PAPER

Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI

IOP Publishing Journal of Neural Engineering
J. Neural Eng. **16** (2019) 026007 (12pp) https://doi.org/10.1088/1741-2552/aaf3f6

Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI

Fatemeh Fahimi[1,2], Zhuo Zhang[2], Wooi Boon Goh[1], Tih-Shi Lee[3], Kai Keng Ang[2] and Cuntai Guan[1]

[1] School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
[2] Institute for Infocomm Research, Agency for Science, Technology and Research (A∗STAR), Singapore
[3] Duke-NUS Medical School, Singapore

E-mail: s150047@e.ntu.edu.sg

Received 8 July 2018, revised 12 November 2018
Accepted for publication 26 November 2018
Published 23 January 2019

Abstract

Objective. Despite the effective application of deep learning (DL) in brain–computer interface (BCI) systems, the successful execution of this technique, especially for inter-subject classification, in cognitive BCI has not been accomplished yet. In this paper, we propose a framework based on the deep convolutional neural network (CNN) to detect the attentive mental state from single-channel raw electroencephalography (EEG) data. *Approach.* We develop an end-to-end deep CNN to decode the attentional information from an EEG time series. We also explore the consequences of input representations on the performance of deep CNN by feeding three different EEG representations into the network. To ensure the practical application of the proposed framework and avoid time-consuming re-training, we perform inter-subject transfer learning techniques as a classification strategy. Eventually, to interpret the learned attentional patterns, we visualize and analyse the network perception of the attention and non-attention classes. *Main results.* The average classification accuracy is 79.26%, with only 15.83% of 120 subjects having an accuracy below 70% (a generally accepted threshold for BCI). This is while with the inter-subject approach, it is literally difficult to output high classification accuracy. This end-to-end classification framework surpasses conventional classification methods for attention detection. The visualization results demonstrate that the learned patterns from the raw data are meaningful. *Significance.* This framework significantly improves attention detection accuracy with inter-subject classification. Moreover, this study sheds light on the research on end-to-end learning; the proposed network is capable of learning from raw data with the least amount of pre-processing, which in turn eliminates the extensive computational load of time-consuming data preparation and feature extraction.

Keywords: attention, BCI, convolutional neural network, deep learning, EEG, end-to-end learning, inter-subject transfer learning

(Some figures may appear in colour only in the online journal)

1. Introduction

With the advent of deep learning (DL), state-of-the-art classification strategies and many other artificial intelligence tasks have been vastly improved. The emergence of DL can be associated with the advancement of the neural network, which itself dates back to the time when researchers had a desire to model the human brain [1]. The most popular types of deep neural networks (DNNs) include deep belief nets [2], recurrent neural networks [3], and convolutional neural networks

(CNNs). By achieving notable success in the ImageNet challenge [4, 5], deep CNN has become the centre of attention. In this paper, we have also built our methodology on the CNN.

DL first found successful applications in the fields of speech recognition and computer vision [6] and then gained attention in other research areas such as the brain–computer interface (BCI) [7, 8], which is the domain of our research in this paper. A BCI system records, processes, and translates brain signals into output commands for a wide variety of applications, such as assistive technology, neuro-rehabilitation, and cognitive enhancement [9]. Among different techniques for brain signal recording, electroencephalography (EEG) is the most studied modality in BCI research [10]. It provides a portable, non-invasive, and low-cost solution to capture the signal with high temporal resolution.

EEG-based cognitive BCI, which is the scope of this study, aims for assessment and enhancement of cognitive functions such as attention [11–14]. In these types of BCI systems, where the subject's attention level serves as a control signal, it is crucial to precisely detect the attentive mental state from the EEG. In this paper, following our previous work [15], we addressed the problem of attention detection from single-channel EEG by introducing a novel framework.

The prior-art methods for monitoring attentive mental state are mostly associated with specific fluctuations in EEG frequency bands. Many studies have investigated attention-induced fluctuations in beta [16, 17], alpha [18–20], and engagement between different frequency bands [21, 22]. Overall, they report that increased activity in high-frequency bands, such as beta, is an indicator of attentional arousal. Decreased theta/beta ratio (TBR), alpha activity, and theta activity also indicate higher attentive behaviour. In these studies, attentional information stored in spatial EEG has been underestimated.

Taking the importance of spatial information into account, Hamadicharef *et al* introduced a novel approach for attention level measurement from EEG [23]. Using two filters in a row, including a filter bank and a common spatial pattern (CSP), they extracted spectral-spatial features from an EEG, which was recorded using multiple electrodes placed in various brain regions. Then, the extracted features were sent to a fisher linear discriminant classifier for classification [23]. Their approach outperformed the conventional methods based on only spectral features. In the case of a lack of spatial information (i.e. single-channel BCI), Fahimi *et al* introduced a framework to differentiate attention from non-attention in a subject-specific manner [24]. They extracted several relative and ratio frequency band powers and performed mutual information based feature selection to find the most informative features for each individual.

Overall, in current methods of feature extraction, reduction of the signal into a few values neglects the dynamics of the signal and its temporal information. In addition to this problem, building a classification framework which is able to deal with the non-stationarity and high-dimensionality of EEG has always been a big challenge [25]. Deep CNNs, with their ability for handling high-volume datasets, better learning algorithms and faster computational resources, are becoming a superior alternative to the EEG classification task.

Although, to the best of our knowledge, DL has not been utilized so far for the detection of mental attention from EEG, there have been some attempts to apply DL for other purposes in EEG-based BCIs. Tabar and colleagues boosted the classification accuracy of motor imagery (MI) BCI by proposing a deep network composed of a CNN and stacked auto-encoders (SAE). In their work, an EEG was converted into images using short time Fourier transform (STFT). Then, these images were fed into a 1D CNN (convolution over time) for feature learning. The learned features were then sent into a SAE network for classification [26]. The performance of their proposed network was investigated in a BCI competition IV-2b dataset. The authors report that their methodology achieves a higher classification accuracy than the winner of the competition [26]. Jirayucharoensak *et al* also used SAE to build a DL network [27]. They extracted principal components of power spectral densities from 32 EEG channels as input to their proposed DL network, comprised of three auto-encoders and two Softmax layers, in order to classify different levels of emotion [27].

In a more recent study, Sakhavi *et al* developed a new classification framework for MI-based BCI by introducing envelope representation of EEG using Hilbert transformation and passing it through a CNN [28]. Their data representation was inspired by a filter bank common spatial pattern (FBCSP). They claimed that by applying their algorithm to a BCI competition IV-2a dataset, they beat the state-of-the-art classification accuracy reported so far [28].

In another work, Lu *et al* introduced a DL network based on a restricted Boltzmann machine (RBM) for MI classification. They called it a frequential deep belief network (FDBN). In FDBN, frequency representation of an EEG, generated using fast Fourier transform (FFT) and wavelet decomposition techniques, passes through three RBMs and an extra output layer for classification [8]. Zhang and Li also employed an RBM to develop a DL scheme, but for a different purpose: mental workload (MWL) classification [29]. They considered EEG channels with relatively higher importance simply based on the network weights between the input layer and the first hidden layer. Another study used recurrent-CNN for MWL classification [30]. In their approach, EEG time series were transformed into spectral images before being used in the deep recurrent-convolutional network. They suggested that such a representation of the data preserves temporal, spectral and spatial information [30].

