Inter-subject Transfer Learning with 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

Abstract

***Objective*:** Despite the effective application of deep learning 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 deep convolutional neural network (CNN) to detect 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 EEG time-series. We also explore the consequence 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-trainings, we perform inter-subject transfer learning techniques as classification strategy. Eventually, to interpret the learned attentional patterns, we visualize and analyze the network perception of attention and non-attention classes. ***Main results*:** The average classification accuracy is 79.26% with only 15.83% of 120 subjects having the accuracy below 70% (a generally accepted threshold for BCI). This is while with inter-subject approach, it is literally hard to output high classification accuracy. This end-to-end classification framework surpasses the conventional classification methods for attention detection. The visualization results validate that the learned patterns from raw data are meaningful. ***Significance*:** This framework significantly improves the attention detection accuracy with inter-subject classification. Moreover, this study sheds light into the research on end-to-end learning; the proposed network is capable to learn 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

1. Introduction

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

 © xxxx IOP Publishing Ltd

types of deep neural networks include deep belief nets [2], recurrent neural networks [3], and convolutional neural networks (CNN). By achieving notable success in ImageNet challenge [4, 5], deep CNN has become the centre of attention. In this paper, we have also built our methodology on CNN.

Deep learning first found successful applications in the fields of speech recognition and computer vision [6] and then gained attraction in other research areas like 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 at assessment and enhancement of cognitive functions such as attention [11-14]. In these kinds of BCI systems where the subject's attention level serves as a control signal, it is crucial to precisely detect attentive mental state from 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. Plenty of 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 like beta is an indicator of attentional arousal. Decreased theta beta ratio, 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 filter bank (FB) and common spatial pattern (CSP), they extracted spectral-spatial features from EEG which was recorded using multiple electrodes placed in various brain regions. Then, the extracted features were sent to a fisher linear discriminant (FLD) classifier for classification task [23]. Their approach outperformed the conventional methods based on only spectral features. In case of 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 (MI)-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 been always a big challenge [25]. Deep convolutional neural networks with their ability in handling high-volume datasets, better learning algorithms and faster computational resources are becoming a superior alternative for 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. Rezaei Tabar and his colleagues boosted the classification accuracy of motor imagery (MI) BCI by proposing a deep network composed of CNN and stacked auto-encoders (SAE). In their work, 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 on 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 deep learning network [27]. They extracted principal components of power spectral densities from 32 EEG channel 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 envelop representation of EEG using Hilbert transformation and passing it through a CNN [28]. Their data representation was inspired by filter-bank common spatial pattern (FBCSP). They claim that by applying their algorithm on 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 deep learning network based on restricted Boltzmann machine (RBM) for MI classification. They named it frequential deep belief network (FDBN). In FDBN, frequency representation of EEG, generated using fast furrier transform (FFT) and wavelet decomposition techniques, passes through three RBMs and an extra output layer for classification [8]. Zhang and Li also employed RBM to develop a deep learning 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 input layer and the first hidden layer. Another study used recurrent-convolutional neural network 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 suggest that such representation of data preserves temporal, spectral and spatial information [30].

Ma et al. targeted at learning discriminative motion-onset visual evoked potentials (mVEP) features by using a combination of multi-level compressed sensing and RBM [31]. They report that deep features, obtained from this method, perform better than conventional amplitude-based features. They used support vector machine for classification [31]. The 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.

In order to provide an insight into the neurophysiological phenomena affect the decision of deep neural network (DNN), Strum and colleagues put forward the idea of using layer-wise relevance propagation (LRP) with 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 term of classification accuracy, their methodology did not outperform the common spatial pattern 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 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) interpretability of what 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 new subject's samples. In this way, the problems of time-consuming re-trainings 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 been also applied on a multi-channel dataset to investigate the performance and generalizability of the method. The results suggest that the end-to-end framework is promising for multi-electrode setting as well.

The rest of this paper is organized as follow: Section 2 describes the data and recording protocol. It then continues with presenting the proposed methodology including pre-processing, data representations and 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 in clinicaltrials.gov. Note that this study is not a clinical trial and does not report on clinical outcomes, it only uses the EEG data.

