
Article:
Azab, A., Ahmadi, H., Mihaylova, L. orcid.org/0000-0001-5856-2223 et al. (1 more author) (2020) Dynamic time warping-based transfer learning for improving common spatial patterns in brain-computer interface. Journal of Neural Engineering, 17 (1). 016061. ISSN 1741-2560

https://doi.org/10.1088/1741-2552/ab64a0

Dynamic time warping-based transfer learning for improving common spatial patterns in brain-computer interface

Dynamic Time Warping-based Transfer Learning for Improving Common Spatial Patterns in Brain-computer Interface

Ahmed M. Azab[1], Hamed Ahmadi[2], Lyudmila Mihaylova[1], and Mahnaz Arvaneh[1]

[1]Department of Automatic Control and System Engineering, Sheffield University, UK
[2] School of Computer Science and Electronic Engineering, University of Essex, UK

E-mail: ammazab1@sheffield.ac.uk

Abstract.
Objective. Common spatial patterns (CSP) is a prominent feature extraction algorithm in motor imagery (MI)-based brain-computer interfaces (BCIs). However, CSP is computed using sample-based covariance-matrix estimation. Hence, its performance deteriorates if the number of training trials is small. To address this problem, this paper proposes a novel regularized covariance matrix estimation framework for CSP (i.e. DTW-RCSP) based on dynamic time warping (DTW) and transfer learning.Approach. The proposed framework combines the subject-specific covariance matrix ($\mathbf{\Sigma}_{\text{ss}}$) estimated using the few available trials from the new subject, with a novel DTW-based transferred covariance matrix ($\mathbf{\Sigma}_{\text{DTW}}$) estimated using previous subjects' trials. In the proposed $\mathbf{\Sigma}_{\text{DTW}}$, the available labelled trials from the previous subjects are temporally aligned to the average of the few available trials of the new subject from the same class using DTW. This alignment aims to reduce temporal variations and non-stationarities between previous subjects trials and the available few trials from the new subjects. Moreover, to tackle the problem of regularization parameter selection when only few trials are available for training, an online method is proposed, where the best regularization parameter is selected based on the confidence scores of the trained classifier on upcoming first few labelled testing trials. Main results. The proposed framework is evaluated on two datasets against two baseline algorithms. The obtained results reveal that DTW-RCSP significantly outperformed the baseline algorithms at various testing scenarios, particularly, when only a few trials are available for training. Significance. Impressively, our results show that successful BCI interactions could be achieved with a calibration session as small as only one trial per class.

Keywords: Brain-computer Interface, Transfer learning, Common spatial patterns, Calibration time, Dynamic time warping.

1. Introduction

Brain-computer interface (BCI) allows a direct communication to control an electronic device using a person's brain signals without any muscular means [1]. Electroencephalogram (EEG) is the most popular brain signals used in BCI as it is recorded non-invasively with high temporal resolution and low cost [2]. Users' mental states are identified and converted to control signals in EEG-based BCIs by classifying the features extracted from the recorded brain signals. In motor imagery (MI)-based BCIs, these features are commonly extracted using common spatial patterns (CSP) algorithm [3].

CSP mainly aims to reduce the high dimensionality of the multi-channel EEG signals by maximising the difference between variances of the two classes of EEG signals. The quality of CSP features can be affected by several issues, such as noise due to movement artifacts, and non-stationarity of EEG signals. Moreover, CSP is computed based on covariance matrix estimation. Thus, it is likely to overfit when few trials are available from the user to train the CSP-based BCI model [4, 5]. This issue leads to one of the main challenges that prevents BCI systems from being used in daily-basis applications which is the long calibration time. Calibration time is the time required to record sufficient number of labelled trials to train the CSP-based BCI model. Typically, the calibration time is 20-30 minutes for each single session. This long calibration time leaves BCI users mentally exhausted before starting the real interactions.

For using a BCI system in daily life-based applications, it must be accurate across sessions and subjects, and with the shortest possible calibration time. The aforementioned challenges could be tackled at different stages by improving either the BCI user training part [6, 7], or the signal processing part. Regarding the EEG signal processing part, developing accurate and more robust CSP-based algorithms which can be calibrated with the minimum possible training data is greatly desirable for MI-based BCIs [8, 9].