Ma *et al* targeted learning discriminative motion-onset visual evoked potential (mVEP) features by using a combination of multi-level compressed sensing and RBM [31]. They reported that deep features, obtained from this method, performed better than conventional amplitude-based features. They used a support vector machine (SVM) for classification [31]. An aspect which should have been further considered in their work is optimal channel selection. It is more efficient to consider only channels with strong visual evoked potentials and exclude those with irrelevant information.

To provide an insight into the neurophysiological phenomena affecting the decision of DNNs, Strum and colleagues put forward the idea of using layer-wise relevance propagation (LRP) with a DNN. In their methodology, LRP, in a backward way, decomposes the network decision into some values which are defined as the relevance of each input component with the decision [32]. In terms of classification accuracy, their methodology did not outperform the CSP with linear discriminant analysis (LDA) classifier.

In the present paper, as a follow-up to our previous work [15], we enhance the detection of attentive mental state from an EEG signal by building an effective CNN-based classification framework. We develop a framework which addresses the problems of: (1) deterioration of classification accuracy due to information loss caused by feature extraction, (2) inter-subject transfer learning, and (3) the interpretability of what the CNN learns. To address the first problem, we develop an end-to-end network that can efficiently learn from raw EEG data instead of pre-extracted properties. This also removes the computational load of unnecessary processing. To solve the second issue, we implement the classification strategy with inter-subject transfer learning techniques. In one approach, the network learns a general model based on the data from a pool of subjects. Then, it transfers the knowledge to a new subject. In a more adaptive approach, the model will be updated based on a subset of a new subject's samples. In this way, the problems of time-consuming re-training and low inter-subject classification accuracy will be addressed. It also guarantees the application of the proposed framework for real time BCI systems. Finally, to interpret the features learned through the network, we visualize the network perception of each class (attention/non-attention). The comparison of the proposed method with the baseline methods [22] verifies that the introduced framework outperforms the state-of-the-art performance. The proposed framework has also been applied to a multi-channel dataset to investigate the performance and generalizability of the method. The results suggest that the end-to-end framework is also promising for multi-electrode settings.

The rest of this paper is organized as follows: section 2 describes the data and recording protocol. It then continues with presenting the proposed methodology, including pre-processing, data representations, and the deep CNN structure. Section 3 presents the results and section 4 provides a comprehensive discussion. Finally, section 5 concludes the study.

2. Materials and methods

2.1. Data

This study uses EEG data collected from healthy subjects as part of a clinical trial registered under NCT02228187 at clinicaltrials.gov. Note that this study is not a clinical trial and does not report on clinical outcomes; it only uses the EEG data.

A total of 120 healthy subjects performed the Stroop color test, which is a well-known task for the study of attention [33, 34]. It can be traced back to John Ridley Stroop, who reported the Stroop effect in his work in 1935 [35]. Then, it gained great attention in the fields of cognitive sciences and psychology, such that a wide variety of experiments based on the Stroop effect have been studied in these fields [33, 36].

During the test, a colored word was presented on a screen and subjects were asked to name the color in which the word was written. In fact, subjects were experiencing a conflict of information; what the word said and what was the color of the word. Thus, subjects needed to obtain and maintain their attention during the Stroop color task [37]. During each session, participants performed 40 repetitions of the Stroop test (attention) followed by a rest period (non-attention). Therefore, they underwent a change of mental state (attentive/non-attentive) during the task. Overall, each session took approximately 10 min. Figure 1 shows the recording protocol and an example of a task demonstration.

To ensure the easement of elderly participants in the long-term treatment program, their brain activity was recorded using a dry EEG headband with a single bi-polar channel, which was positioned at the frontal area (Fp1–Fp2). The sampling frequency was 256 Hz. There is strong evidence from several studies which proves the efficiency of frontal EEG channels for studying attention-related tasks [11, 14, 22, 24, 38].

2.2. Pre-processing

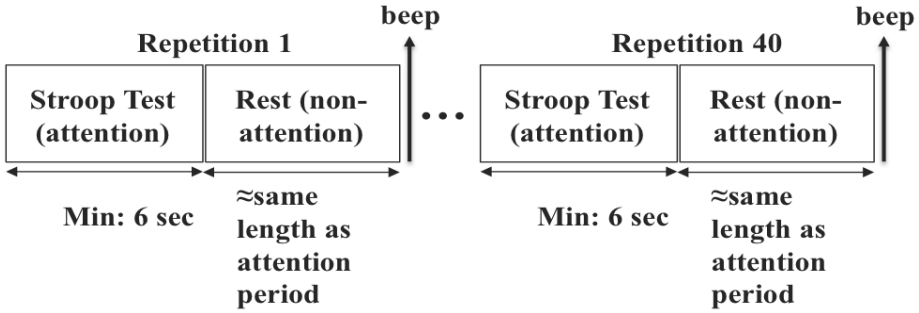
We applied a 2 s sliding window with 50% overlapping to segment the continuous EEG time series. The rationale behind choosing 2 s for the EEG segment length was that the subjects took 2 s on average to respond to each question in the Stroop task. Data were visually screened to discard any noisy trials. Additionally, since the maximum amplitude of an EEG recorded from the scalp is 100 μv [39], we set a threshold at ± 100 μv to discard the segments affected by ocular artefacts or other noises. We also filtered the EEG above 0.5 Hz to eliminate any plausible low-frequency artefacts that remained.

2.3. Deep CNN

2.3.1. Input representation. To preserve the information and minimize the computational load, we prepare the raw EEG with a minimal amount of processing as input. In fact, we avoid any pre-feature extraction and/or transferring the EEG into images, which are the main sources of information loss and computational costs. Based on the single-channel EEG data which provides a 1D input for the network, we defined three input representations for the network without pre-extracted features. In all representations, the segments were down-sampled by three with respect to the original value of 256 Hz, resulting in 171 time points for 2 s intervals.

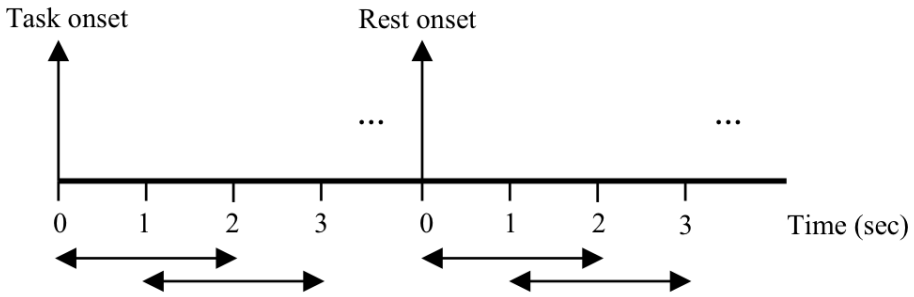
- Data Representation 1 (DR1): Raw EEG data were pre-processed as described in section 2.2.
- Data Representation 2 (DR2): Raw EEG segments were band-pass filtered at 0.5–40 Hz.
- Data Representation 3 (DR3): Raw EEG segments were filtered at five classical bands; δ (0.5–4 Hz), θ (4–8 Hz), α (8–12 Hz), β (12–30 Hz), and low γ (30–40 Hz).

