2017). Personalized features for attention detection in children with attention deficit hyperactivity disorder. In *2017 39th annual international conference of the IEEE engineering in medicine and biology society* (pp. 414–417). IEEE.

Fahimi, F., Zhang, Z., Goh, W. B., Lee, T.-S., Ang, K. K., & Guan, C. (2019). Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. *Journal of Neural Engineering*, *16*(2), Article 026007.

Fuchs, T., Birbaumer, N., Lutzenberger, W., Gruzelier, J. H., & Kaiser, J. (2003). Neurofeedback treatment for attention-deficit/hyperactivity disorder in children: a comparison with methylphenidate. *Applied Psychophysiology and Biofeedback*, *28*(1), 1–12.

Hamadicharef, B., Zhang, H., Guan, C., Wang, C., Phua, K. S., Tee, K. P., et al. (2009). Learning EEG-based spectral-spatial patterns for attention level measurement. In *2009 IEEE international symposium on circuits and systems* (pp. 1465–1468). IEEE.

Jiang, Y., Abiri, R., & Zhao, X. (2017). Tuning up the old brain with new tricks: attention training via neurofeedback. *Frontiers in Aging Neuroscience*, *9*, 52.

Kingma, D. P., & Ba, J. A. (2014). A method for stochastic optimization. ArXiv preprint arXiv:1412.6980 434.

Kwon, O.-Y., Lee, M.-H., Guan, C., & Lee, S.-W. (2019). Subject-independent brain-computer interfaces based on deep convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(10), 3839–3852.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *15*(5), Article 056013.

Lee, D.-Y., Jeong, J.-H., Shim, K.-H., & Kim, D.-J. (2020). Classification of upper limb movements using convolutional neural network with 3D inception block. In *2020 8th international winter conference on brain-computer interface* (pp. 1–5). IEEE.

Lim, C. G., Lee, T. S., Guan, C., Fung, D. S. S., Zhao, Y., Teng, S. S. W., et al. (2012). A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder. *PLoS One*, *7*(10), Article e46692.

Liu, N.-H., Chiang, C.-Y., & Chu, H.-C. (2013). Recognizing the degree of human attention using EEG signals from mobile sensors. *Sensors*, *13*(8), 10273–10286.

Ma, X., Qiu, S., Du, C., Xing, J., & He, H. (2018). Improving EEG-based motor imagery classification via spatial and temporal recurrent neural networks. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society* (pp. 1903–1906). IEEE.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719.

Qian, X., Castellanos, F. X., Uddin, L. Q., Loo, B. R. Y., Liu, S., Koh, H. L., et al. (2019). Large-scale brain functional network topology disruptions underlie symptom heterogeneity in children with attention-deficit/hyperactivity disorder. *NeuroImage: Clinical*, *21*, Article 101600.

Qian, X., Loo, B. R. Y., Castellanos, F. X., Liu, S., Koh, H. L., Poh, X. W. W., et al. (2018). Brain-computer-interface-based intervention re-normalizes brain functional network topology in children with attention deficit/hyperactivity disorder. *Translational Psychiatry*, *8*(1), 1–11.

Raghu, S., Sriraam, N., Temel, Y., Rao, S. V., & Kubben, P. L. (2020). EEG based multi-class seizure type classification using convolutional neural network and transfer learning. *Neural Networks*, *124*, 202–212.

Robinson, N., Lee, S.-W., & Guan, C. (2019). EEG representation in deep convolutional neural networks for classification of motor imagery. In *2019 IEEE international conference on systems, man and cybernetics* (pp. 1322–1326). IEEE.

Sakhavi, S., Guan, C., & Yan, S. (2018). Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(11), 5619–5629.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*(11), 5391–5420.

Shin, J., von Lühmann, A., Blankertz, B., Kim, D.-W., Jeong, J., Hwang, H.-J., et al. (2016). Open access dataset for EEG+NIRS single-trial classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(10), 1735–1745.

Shin, J., Von Lühmann, A., Kim, D.-W., Mehnert, J., Hwang, H.-J., & Müller, K.-R. (2018). Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. *Scientific Data*, *5*, Article 180003.

Stieger, J., Engel, S., Suma, D., & He, B. (2021). Benefits of deep learning classification of continuous noninvasive brain-computer interface control. *Journal of Neural Engineering*, *18*(4), Article 046082.

Tsiouris, K. M., Pezoulas, V. C., Zervakis, M., Konitsiotis, S., Koutsouris, D. D., & Fotiadis, D. I. (2018). A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. *Computers in Biology and Medicine*, *99*, 24–37.

Wai, A. A. P., Dou, M., & Guan, C. (2020). Generalizability of EEG-based mental attention modeling with multiple cognitive tasks. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society* (pp. 2959–2962). IEEE.

Wei, X., Zhou, L., Chen, Z., Zhang, L., & Zhou, Y. (2018). Automatic seizure detection using three-dimensional CNN based on multi-channel EEG. *BMC Medical Informatics and Decision Making*, *18*(5), 111.

Yeo, S. N., Lee, T. S., Sng, W. T., Heo, M. Q., Bautista, D., Cheung, Y. B., et al. (2018). Effectiveness of a personalized brain-computer interface system for cognitive training in healthy elderly: A randomized controlled trial. *Journal of Alzheimer's Disease*, *66*(1), 127–138.

Zeng, D., Huang, K., Xu, C., Shen, H., & Chen, Z. (2021). Hierarchy graph convolution network and tree classification for epileptic detection on electroencephalography signals. *IEEE transactions on cognitive and developmental systems* (in press), http://dx.doi.org/10.1109/TCDS.2020.3012278.

Zhang, J., & Li, S. (2017). A deep learning scheme for mental workload classification based on restricted Boltzmann machines. *Cognition, Technology & Work*, *19*(4), 607–631.

Zhang, K., Robinson, N., Lee, S.-W., & Guan, C. (2021). Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network. *Neural Networks*, *136*, 1–10.

Zhao, X., Zhang, H., Zhu, G., You, F., Kuang, S., & Sun, L. (2019). A multi-branch 3D convolutional neural network for EEG-based motor imagery classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(10), 2164–2177.