Transfer learning could be potentially used to reduce the calibration time of BCI systems while the loss in the accuracy is minimised. Using transfer learning approaches, shortage of labelled trials from the current user can be compensated by incorporating other sessions/subjects data in the learning process [10]. Transfer learning can be applied on different domains to improve MI-based BCIs. In raw EEG domain, previously proposed transfer learning algorithms are mostly based on either instance selection [11] or importance sampling [12]. Available transfer learning algorithms on feature domain try to enhance CSP by improving either the estimation method of covariance matrix [13, 14, 15] or the optimization function of CSP [9, 16, 5]. For classification domain, existing transfer learning algorithms use either domain adaptation techniques [17], ensemble learning of classifiers [18, 19], or classifier objective function modification [20].

To the best of our knowledge, none of these studies considered the temporal variations between EEG trials of a new subject and those of previous subjects to

reduce between-subjects non-stationarity during transfer learning. Moreover, most of the proposed algorithms in the feature domain require calculating multiple regularization parameters which is computationally expensive.

This paper proposes a novel transfer learning framework in raw EEG and feature domains, called DTW-based regularized CSP (DTW-RCSP). At first, in the raw EEG domain, we transform previous subjects' trials to be more similar to the target subject's few training trials using a novel alignment method in time domain based on DTW, and hence use these aligned trials to form the transferred covariance matrix. Then, in the feature domain, we propose a novel regularization between the subject-specific and the transferred covariance matrices to improve the CSP covariance matrix estimation. The output of our proposed DTW-RCSP framework is a new regularized CSP matrix which is a combination of the subject-specific covariance matrix and the transferred covariance matrix from other subjects. Finally, to address the issue of regularization parameter selection when very few training trials are available, we propose an online method based on the upcoming first few labelled testing trials, where some predefined regularization parameters are evaluated based on the confidence scores of the trained classifier.

The proposed DTW-RCSP framework is evaluated across different scenarios based on the available subject-specific training trials using two datasets. The proposed DTW-RCSP performance is compared against two state of the art algorithms, standard CSP and Composite CSP (CCSP) [13].

2. Methodology

This section presents our proposed transfer learning framework (DTW-RCSP) to improve the CSP features of EEG signals, when few trials from the target subject and a group of trials recorded previously from other subjects are available. First, we will give a brief description about transfer learning definition. A domain d is defined by its feature space $\mathbf{X}$ and its marginal probability distribution $P(\mathbf{X})$. Subsequently, for each domain, its task consists of label space y and objective classification function f. This classification function can be learnt using the available training trials to find the labels of the testing trials. Generally, two different domains might have different feature space, different marginal probability distributions or both. Similarly, two different tasks have either different label space, different classification function or both.

Definition: Given source domain d_s, source task t_s, target domain d_t, target task t_t, transfer learning aims to help improve the learning of the target classification function f_t in d_t using the knowledge in d_s and t_s, where $d_s \neq d_t$ or $t_s \neq t_t$. Where $d_s \neq d_t$ means $P_s(\mathbf{X}) \neq P_t(\mathbf{X})$ or/and $\mathbf{X}_s \neq \mathbf{X}_t$. Moreover, $t_s \neq t_t$ means $y_t \neq y_s$ and/or $P_s(y|\mathbf{X}) \neq P_t(y|\mathbf{X})$[10]. For more information about transfer learning and its application in BCI, the reader can refer to [21, 10].

In our proposed DTW-RCSP framework, the previously recorded EEG trials from other subjects and sessions are pooled together as one single session s, and referred to as the source domain. Subsequently, the source domain is presented as $d_s = (\mathbf{X}_s^i, y_s^i)_{i=1}^N$,

where $\mathbf{X}_s^i$ and $y_s^i \in \{-1, 1\}$ respectively denote the EEG instance matrix and the class label of the i^{th} trial, and N points to the trials number. In each trial $\mathbf{X}_s^i \subset R^{h \times V}$, h is the EEG samples contained in each trial and V is the channels number. Similarly, the set of labelled trials of the target subject, t, is denoted as $d_t = (\mathbf{X}_t^i, y_t^i)_{i=1}^M$, where M is the number of the available subject-specific trials.