Note that in all representations, the data were first pre-processed as described in section 2.2 to remove artefacts.



(a)

(b)



(c)

Figure 1. The Stroop color task: (a) the recording protocol, (b) the test display, and (c) a segmentation diagram.

2.3.2. Network architecture. The early CNN (LeNet-5) introduced by LeCun [40], was composed of a sequence of convolution and pooling layers. Since then, numerous attempts have been made to upgrade the CNNs through some extensions, such as batch normalization [41] and dropout [42], in order to accelerate training, avoid over-fitting and better preserve the information. In this study, we also exploit some of these techniques.

In convolutional layers, the filter (kernel) convolves over the input and produces element-wise multiplications. These numbers will be summed up and will produce a single value for that receptive field. Repeating this procedure by sliding the filter all over the input generates a single value for each receptive field. It will eventually produce the activation map or feature map as the output of the convolutional layer. Using the subsequent pooling layer aims to reduce the dimension of the feature map by replacing each patch with a single value based on the operation of interest (for example, maximum for max-pooling). As the input passes through the layers, high-level feature maps will be generated. For classification tasks, the last layer of the CNN is a fully connected layer which takes the output of the previous layer and outputs an n-dimensional vector (n is the number of classes). In Softmax, for example, each element of this vector represents the probability that the original input belongs to the corresponding class. In this procedure, the network parameters are learned through back-propagation.

In the present study, the EEG data representations, as described in section 2.3.1, are imported into the network as the input. Since the input data are a time series, a 1D filter has been used across time for convolution. The effectiveness of using a 1D filter across time, even for 2D inputs, has been proved in the literature [26, 28]. To generate high-level features, we inserted three convolutional layers with a 1D filter for the network. The first layer, with 60 filters and a kernel size 1×4, is followed by a max-pooling layer with a pool size 1×2. The output of the max-pooling passes through the second convolution layer with 40 filters and kernel size 1×3. Finally, after the third convolution layer, with 20 filters and a kernel size 1×2, the generated feature maps are flattened into a vector. This vector then passes through a dropout layer with a probability of 20% before being fed into the first fully

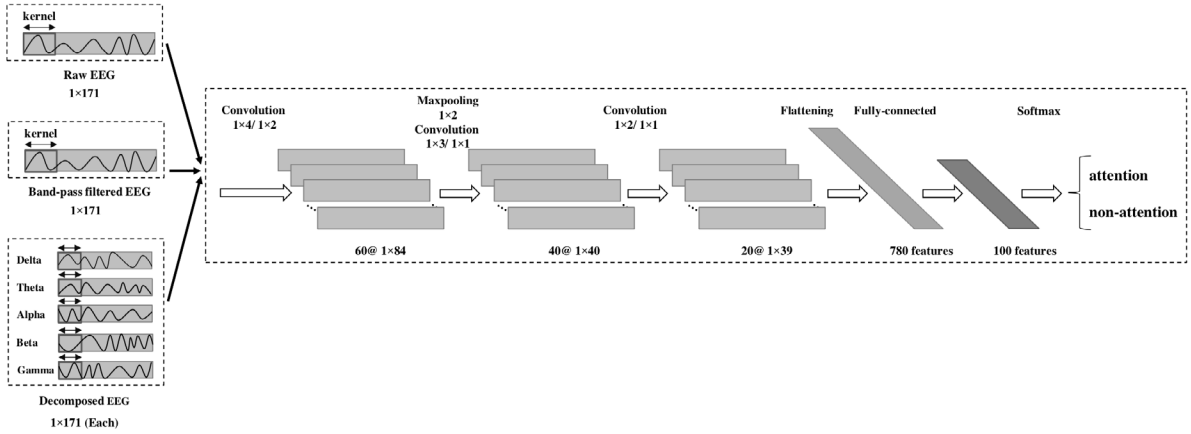


Figure 2. A schematic diagram of the end-to-end CNN-based classification framework for transfer learning. The first tuple under convolution refers to kernel size and the second tuple shows the stride. After the learning process, the learned features will be classified by Softmax. The left boxes are associated with data representation 1, 2, and 3, respectively. Note that in the case of data representation 3, the input EEG fed into the network is stored in five frequency channels but the network structure remains the same.

connected layer of size 100. Then, we inserted the second dropout layer with a probability of 30% before the second fully connected layer (Softmax) to overcome the over-fitting. Finally, the features are fed into the Softmax layer for classification. Note that by decreasing the temporal dimension over layers, a smaller kernel size is used. The activation function of a type rectified linear unit (ReLU) has been employed after each convolution layer and the first fully connected layer. For the optimization algorithm, we applied the ADAM method [43]. Figure 2 depicts a schematic diagram of the end-to-end deep CNN-based classification framework for inter-subject transfer learning in BCI.

3. Results

In this work, the DL experiments were conducted in a Python environment on an Ubuntu system powered by a NVIDIA GeForce GPU. The baseline methods, which have been described below, are implemented in a Matlab R2013b environment on an Intel Xeon CPU @3.5 GHz with 16 GB RAM (except for the classification stage of baseline 1 that is conducted in Python).

3.1. Baseline

To provide a fair baseline for the proposed technique, we implemented the classification framework as introduced in [22] for the single-channel data to differentiate between attention and non-attention states. Additionally, to be consistent with the proposed data representations, we performed the conventional feature extraction and classification method using the same frequency bands as described in data representation 3. Note that other techniques for attention detection/measurement presented in other studies exploit the spatial information of multi-channel EEG [23], which is not feasible to implement in the case of single-channel EEG.

According to the method in [22], frequency band energies including delta (0.5–3 Hz), theta (4–7 Hz), alpha (8–13 Hz), beta (14–30 Hz), and the alpha beta ratio were extracted using FFT and sent into an SVM with a polynomial kernel function for classification.

As the second baseline, we band-pass filtered the data in five subsequent frequency bands, including δ, θ, α, β, and low γ (as described in DR3), using Chebyshev type II. Then, the band powers were computed (mean of the squared values) and sent into LDA for classification. Note that unlike [22], which used *k*-fold cross validation, we used the inter-subject classification approach (leave-one subject-out (LOO)) in both baseline methods to provide a fair comparison with the results of the deep CNN. Baseline 1 reached an average accuracy of only 50.70%. Additionally, to improve the accuracy, we normalized the features of baseline 1. As a result, the average accuracy improved to 67.90%. The left side of table 1 summarizes the baseline results. As can be seen, baseline1 and 2, respectively, led to average accuracies of 67.90 and 68.23 with no statistically significant difference between them ($p\_$ value $= 0.87$). More than 50% of subjects have an accuracy below 70% (as the accepted threshold for BCI performance [44, 45]). It requires a lot of effort to increase the accuracy for these subjects.