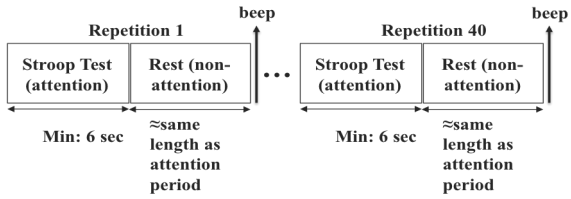
A total number of 120 healthy subjects performed the Stroop color test which is a well-known task to study attention [33, 34]. It can be traced back to John Ridley Stroop who reported the Stroop effect in his work in 1935 [35]. Then, it gained great attraction in the fields of cognitive sciences and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

During test, a colored word was presented on a screen and subjects were asked to name the color in which word is written. In fact, subjects were experiencing a conflict of information; what the word says and what is the color of the word. Thus, subjects needed to obtain and maintain their attention during Stroop color task [37]. During each session, participants performed 40 repetitions of 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 minutes. Figure 1 shows the recording protocol and an example of 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 256Hz. There are strong evidences from several studies which prove the efficiency of frontal EEG channel in studying attention-related tasks [11, 14, 22, 24, 38].

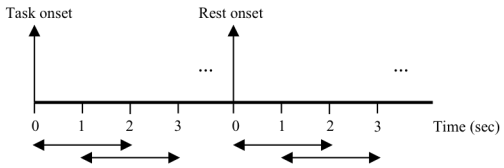
2.2 Pre-processing

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



(a) Recording protocol; Stroop test followed by a rest period.

(b) An example of test demonstration



(c) Segmentation of EEG with reference to question onset.

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

filtered EEG above 0.5Hz 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 raw EEG with minimal amount of processing as input. In fact, we avoid any pre-feature extraction and/or transferring EEG into image which are the main sources of information loss and computational costs. Based on the single-channel EEG data which provide a 1D input for the network, we defined 3 input representations for the network without pre-extracted features. In all representations, the segments were down-sampled by 3 with respect to the original value of 256Hz, resulting in 171 time points for 2s 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 5 classical bands; δ (0.5-4 Hz), θ (4-8Hz), 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 in order to remove artefacts.

2.3.2 Network Architecture

The early convolutional neural network (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 of some of these techniques.

In convolutional layers, the filter (kernel) convolves over input and produces element-wise multiplications. These numbers will be summed up and 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 convolutional layer. Using the subsequent pooling layer targets at reducing the dimension of 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, the high level feature maps will be generated. For classification tasks, the last layer of 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 input. Since the input data are time series, 1D filter has been used across time for convolution. The effectiveness of using 1D filter across time even for 2D inputs has been proven in the literature [26, 28]. To generate high level features, we inserted three convolutional layers with 1D filter for the network. The first layer with 60 filters and kernel size 1×4 is followed by a max-pooling layer with pool size 1×2. The output of 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 kernel size 1×2, the generated feature maps are flattened into a vector. This vector then passes through a dropout layer with the probability of 20% before being fed into the first fully connected layer of size 100. Then, we inserted the second dropout layer with the 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 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 the schematic diagram of end-to-end deep CNN-based