2.1. Dynamic Time Warping-based Transfer Learning Regularized CSP Framework (DTW-RCSP)

To improve CSP covariance matrix estimation when few trials are available for training, regularization based transfer learning techniques could be used. Regularized CSP works by specifying a trade-off between the estimates obtained using few target subject-specific trials and informative estimates obtained using previously recorded trials from other subjects/sessions [22]. In our proposed DTW-RCSP framework, the average CSP covariance matrix $\mathbf{\Sigma}_{\mathrm{TLR_c}}$ for each class c is calculated as follows:

$$\mathbf{\Sigma}_{\mathrm{TLR_c}} = (1 - r)\mathbf{\Sigma}_{\mathrm{SS_c}} + r\mathbf{\Sigma}_{\mathrm{DTW_c}}, \tag{1}$$

where r is the regularization parameter ($0 \leq r \leq 1$). $\mathbf{\Sigma}_{\mathrm{DTW_c}}$ is the proposed DTW-based transferred average covariance matrix of the aligned trials of class c from other subjects which will be explained in 2.2. $\mathbf{\Sigma}_{\mathrm{SS_c}}$ is the average covariance matrix of the few subject-specific trials of class c from the target subject. $\mathbf{\Sigma}_{\mathrm{SS_c}}$ is calculated as

$$\mathbf{\Sigma}_{\mathrm{SS_c}} = \frac{1}{m_c} \sum_{i=1}^{m_c} \frac{\mathbf{X}_t^{i\top} \mathbf{X}_t^i}{\mathbf{tr}(\mathbf{X}_t^{i\top} \mathbf{X}_t^i)}, \tag{2}$$

where m_c is the number of trials per class c, $\top$ is the matrix transpose function, and $\mathbf{tr}$ is the trace function.

The regularization parameter r shrinks the subject-specific covariance matrix towards the DTW-based transferred covariance matrix to neutralize the possible estimation bias due to the availability of few training trials from the target subject. In fact, $\mathbf{\Sigma}_{\mathrm{DTW_c}}$ represents the information on how the covariance matrix for the considered intellectual condition should typically be. Finally the DTW-RCSP filters,$\mathbf{W}_{\textsc{dtw-rcsp}}$, are calculated by maximising the following objective function using joint diagonalization [3]:

$$\mathbf{W}_{\textsc{dtw-rcsp}} = \mathbf{arg}\ \mathbf{max}_{\mathbf{W}}\ \frac{\mathbf{W}\ \mathbf{\Sigma}_{\mathrm{TLR_1}} \mathbf{W}^\top}{\mathbf{W}(\mathbf{\Sigma}_{\mathrm{TLR_1}} + \mathbf{\Sigma}_{\mathrm{TLR_2}})\mathbf{W}^\top}. \tag{3}$$

From (1), the classical CSP can be considered as a special case of DTW-RCSP framework, when $r = 0$.

2.2. Estimation of the Dynamic Time Warping Transferred Covariance Matrix

DTW has been initially proposed as a solution of the time distortion issue between two time series in speech recognition problems in a non-linear fashion. DTW finds an optimal alignment between two given sequences under certain restrictions to compensate the

timing differences between them [23]. After-that, different research areas have applied DTW such as object recognition, motion analysis, and classification of time domain signals including EEG, and ECG [24, 25]. For EEG, DTW is used as a dissimilarity measure between two EEG segments after being optimally aligned. In our previous paper, DTW has been used to reduce subject-specific temporal variations between two EEG segments [26].

In this paper, DTW is used for the purpose of transfer learning. Unlike the previous EEG-based studies, the goal is to align a collection of EEG trials from other subjects or sessions to the average of the few available trials from the new target subject. Thus, to calculate $\mathbf{\Sigma}_{\mathrm{DTW_c}}$, the DTW-based transferred average covariance matrix, the following steps are taken.

First the average of the available few trials of the target subject from class c is computed as follows:

$$\bar{\mathbf{X}}_{t_c} = (1/m_c) \sum_{i=1}^{m_c} \mathbf{X}_t^i, \tag{4}$$

where $\bar{\mathbf{X}}_{t_c}$ and m_c respectively refer to the average and the total number of the target trials of class c.