We found another study which has attempted to classify attention from frontal single-channel EEG data [38]. In this study, the Neurosky device was used for EEG recording. This device generates the attention indicator and other information, such as frequency band powers. The authors simply used the attention indicator obtained from the device to detect the attentive state using an LDA classifier. Initially, ten subjects were involved in the experiment but four of them failed to control their attention level (based on the attention indicator) and were excluded. Thus, the classification was conducted on the small population of six subjects. The average accuracy is 79.5% based on their paper. The main limitation of their work, besides small sample size, is the method of classification that is performed for each subject on each session separately and then averaged over the sessions. This simplified way of classification (within subject and within session) will certainly be deteriorated by subject-to-subject and session-to-session variations. They also reported that including frequency band

Table 1. The average accuracy for the baseline and proposed methods. Std refers to the standard deviation.

	Baseline Methods		End-to-end deep CNN with transfer learning methods					
			CNN-LOO			CNN-subject adaptation		
	FFT-SVM [22]	DR3-LDA	DR1	DR2	DR3	DR1	DR2	DR3
Accuracy (Std)	67.90 (11.02)	68.23 (10.89)	76.20 (8.98)	75.07 (8.50)	76.68 (8.80)	79.26 (7.67)	78.12 (7.75)	79.86 (7.69)
Range (Min–Max)	64.56 (22.06–86.62)	62.06 (26.31–88.37)	44.06 (48.24–92.30)	44.45 (46.84–91.29)	40.46 (51.92–92.38)	35.24 (58.45–93.69)	38.67 (53.15–91.82)	36.02 (58.78–94.80)
Population with accuracy $<70\%$	54.17%	50.84%	26.67%	24.17%	23.34%	15.83%	17.50%	15.83%

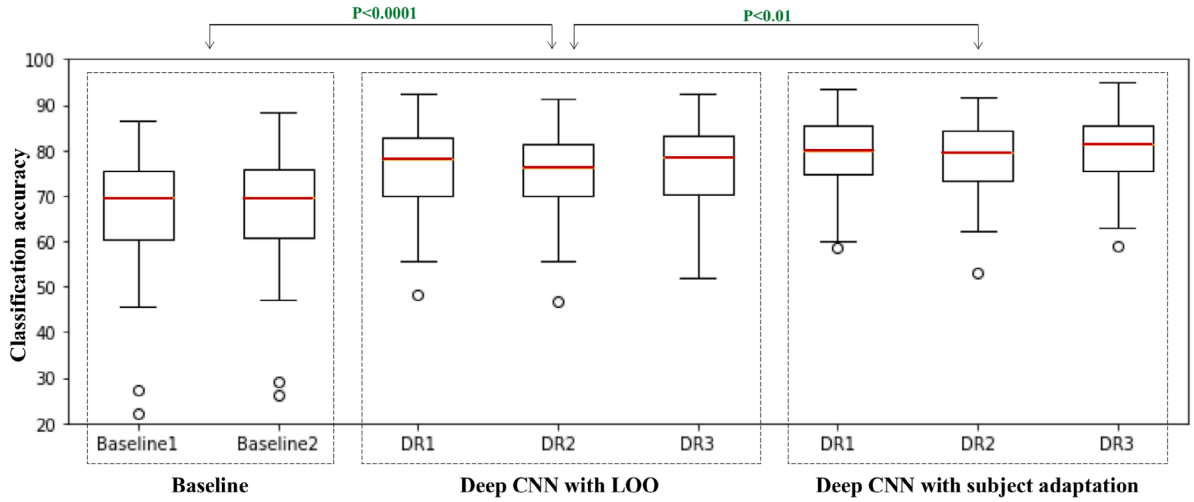


Figure 3. A comparison of the performance of baseline and end-to-end deep CNN methods in attention detection. A classification framework based on deep CNN (both strategies; LOO and subject adaptation) significantly outperforms the baseline methods. Note that there are no statistically significant differences between the methods in each group (baseline, deep CNN with LOO, and deep CNN with subject adaptation). *P*_values are calculated using a Wilcoxon test. The circles are the outliers; subjects with smaller accuracy than the lower extreme.

Table 2. The results on the multi-channel dataset. Std refers to the standard deviation.

	FBCSP [47]		Shallow ConvNet [7]	End-to-end deep CNN	
	Intra-subject	LOO	LOO	LOO	Adaptive
Accuracy (Std)	80.01 (6.43)	60.79 (6.74)	72.82 (6.54)	79.10 (7.60)	89.32(4.47)
Range (Min–Max)	18.83 (72.83–91.67)	19.42 (55.25–74.67)	16.27 (65.33–81.60)	18.30 (72.67–90.97)	11.69 (82.66–94.35)
Population with accuracy <70%	0%	87.50%	50%	0%	0%

powers did not improve the classification accuracy. Note that since the attention indicator used for the classification is generated by the recording device and no details of the algorithm are provided, it was not feasible to implement their methodology as baseline for our data.

3.2. Deep CNN with LOO

In the LOO approach, a generalized network will be learned using the data from a pool of subjects (source) and then the learned knowledge will be transferred to the new subject (target). This is actually a type of inter-subject transfer learning. Since re-training is not required, this method will be relatively less computationally demanding. In this study, we trained the network on the data from all the subjects, excluding the target subject, and transferred the information to the target subject. Execution of this method led to significantly better accuracies than baseline (p_value < 0.0001), with 7.92% improvement on average. The average accuracies for DR1, DR2, and DR3 are, respectively, 76.20%, 75.07%, and 76.68% with no statistically significant difference between them. This method also showed a considerable drop in the percentage of subjects with accuracies below 70% (as threshold [44, 45]) with only 26.67%, 24.17%, and 23.34% of the total 120 subjects for DR1, DR2 and DR3, respectively.

3.3. Deep CNN with subject adaptation

Although the zero-shot learning method avoids long time training for a new subject's data, this approach might encounter the problem of information change/shift when transferring the knowledge from the source to the target. To resolve this issue, we conducted the adaptive method in which re-training is carried out on a small sample size of a new subject's data. In this way, the problems of excessive re-training time and information shift can both be addressed.

In this study, we used half of the new subject's samples for adaptation (two-fold). This strategy surpasses the baseline and LOO methods by achieving 79.26%, 78.12%, and 79.86% average accuracies for DR1, DR2, and DR3, respectively. This means, on average, an 11.02% increase compared to baseline (p_value < 0.0001) and a 3.10% increase compared to LOO (p_value < 0.01). The population of subjects with poor performance decreased to only 15.83%, 17.50%, and 15.83% of the total 120 subjects for DR1, DR2, and DR3, respectively. Table 1 summarizes the results of the baseline and end-to-end deep CNN methods. The performance of the different methods discussed can be visually compared in the box plot of the results shown in figure 3. Overall, the CNN with the subject adaptation technique achieves the best performance. Although there is a statistically significant difference between the CNN methods (LOO and subject adaptation), there is no

Table 3. The confusion matrix of the deep CNN classification results.