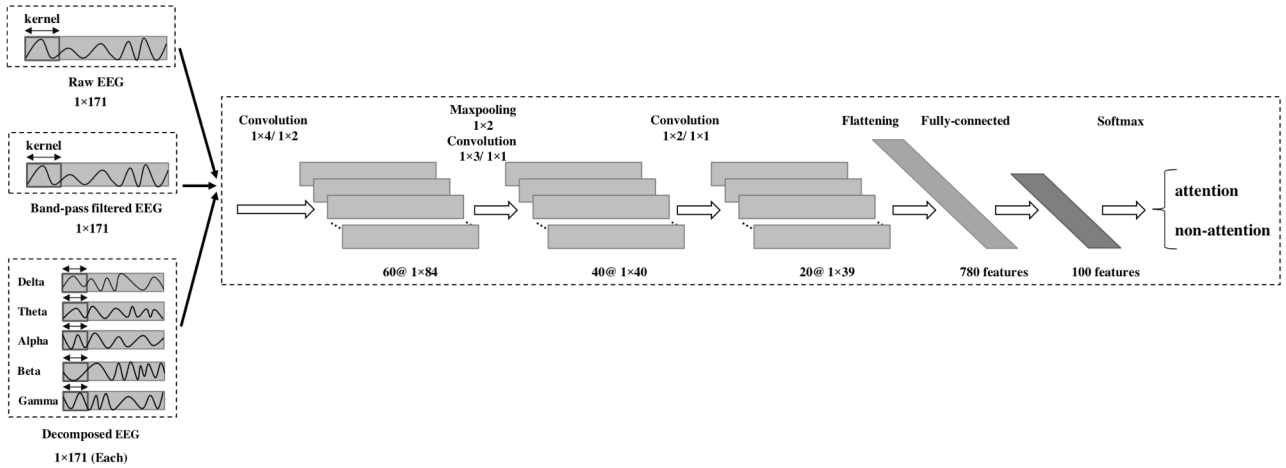


Figure 2. Schematic diagram of end-to-end CNN-based classification framework for transfer learning. The first tuple under convolution refers to kernel size and the second tuple shows the stride. After learning process, the learned features will be classified by softmax. The left boxes are respectively associated with data representation 1, 2, and 3. Note that in the case of data representation 3, the input EEG fed into the network is stored in 5 frequency channels but the network structure remains the same.

classification framework for inter-subject transfer learning in BCI.

3. Results

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

3.1 Baseline

In order to provide a fair baseline for the proposed technique, we implemented the classification framework as introduced in [22] for the single-channel data to classify between attention and non-attention. 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 (0.5-3Hz), theta (4-7Hz), alpha (8-13Hz), beta (14-30Hz) and alpha beta ratio were extracted using fast furrier transform (FFT) and sent into support vector machine (SVM) with polynomial kernel function for classification.

As second baseline, we band-pass filtered the data in 5 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 squared values) and sent into LDA for classification. Note that unlike [22] which used k-fold cross validation, we performed inter-subject classification approach (leave-one subject-out) in both baseline methods to provide fair comparison with the results of deep CNN. The 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%. Table 1 left side summarizes the baseline results. As can be seen, baseline1 and 2 respectively led to the 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 accuracy below 70% (as 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 [46]. In this study, the neurosky device was used for EEG recording. This device generates the attention indicator and some other information such as frequency band powers. The authors simply used the attention indicator obtained from the device to detect attentive state using LDA classifier. Initially, 10 subjects were involved in the experiment but 4 of them failed to control their attention level (based on the attention indicator) and were excluded. Thus, the classification was done on the small population of 6 subjects. The average accuracy is 79.5% based on table 7 in their paper. The main limitation of their work, beside small sample size, is the way of classification that is done for each subject on each session separately and then averaged over sessions. This simplified way of classification (within subject and within session) will be certainly deteriorated by subject-to-subject and session-to-session variations. They also reported that including frequency band 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 leave-one subject-out

In leave-one subject-out (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 retraining 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 target subject and transferred the information to the target subject. Execution of this method led to the 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 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 total 120 subjects for DR1, DR2 and DR3 respectively.

3.3 Deep CNN with subject Adaptation

Although zero-shot learning method evades long time trainings for 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 retraining is done on a small sample size of new subject's data. In this way, the problems of excessive re-training time and information shift can be both addressed.

In this study, we used half of new subject's samples for adaptation (2-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, 11.02% increase compared to baseline (p_value<0.0001) and 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 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, CNN with subject

Table 1. Average Accuracy for Baseline and Proposed Methods. Std Refers to 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%