CNN with LOO						
	Class 1			Class 2		
	DR1	DR2	DR3	DR1	DR2	DR3
Class 1	81.32	77.45	82.02	18.67	22.54	17.97
Class 2	28.92	27.25	28.59	71.07	72.74	71.40
CNN with subject adaptation						
	Class 1			Class 2		
	DR1	DR2	DR3	DR1	DR2	DR3
Class 1	78.77	77.81	79.26	21.22	22.18	20.73
Class 2	21.17	22.55	20.40	78.82	77.44	79.59

significant difference between the data representations within each method.

3.4. Results on a multi-channel public dataset

To investigate the generalizability of the proposed framework, we applied the network on a multi-channel dataset. The data has been collected for a study on covert attention [46]. A total number of eight healthy subjects (18–27 years old) participated in the experiment and their EEG was recorded using a 64-channel cap with the electrodes placed based on the international 10–10 system. The sampling frequency during recording was set at 1000 Hz, which was later down-sampled to 200 Hz. The experiment includes the sequences of attention, response, and rest. We have segmented the EEG during the attention and rest parts for the classification task. Based on the original study of this dataset [46], a subset of nine electrodes, including PO3, 4, 7–10, Oz, O1, and O2, are the optimal electrodes for studying attention. We have also used these nine recommended electrodes in our study.

As the first baseline for the multi-electrode dataset, we implemented the popular method of FBCSP [47]. The methods of mutual information-based best individual feature (MIBIF) and the naive Bayesian Parzen window (NBPW) have been used for feature selection and classification, respectively, as in [47]. In addition to classification with LOO, which provides the results for fair comparison with the end-to-end framework, we also performed intra-subject classification with 10-fold cross validation.

The second baseline we used is the method of shallow CNN, as introduced in [7]. This network, which is called shallow ConvNet, is inspired by the FBCSP method. Briefly, it has two hidden layers that perform temporal convolution and spatial filtering for band power feature decoding. They report that, unlike FBCSP, this method jointly optimizes all the computational steps through a single network [7].

Table 2 presents the results. Overall, shallow ConvNet, which is built based on FBCSP, beats the FBCSP method and end-to-end deep CNN outperforms both baseline methods. Comparing the LOO results, the performance of the proposed method is significantly better than the FBCSP (+18.31%, $P\_value < 0.001$) and shallow ConvNet (+6.28%, $P\_value < 0.001$). In fact, the results of the LOO classification with the end-to-end framework are as good as the results of the intra-subject classification with the FBCSP. This shows that although FBCSP demonstrates a good performance in intra-subject classification, it fails to produce acceptable results when it comes to inter-subject classification (19.21% decrease). The observations suggest that CNN-based methods can potentially be used to address this problem. The proposed end-to-end deep CNN decodes more than 70% of the EEG trials correctly for all eight subjects.

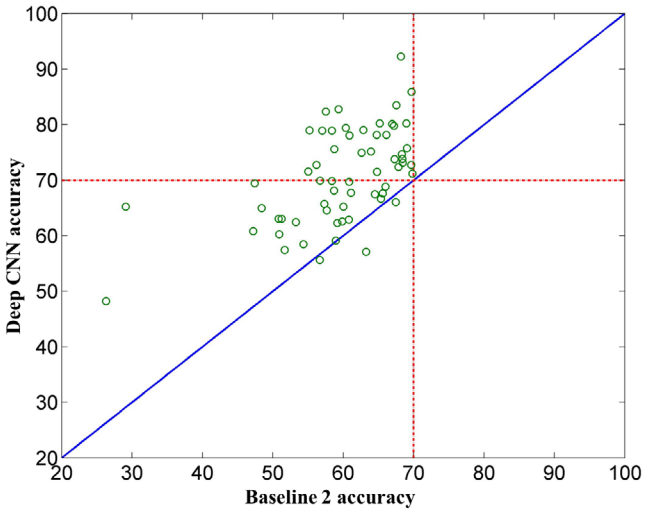


Figure 4. Classification accuracy for subjects with poor performance (<70%) at baseline. For simplicity in the comparison only baseline2 and the deep CNN with LOO on DR1 are plotted. The end-to-end deep CNN framework dramatically increases the performance of these subjects by a 10.84% increase in the average accuracy and a 50.82% decrease in the number of these subjects (61 reduced to 30).

4. Discussion

4.1. End-to-end CNN by learning from raw EEG data preserves the information and boosts the classification accuracy

EEG classification with minimal pre-processing and feature extraction is always a worthy goal. For this reason, we conducted an exploratory evaluation of several data representations without pre-extracted features as an input to a CNN. The objective was to learn from the raw EEG; an end-to-end study. The first representation (DR1) is a raw EEG with the least amount of pre-processing (to remove artefacts). The CNN with this representation as the input outperforms the baseline ($p < 0.0001$) with 8.14% (LOO) and 11.20% (adaptive) improvement in the average accuracy. Going one step further in data preparation, we band-pass filtered the data at 0.5–40 Hz (DR2) and fed it into the CNN for classification. Interestingly, the average classification accuracy dropped by 1.13% in LOO ($p > 0.1$) and by 1.14% in the adaptive method ($p > 0.1$). Given the knowledge that the most used EEG frequency bands are δ, θ, α, β, and low γ, we extracted these bands from the EEG to obtain the third representation (DR3).

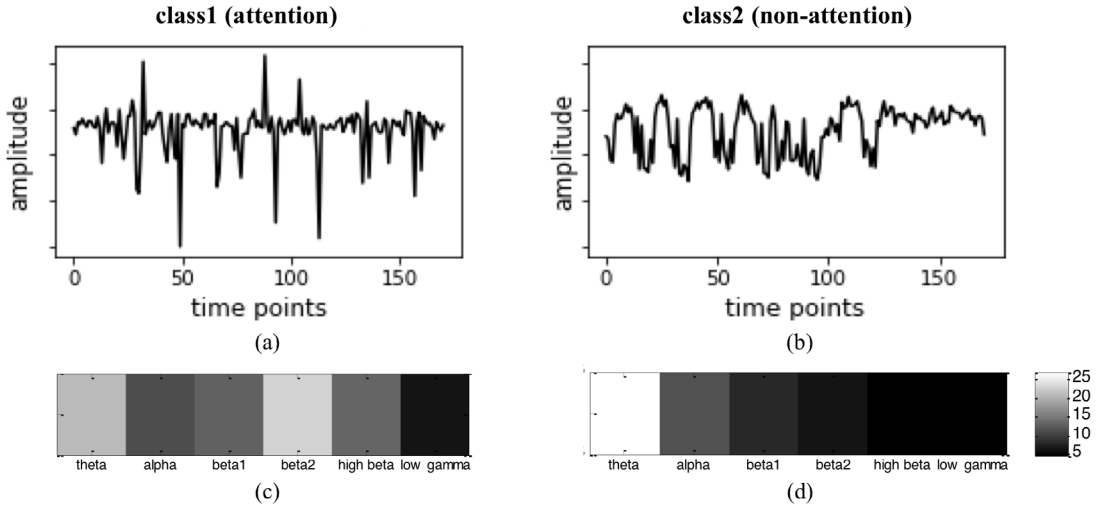


Figure 5. The visualization results. The plot in (a) is the network perception of class 1 (attention) and the plot in (b) is the network perception of class 2 (non-attention). The attention class shows high-frequency oscillations while these components have disappeared in the non-attention pattern. The PSD of signals in (a) and (b), over several frequency bands, including alpha, beta1, beta2, high beta, and low gamma, is demonstrated in (c) and (d), respectively. As can be seen, beta, especially beta 2, has higher activity and theta has lower activity in the attention class than in the non-attention class. These observations are validation that the attentional information the network has learned is meaningful.

Using DR3 as the input produces slightly better results than DR1 (+0.48% in LOO and +0.60% in the adaptive method) which are not statistically significant ($p > 0.1$). Based on the impact of data representation on classification performance, we can infer that the deep CNN classification framework is capable of efficiently differentiating between attentive mental classes by learning from raw EEG data. It meaningfully removes the data preparation burden and sheds light on the utility of the raw EEG time series for classification tasks.

4.2. Inter-subject transfer learning

Transferring knowledge from one subject to another deteriorates the classification accuracy. For this reason, most of the studies usually perform intra-subject classification. However, due to time-consuming calibration and re-training sessions, it has always been a priority for BCI systems to transfer the knowledge learned from multiple subjects to the new target subject. In this study, we put forward a framework with inter-subject transfer learning techniques. It achieved an accuracy above 70% for 84.17% of the subjects, while the baseline methods with inter-subject transfer learning could hardly reach 70% (see table 1). Table 3 represents the confusion matrix in which class 1 and class 2 refer to the attentive and non-attentive mental states, respectively. For all data representations, the CNN with subject adaptation demonstrated less confusion between non-attentive and attentive mental states than LOO. This is indeed important when it comes to the application of EEG in diagnosis. Based on the average classification accuracy and confusion matrix, it can be seen that the adaptive technique demonstrates a better performance. This indicates that unlike LOO, with naive knowledge transfer that faces the problem of information shift/change, the adaptive method efficiently conquers this problem without losing the time optimality.

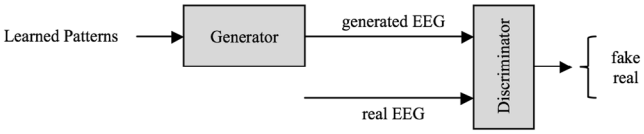


Figure 6. Training a GAN to generate EEG from the signals learned by a deep CNN.

To evaluate the performance of the proposed framework in the subjects with poor performance at baseline, we consider a threshold accuracy at 70% [44, 45]. A total number of 61 subjects out of 120 had accuracy below the threshold at baseline 2 (the better baseline). The proposed end-to-end framework results in a dramatic increase of 10.84% and 15.09% in the average classification accuracy of this group by the LOO and the adaptive method, respectively. Also, the CNN with LOO and the adaptive method decrease the size of this population from 50.84% to only 26.67% and 15.83%, respectively (see table 1). Notice that only the results of DR1 (raw EEG) are mentioned here. Figure 4 shows how the end-to-end deep CNN method enhances the detection accuracy for those 61 subjects with accuracy below 70% at baseline. As can be seen, the classification accuracy for 58 subjects out of 61 has been boosted, which means a 95.08% improvement.

4.3. The learned patterns are interpretable

In addition to quantitative analysis, it is important to obtain an understanding of what the network has learned from the input EEG data. Inspired by visualization techniques in image processing, we used a back-propagation-based method to gain an insight into the network learning. To do so, we performed an activation maximization technique to visualize the perceived input from the network [48]. In this method, we look for an input pattern that maximizes the activation of class c. In other

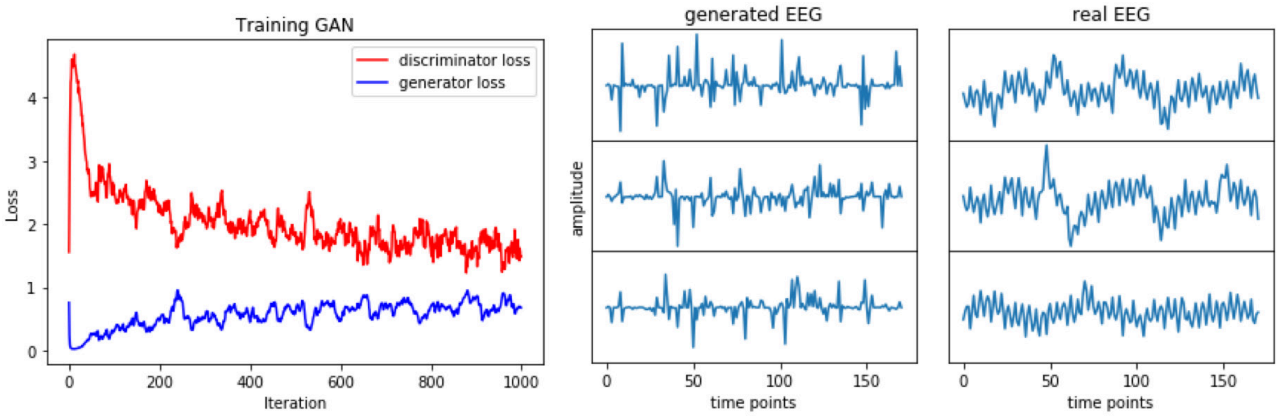


Figure 7. Results of GAN training; (a) the generator and discriminator losses over iterations, and (b) a few samples of generated and real EEG.

words, we solve the below optimization problem by means of a back-propagation technique:

$$x^* = \text{argmax}_x(a_c(x, \varphi) - R_\theta(x)). \quad (1)$$

Where a_c is the activation of the input signal x with the network parameters φ, $R_\theta(x)$ is the regularization term with parameters θ, and x^* is the desired input pattern. In fact, x^* is an input that, when fed to the network, results in class c. That is to say, this perceived input is what the network recognizes as class c. We used LP-norm (in our case, $p = 6$) as the regularization function.

The perception of the network from each class is plotted in figure 5. The aim is to understand what the network learns from neural data (EEG) and whether the learned information is meaningful. Interestingly, we observed that the network constructed a perceived input which has similar manifold to the original data. The patterns the network has learned from raw data (DR1) for attentive and non-attentive states are easy to distinguish. The attention class encompasses high-frequency components while the non-attention class shows low-frequency oscillations in its pattern. For further investigation, power spectral density (PSD) of these perceived inputs is computed using the Burg algorithm. Figures 5(c) and (d) demonstrate the PSD over the most common frequency bands, namely theta (4–8 Hz), alpha (8–12 Hz), beta1 (12–16 Hz), beta2 (16–20 Hz), high beta (20–30 Hz), and low gamma (30–40 Hz). Interestingly, we observed that with a change in mental state from non-attentive (class 2) to attentive (class 1):

(1) Beta activity increases.
(2) This excess in the beta band is bolder in beta2.
(3) Theta activity diminishes.
(4) The TBR, which has been known as an attention indicator, decreases. This can be inferred from observations 1 to 3.