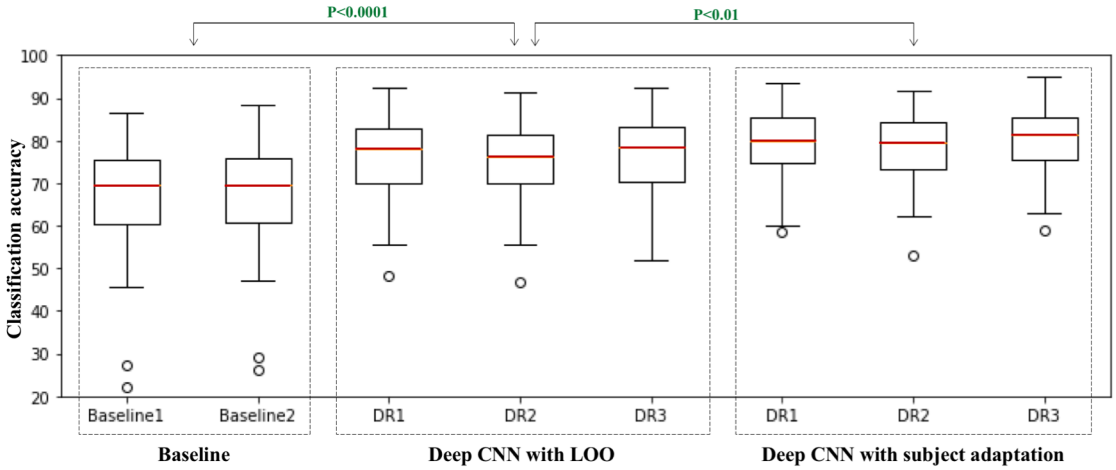


Figure 3. Comparing the performance of baseline and end-to-end deep CNN methods in attention detection. Classification framework based on deep CNN (both strategies; LOO and subject adaptation) significantly outperforms the baseline methods. Note that there is no statistically significant difference between methods in each group (baseline, Deep CNN with LOO, and Deep CNN with subject sdaptation). P_values are calculated using Wilcoxon test. The circles are the outliers; subjects with smaller accuracy than the lower extreme.

adaptation technique achieves the best performance. Although there is a statistically significant difference between CNN methods (LOO and subject adaptation), there is no significant difference between data representations within each method.

3.4 Results on a multi-channel public dataset

In order 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 [47]. A total number of 8 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 international 10-10 system. The sampling frequency during recording was set at 1000Hz which later it was down-sampled to 200Hz. The experiment includes the sequences of attention, response, and rest. We have segmented the EEG during attention and rest parts for classification task. Based on the original study on this dataset [47], a subset of 9 electrodes including PO3, 4, 7-10, Oz, O1, and O2 are the optimal electrodes for studying attention. We have also used these 9 recommended electrodes in our study.

As the first baseline for multi-electrode dataset, we implemented the popular method of Filter Bank Common Spatial Pattern (FBCSP) [48]. The methods of Mutual Information-based Best Individual Feature (MIBIF) and Naïve Bayesian Parzen Window (NBPW) have been used for feature selection and classification respectively, the same as in [48]. Beside classification with LOO, which provides the results for fair comparison with 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 they call it shallow ConvNet, is inspired by the method of FBCSP. 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 method of FBCSP and end-to-end deep CNN outperforms both baseline methods. Comparing LOO results, the performance of the proposed method is significantly better than FBCSP (+18.31%, P_value<0.001) and shallow ConvNet (+6.28%, P_value<0.001). In fact, the results of LOO classification with end-to-end framework are as good as the results of intra-subject classification with FBCSP. This shows although FBCSP has 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 be potentially used to address this problem. The proposed end-to-end deep CNN decodes more than 70% of the EEG trials correctly for all the 8 subjects.

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 input to CNN. The objective was to learn from raw EEG; end-to-end study. The first representation (DR1) is raw EEG with the least amount of pre-processing (to remove artefacts). CNN with this representation as input outperforms the baseline ($p<0.0001$) with 8.14% (LOO) and 11.20% (Adaptive) improvement in average accuracy. Going one step further in data preparation, we band-pass filtered the data at 0.5-40Hz (DR2) and fed it into CNN for classification. Interestingly, the average classification accuracy dropped by 1.13% in LOO ($p>0.1$) and by 1.14% in adaptive ($p>0.1$). Given the knowledge that the most used EEG frequency bands are δ, θ, α, β and low γ, we extracted these bands from EEG to obtain the third representation (DR3). Using DR3 as input produces slightly better results than DR1 (+0.48% in LOO and +0.60% in adaptive) which are not statistically significant ($p>0.1$). Based on the impact of data representations in classification performance, we can infer that deep CNN classification framework is capable to efficiently differentiate between attentive mental classes by learning from raw EEG data. It meaningfully removes the data preparation burden and sheds

Table 2. Results on multi-channel dataset. Std Refers to Standard Deviation.

	FBCSP [48]		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%

light on the utility of 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's been always 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 class1 and class2 respectively refer to the attentive and non-attentive mental states. For all data representations, 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 adaptive technique has better performance. It 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.

To evaluate the performance of proposed framework in the case of 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 threshold at baseline 2 (the better baseline). The proposed end-to-end framework makes a dramatic increase of 10.84% and 15.09% in average classification accuracy of this group by LOO and adaptive method, respectively. Also, CNN with LOO and adaptive method decreases 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 end-to-end deep CNN method enhances the detection accuracy for those 61 subjects with accuracy below 70% at baseline. As it can be seen, the classification accuracy for 58 subjects out of 61 has been boosted which means 95.08% improvement.