These observations are consistent with the results of studies on attention-induced frequency oscillations [17, 21]. In our previous study (on a different dataset), we applied mutual information-based feature selection to discover the most discriminative attention-representative features [24]. Eventually, we found out that beta power and TBR are the most informative attributes for attention detection while theta power is not discriminative by itself [24]. Here, as a result of visualization, we ended up with similar observations but without any effort for feature extraction and selection. These findings suggest that the proposed network can successfully learn meaningful information from raw EEG data. It should be mentioned that EEG decomposition into frequency bands might affect the morphology of the signal, cause loss of information, and form misleading information. By learning directly from raw EEG, the end-to-end CNN is capable of automatically detecting the important frequency bands in attention detection without encountering the problems associated with EEG decomposition and feature extraction.

To further investigate whether the learned signals lie on the manifold of real EEG signals, we applied the method of generative adversarial networks (GAN) to generate EEG from these learned signals instead of noise. The overall framework is presented in figure 6.

After successful application of GANs in image generation [49], it has also recently been used in a few studies on time series data [50]. An interesting direction for the use of GANs in EEG is to generate naturalistic EEG signals. This EEG generation has the potential to be used in a range of generative applications, such as restoration of corrupted EEG segments and EEG augmentation for BCI tasks. Here, we used GAN to further analyse the learned signals obtained from the activation maximization technique. We hypothesized that if the discriminator fails in recognizing the fake EEG, this would be further evidence for the similarity between the learned signals' manifold and the manifold of the EEG.

The generator network consisted of three transposed convolution layers, each followed by batch normalization. The discriminator network has two convolution layers, each layer similarly followed by batch normalization. We used leaky ReLU activation and ADAM optimizer methods in both the generator and discriminator networks. Figure 7 shows the preliminary results; (a) the generator and discriminator losses

over iterations, and (b) a few samples of generated and real EEG. The outputs suggest that it is feasible to generate EEG from the learned signals by training a GAN.

5. Conclusion

The emergence of DL techniques has greatly enhanced classification tasks in several areas, such as speech and vision. In recent years, these networks have also found meaningful applications in BCI systems. Huge amounts of EEG time series can be fed into DNNs for classification tasks. EEG classification methods are prone to a notable drop in classification accuracy due to (1) loss of information and (2) transferring the knowledge inter-subjects. When it comes to DL frameworks, another challenge arises; (3) the interpretation of what the network learns. To address the three challenges listed, we proposed a deep CNN framework for the classification of EEG into attentive/non-attentive mental states with applications in cognitive BCI, game-based BCI, and neuro-rehabilitation.

This technique avoids the loss of information by learning from raw EEG (addressing problem 1). The combination of convolutional, max-pooling, and dropout layers builds a network that outputs a significantly higher accuracy than conventional feature extraction and classification techniques. Furthermore, this framework greatly reduced the percentage of subjects with accuracy less than 70% (as a threshold for BCI). We investigated the performance of the network by importing two other EEG representations into the deep CNN and comparing the results with those from raw EEG. No statistically significant improvement was found in the average accuracies. This means that the proposed classification framework does not benefit from the processed EEG representations and, except for artefact removal, any further processing is redundant.

Unlike baseline methods, the end-to-end deep CNN framework does not suffer from transferring the learned knowledge to a new subject (addressing problem 2). We implemented inter-subject transfer learning methodologies (LOO and subject adaptation) by training a generalized model for a pool of subjects and transferring the knowledge to the new subject or adapting the trained model based on a small amount of the new subject's data. This strategy is beneficial in the implementation of real time BCI systems. The results also showed that the adaptive technique outperforms the LOO technique. In particular, in subjects with relatively lower accuracy, adaptation helps the network to learn more optimal patterns for attention detection.

The visualizations verify that the learned attentive/non-attentive patterns from raw EEG data are discriminative and meaningful; the presence of high-frequency elements can be seen in the attention class but not in the non-attention class (addressing problem 3). When the brain is involved in attentional tasks, the EEG has higher activity in the beta band, especially in beta2, and lower activity in the theta band. In other words, the network, without being directly trained on these features, will recognize that decreased theta power, increased beta power, and decreased TBR are indicators of attentive mental state.

Furthermore, the sufficient number of samples and regularization techniques, such as dropout, guarantee that our network does not face the over-fitting problem. Another advantage of this work is that, unlike many other methods in which the input preparation stage is independent from the classification network, the proposed algorithm is an end-to-end unified framework.

One limitation of this study is that for the adaptive method we used part of a new subject's samples to adapt the trained model. This means that compared to LOO, the size of the training set is slightly larger (i.e. 0.4% larger, if we suppose all subjects have an equal number of samples). This might be a possible reason for the improved performance. Implementation of the adaptive method in a semi-supervised manner would effectively address this problem and is worth further investigation. Another caveat is the lack of automatic optimal parameter selection. This can potentially be addressed by using hyper-parameter optimization algorithms.

Overall, this study indicates that DL by means of a CNN is a promising classification technique for EEG which outperforms other techniques such as LDA, SVM, and FBCSP. The observations suggest that by employing a deep CNN, it is possible to learn from raw EEG and successfully transfer the learned knowledge to a new target subject. The presented work can be applied to attention-based BCI systems and extended to other types of EEG-based BCIs.

Acknowledgments

The authors would like to thank the I2R-ASTAR and Duke-NUS members for data acquisition. Also, the participation of elderly subjects and their caregivers is greatly appreciated.