4.3 The learned patterns are interpretable

Beside quantitative analysis, it is important to obtain an understanding of what the network has learned from the input EEG data. Inspired by the 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 activation maximization technique to visualize the perceived input from the network [49]. In this method, we look for an input pattern that maximizes the activation of class c. In other words, we solve the below optimization problem by means of back-propagation technique:

$$x^* = argmax_x(a_c(x, \varphi) - R_\theta(x)) \quad (1)$$

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 regularization function.

The perception of the network from each class is plotted in figure 5. The interest 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

Table 3. 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

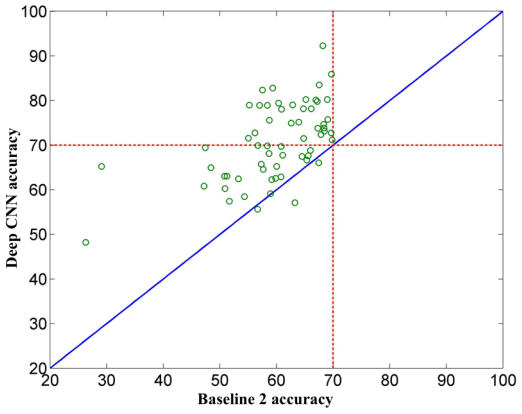


Figure 4 Classification accuracy for subjects with poor performance (<70%) at baseline. For simplicity in comparison only baseline2 and Deep CNN with LOO on DR1 are plotted. The end-to-end deep CNN framework dramatically increases the performance of these subjects by 10.84% increase in the average accuracy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

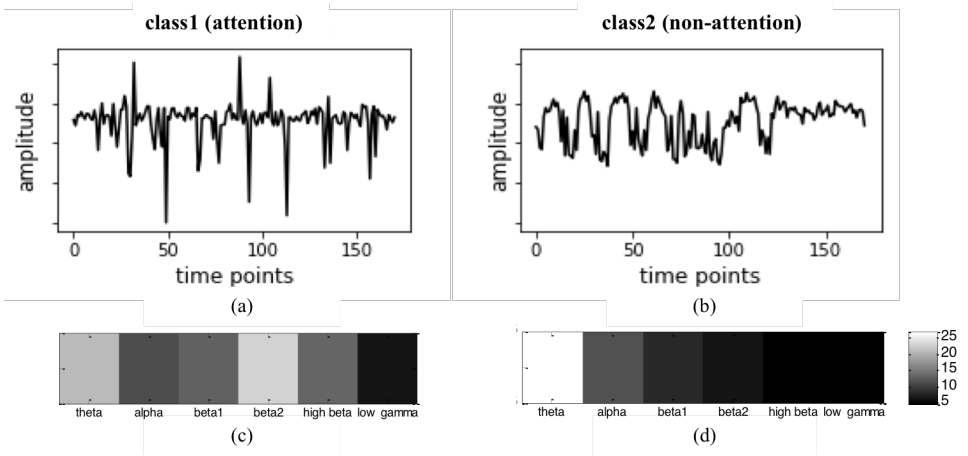


Figure 5. The visualization results. The plot in (a) is the network perception of class 1 (attention) and the plot in (b) is the network perception of class 2 (non-attention). Attention class shows high-frequency oscillations while these components are disappeared in non-attention pattern. Power spectral density 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 attention class than non-attention class. These observations validate that the attentional information the network has learned is meaningful.

low-frequency oscillations in its pattern. For further investigation, power spectral density (PSD) of these perceived inputs are computed using burg algorithm. Figure 5 (c and d) demonstrates the PSD over the most common frequency bands namely theta (4-8Hz), alpha (8-12Hz), beta1 (12-16Hz), beta2 (16-20Hz), high beta (20-30Hz), and low gamma (30-40Hz). Interestingly, we observed that with change in mental state from non-attentive (class2) to attentive (class1):

1) Beta activity increases.

2) This excess in beta band is bolder in beta2.

3) Theta activity diminishes.

4) Theta beta ratio (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 (MI)-based feature selection to discover the most discriminative attention-representative features [24]. Eventually, we found out that beta power and theta beta ratio 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 signal, cause loss of information, and form misleading information. By learning directly from raw EEG, the end-to-end CNN is capable to automatically detect 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 [50], it has recently been used in a few studies on time-series data as well [51]. An interesting direction for the use case of GANs in EEG is to generate naturalistic EEG signals. This EEG generation has 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 a further evidence for the similarity between learned signals' manifold and the manifold of EEG.

The generator network consisted of 3 transposed convolution layers, each followed by batch-normalization. The discriminator network has 2 convolution layers, similarly each layer followed by batch-normalization. We used Leaky ReLU activation and Adam optimizer methods in both 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.

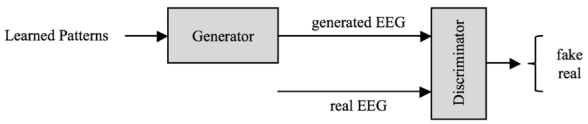
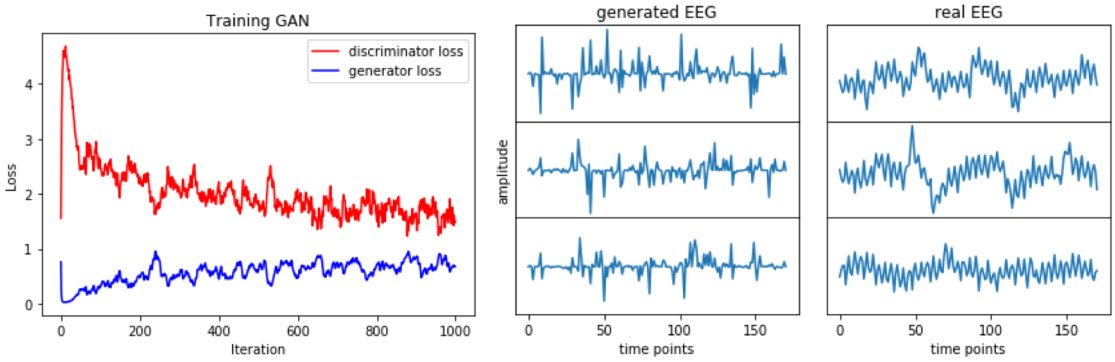


Figure 6 Training GAN to generate EEG from the signals learned by Deep CNN.



(a) **(b)**

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

5. Conclusion

The emergence of deep learning techniques has highly enhanced the classification tasks in several areas such as speech and vision. In recent years, these networks have found meaningful applications in BCI systems as well. Huge amount of EEG time series can be fed into deep neural networks 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 deep learning 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 the 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 the significantly higher accuracy than conventional feature extraction and classification techniques. Furthermore, this framework majorly lessened 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 the ones 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 (leave-one subject-out and subject adaptation) by training a generalized model for a pool of subjects and transferring the knowledge to the new subject or adapt the trained model based on the small amount of new subject's data. This strategy is beneficial in implementation of real time BCI systems. The results also showed that the adaptive technique outperforms the LOO technique. Especially in case of subjects with relatively lower accuracy, adaptation helps the network to learn more optimal pattern for the 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 brain is involved in attentional task, EEG has higher activity in beta band, especially in beta2, and lower activity in 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 theta beta ratio are the 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 a part of new subject's samples to adapt the trained model. This means that compared to LOO, the size of training set is slightly larger (i.e. 0.4% larger, if we suppose all subjects have equal number of samples). This might be a possible reason for the improved performance. Implementation of 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 be potentially addressed by using hyper-parameter optimization algorithms.

Overall, this study indicates that deep learning by means of CNN is a promising classification technique for EEG which outperforms other techniques like LDA, SVM, and FBCSP. The observations suggest that by employing 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 for attention-based BCI systems and extended to other types of EEG-based BCIs.