ORCID iDs

Fatemeh Fahimi https://orcid.org/0000-0002-8516-7285

References

[1] Hebb D O 1949 *The Organization of Behavior: a Neuropsychological Theory* (London: Psychology Press)
[2] Hinton G E, Osindero S and Teh Y W 2006 A fast learning algorithm for deep belief nets *Neural Comput.* **18** 1527–54
[3] Lipton Z C, Berkowitz J and Elkan C 2015 A critical review of recurrent neural networks for sequence learning (arXiv:1506.00019)
[4] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *Proc. 25th Int. Conf. on Neural Information Processing Systems* (Lake Tahoe, NV: Curran Associates Inc.) pp 1097–105
[5] Deng J, Dong W, Socher R, Li L J, Kai L and Li F-F 2009 ImageNet: a large-scale hierarchical image database *2009 IEEE Conf. on Computer Vision and Pattern Recognition* pp 248–55
[6] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436
[7] 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
[8] Lu N, Li T, Ren X and Miao H 2017 A Deep learning scheme for motor imagery classification based on restricted boltzmann machines *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 566–76
[9] Wolpaw J and Wolpaw E W 2012 *Brain–Computer Interfaces: Principles and Practice* (Oxford: Oxford University Press)
[10] Nicolas-Alonso L F and Gomez-Gil J 2012 Brain computer interfaces, a review *Sensors* **12** 1211–79
[11] Lee T-S *et al* 2013 A brain–computer interface based cognitive training system for healthy elderly: a randomized control pilot study for usability and preliminary efficacy *PLoS One* **8** e79419
[12] Perego P, Turconi A C, Andreoni G, Maggi L, Beretta E, Parini S and Gagliardi C 2011 Cognitive ability assessment by brain–computer interface: validation of a new assessment method for cognitive abilities *J. Neurosci. Methods* **201** 239–50
[13] Jiang Y, Abiri R and Zhao X 2017 Tuning up the old brain with new tricks: attention training via neurofeedback *Frontiers Aging Neurosci.* **9** 52
[14] Lim C G, Lee T S, Guan C, Fung D S S, Zhao Y, Teng S S W, Zhang H and Krishnan K R R 2012 A brain–computer interface based attention training program for treating attention deficit hyperactivity disorder *PLoS One* **7** e46692
[15] Fahimi F, Zhang Z, Lee T S and Guan C 2018 Deep convolutional neural network for the detection of attentive mental state in elderly *The Seventh Int. BCI Meeting (Alisomar, USA)*
[16] MacLean M H, Arnell K M and Cote K A 2012 Resting EEG in alpha and beta bands predicts individual differences in attentional blink magnitude *Brain Cogn.* **78** 218–29
[17] Kamiński J, Brzezicka A, Gola M and Wróbel A 2012 Beta band oscillations engagement in human alertness process *Int. J. Psychophysiol.* **85** 125–8
[18] Klimesch W 2012 Alpha-band oscillations, attention, and controlled access to stored information *Trends Cogn. Sci.* **16** 606–17
[19] Hanslmayr S, Gross J, Klimesch W and Shapiro K L 2011 The role of alpha oscillations in temporal attention *Brain Res. Rev.* **67** 331–43
[20] Klimesch W, Sauseng P and Hanslmayr S 2007 EEG alpha oscillations: the inhibition-timing hypothesis *Brain Res. Rev.* **53** 63–88
[21] Martijn A, Conners C K and Helena C K 2012 A decade of EEG theta/beta ratio research in ADHD: a meta-analysis *J. Attention Disord.* **17** 374–83
[22] Liu N-H, Chiang C-Y and Chu H-C 2013 Recognizing the degree of human attention using EEG signals from mobile sensors *Sensors* **13** 10273–86
[23] Hamadicharef B, Zhang H, Guan C, Chuanchu W, Phua K S, Tee K P and Ang K K 2009 Learning EEG-based spectral-spatial patterns for attention level measurement *2009 IEEE Int. Symp. on Circuits and Systems* pp 1465–8
[24] Fahimi F, Guan C, Ang K K, Goh W B and Lee T S 2017 Personalized features for attention detection in children with attention deficit hyperactivity disorder *IEEE Engineering in Medicine and Biology Society* (*Jeju Island, South Korea*) pp 414–7
[25] Shenoy P, Krauledat M, Blankertz B, Rao R P and Muller K R 2006 Towards adaptive classification for BCI *J. Neural Eng.* **3** R13–23
[26] Tabar Y R and Halici U 2017 A novel deep learning approach for classification of EEG motor imagery signals *J. Neural Eng.* **14** 016003
[27] Jirayucharoensak S, Pan-Ngum S and Israsena P 2014 EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation *Sci. World J.* **2014** 10
[28] Sakhavi S, Guan C and Yan S 2018 Learning temporal information for brain–computer interface using convolutional neural networks *IEEE Trans. Neural Netw. Learn. Syst.* **29** 5619–29
[29] Zhang J and Li S 2017 A deep learning scheme for mental workload classification based on restricted Boltzmann machines *Cogn. Technol. Work* **19** 607–31
[30] Pouya Bashivan I R, Yeasin M and Codella N 2016 Learning representations from EEG with deep recurrent-convolutional neural networks (arXiv:1511.06448v3)
[31] Ma T, Li H, Yang H, Lv X, Li P, Liu T, Yao D and Xu P 2017 The extraction of motion-onset VEP BCI features based on deep learning and compressed sensing *J. Neurosci. Methods* **275** 80–92
[32] Sturm I, Lapuschkin S, Samek W and Müller K-R 2016 Interpretable deep neural networks for single-trial EEG classification *J. Neurosci. Methods* **274** 141–5
[33] MacLeod C M 1991 Half a century of research on the Stroop effect: an integrative review *Psychol. Bull.* **109** 163–203
[34] MacLeod C M and MacDonald P A 2000 Interdimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention *Trends Cogn. Sci.* **4** 383–91
[35] Stroop J R 1935 Studies of interference in serial verbal reactions *J. Exp. Psychol.* **18** 643–62
[36] Dyer F N 1973 The Stroop phenomenon and its use in the stlldy of perceptual, cognitive, and response processes *Mem. Cogn.* **1** 106–20
[37] Marie T B 2009 Executive function: the search for an integrated account *Curr. Dir. Psychol. Sci.* **18** 89–94
[38] Molina-Cantero A, Guerrero-Cubero J, Gómez-González I M, Merino-Monge M and Silva-Silva J I 2017 Characterizing computer access using a one-channel EEG wireless sensor *Sensors* **17** 1525
[39] Malmivuo J and Plonsey R 1995 Bioelectromagnetism *Electroencephalography* (New York: Oxford University Press) p 13
[40] Lecun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324
[41] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *32nd Int. Conf. on Machine Learning (Lille, France)*
[42] 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 (http://jmlr.org/papers/v15/srivastava14a.html)
[43] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *3rd Int. Conf. for Learning Representations (San Diego, USA)*
[44] Kübler A, Neumann N, Wilhelm B, Hinterberger T and Birbaumer N 2004 Predictability of brain–computer communication *J. Psychophysiol.* **18** 121–9
[45] Vidaurre C and Blankertz B 2010 Towards a cure for BCI illiteracy *Brain Topogr.* **23** 194–8
[46] Treder M S, Bahramisharif A, Schmidt N M, van Gerven M A and Blankertz B 2011 Brain–computer interfacing using modulations of alpha activity induced by covert shifts of attention *J. NeuroEng. Rehabil.* **8** 24
[47] Ang K K, Chin Z Y, Wang C, Guan C and Zhang H 2012 Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b *Frontiers Neurosci.* **6** 39
[48] Erhan D, Bengio Y, Courville A and Vincent P 2009 Visualizing higher-layer features of a deep network *Technical Report* University of Montreal
[49] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Proc. 27th Int. Conf. on Neural Information Processing Systems (Montreal, Canada)*
[50] Corley I A and Huang Y 2018 Deep EEG super-resolution: upsampling EEG spatial resolution with generative adversarial networks *2018 IEEE EMBS Int. Conf. on Biomedical & Health Informatics (BHI)* pp 100–3