Acknowledgements

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

References

[1] Hebb D O 1949 The organization of behavior: A neuropsychological theory *Psychology press*

[2] Hinton G E, Osindero S and Teh Y W 2006 A fast learning algorithm for deep belief nets *Neural computation* **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. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems* (Lake Tahoe, Nevada: 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. In: *2009 IEEE Conference 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 *Human brain mapping* **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 Transactions on Neural Systems and Rehabilitation Engineering* **25** 566-76

[9] Wolpaw J and Wolpaw E W 2012 *Brain-Computer Interfaces: Principles and Practice*: 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, Goh S J A, Quek S Y, Phillips R, Guan C, Cheung Y B, Feng L, Teng S S W, Wang C C, Chin Z Y, Zhang H, Ng T P, Lee J, Keefe R and Krishnan K R R 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 *Journal of neuroscience 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 in Aging Neuroscience* **9**

[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. In: *The Seventh International 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 and cognition* **78** 218-29

[17] Kamiński J, Brzezicka A, Gola M and Wróbel A 2012 Beta band oscillations engagement in human alertness process *International Journal of Psychophysiology* **85** 125-8

[18] Klimesch W 2012 alpha-band oscillations, attention, and controlled access to stored information *Trends in cognitive sciences* **16** 606-17

[19] Hanslmayr S, Gross J, Klimesch W and Shapiro K L 2011 The role of alpha oscillations in temporal attention *Brain Research Reviews* **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 *Journal of Attention Disorders* **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 (Basel, Switzerland)* **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. In: *2009 IEEE International Symposium on Circuits and Systems,* pp 1465-8

[24] F. Fahimi, C. Guan, K. K. Ang, W. B. Goh and T. S. Lee 2017 Personalized features for attention detection in children with Attention Deficit Hyperactivity Disorder. In: *IEEE Eng Med Biol Soc,* (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 *Journal of neural engineering* **3** R13-23

[26] Tabar Y R and Halici U 2017 A novel deep learning approach for classification of EEG motor imagery signals *Journal of neural engineering* **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 *The Scientific World Journal* **2014** 10

[28] Sakhavi S, Guan C and Yan S 2018 Learning Temporal Information for Brain-Computer Interface Using Convolutional Neural Networks *IEEE Transactions on Neural Networks and Learning Systems* **PP** 1-11

[29] Zhang J and Li S 2017 A deep learning scheme for mental workload classification based on restricted Boltzmann machines *Cognition, Technology & Work* **19** 607-31

[30] Pouya Bashivan I R, Mohammed Yeasin, Noel Codella 2015 Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks *arXiv:1511.06448*

[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 *Journal of neuroscience 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 *Journal of neuroscience methods* **274** 141-5

[33] MacLeod C M 1991 Half a century of research on the Stroop effect: an integrative review *Psychological bulletin* **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 in cognitive sciences* **4** 383-91
[35] Stroop J R 1935 Studies of interference in serial verbal reactions *Journal of Experimental Psychology* **18** 643-62
[36] Dyer F N 1973 The Stroop phenomenon and its use in the stlldy of perceptual, cognitive, and response processes *Memory & Cognition* **1** 106-20
[37] Marie T B 2009 Executive Function: The Search for an Integrated Account *Current Directions in Psychological Science* **18** 89-94
[38] Molina-Cantero A J, 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 (Basel, Switzerland)* **17** 1525
[39] Malmivuo J and Plonsey R 1995 *Bioelectromagnetism. 13. Electroencephalography*
[40] Lecun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proceedings of the IEEE* **86** 2278-324
[41] Ioffe S and Szegedy C 2015 Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *32nd International Conference 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
[43] Diederik P. Kingma and Ba J 2015 Adam: A Method for Stochastic Optimization. In: *3rd International Conference for Learning Representations,* (San Diego, USA
[44] Kübler A, Neumann N, Wilhelm B, Hinterberger T and Birbaumer N 2004 Predictability of Brain-Computer Communication *Journal of Psychophysiology* **18** 121-9
[45] Vidaurre C and Blankertz B 2010 Towards a Cure for BCI Illiteracy *Brain Topography* **23** 194-8
[46] Molina-Cantero A, Guerrero-Cubero J, Gómez-González I, Merino-Monge M and Silva-Silva J 2017 Characterizing Computer Access Using a One-Channel EEG Wireless Sensor *Sensors* **17** 1525
[47] 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 *Journal of NeuroEngineering and Rehabilitation* **8** 24
[48] 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 in Neuroscience* **6**
[49] Erhan D, Bengio Y, Courville A and Vincent P 2009 Visualizing Higher-Layer Features of a Deep Network. (Technical Report, University of Montreal
[50] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative Adversarial Nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems,* (Montreal, Canada
[51] Corley I A and Huang Y 2018 Deep EEG super-resolution: Upsampling EEG spatial resolution with Generative Adversarial Networks. In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI),* pp 100-3