Next, each trial from the source session gets aligned to the average of the few target trials from the same class, $\bar{\mathbf{X}}_{t_c}$, using DTW. To align $\mathbf{X}_s^i \subset R^{h\times V}$ to $\bar{\mathbf{X}}_{t_c} \subset R^{h\times V}$, we construct a distance matrix $\mathbf{D}^{h\times h}$, where $\mathbf{D}(a,b)$ is the Euclidean distance between the EEG signals of two time instances of a and b from $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$ respectively,

$$\mathbf{D}(a,b) = \sqrt{\sum_{v=1}^{V} (\mathbf{X}_s^i(a,v) - \bar{\mathbf{X}}_{t_c}(b,v))^2}. \tag{5}$$

Thereafter, the elements of $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{\mathbf{t_c}}$ are mapped through the matrix D by finding an optimum warping path, whereby the cumulative distance between the two above-mentioned EEG trials is minimised. Generally, a warping path, $\mathbf{P}$, defines a mapping between $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$, and its elements are presented as

$$\mathbf{P} = [p(1), .., p(k), ..., p(K)] \quad h \leq K < 2h-1 \tag{6}$$

where $p(k)$=$\mathbf{D}(a_k, b_k)$. a_k and b_k belong to $\{1, 2, ..., h\}$, and remap the time indices of $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$ respectively. A warping path requires to be subject to the following constraints:
1- Boundary conditions: $p(1)$=$\mathbf{D}(1,1)$ and $p(K)$=$\mathbf{D}(h,h)$. In other words, $a_1 = b_1 = 1$ and $a_K = b_K = h$.
2- Continuity and monotonicity: $0 \leq a_k - a_{k-1} \leq 1$ and $0 \leq b_k - b_{k-1} \leq 1$.
3-In addition to the above mentioned constraints, there are some other global constraints on the warping path. These constraints limit how far the warping path from the diagonal path, could be. Global constraints are generally applied to prevent pathological warpings, where a relatively small section from one time sequence being mapped to a relatively large section of another, and to calculate the DTW distance matrix slightly faster. The two most frequently used global constraints are the Sakoe-Chiba band [27] and the Itakura parallelogram [28].

Numerous warping paths can satisfy the above-mentioned conditions. However, we are interested in the optimum warping path, $\mathbf{P}^*$, with the shortest non-linear alignment between $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$, as follows [29, 30]

$$\mathbf{P}^* = \mathbf{arg}\min_{\mathbf{P}} \ (\frac{1}{K}\sqrt{\sum_{k=1}^{K} p(k)}). \tag{7}$$

To reduce the computational time, $\mathbf{P}^*$ is computed using dynamic programming to assess the following recurrence [25], where the cumulative distance $\boldsymbol{\gamma}(a,b)$ is defined as the distance between two time instances a and b from $\mathbf{X}_s^i$ and $\bar{\mathbf{X}}_{t_c}$, $\mathbf{D}(a,b)$, and the minimum of the cumulative distances of the adjacent elements:

$$\boldsymbol{\gamma}(a,b) = \mathbf{D}(a,b) + min[\boldsymbol{\gamma}(a-1,b-1), \boldsymbol{\gamma}(a-1,b), \boldsymbol{\gamma}(a,b-1)] \tag{8}$$

Given $\mathbf{P}^*$, $\mathbf{X}_s^i$ is aligned to $\bar{\mathbf{X}}_{t_c}$ as:

$$\mathbf{X}^i_{s_{aligned}} = \begin{bmatrix} \mathbf{X}_s^i(a_1^*,1) & \mathbf{X}_s^i(a_1^*,2) & \cdots & \mathbf{X}_s^i(a_1^*,V) \\ \mathbf{X}_s^i(a_2^*,1) & \mathbf{X}_s^i(a_2^*,2) & \cdots & \mathbf{X}_s^i(a_2^*,V) \\ \vdots & & \ddots & \vdots \\ \mathbf{X}_s^i(a_K^*,1) & \mathbf{X}_s^i(a_K^*,2) & \cdots & \mathbf{X}_s^i(a_K^*,V) \end{bmatrix} \tag{9}$$

where $[a_1^*, a_2^*, ..., a_K^*]$ are the time indices of $\mathbf{X}_s^i$ forming the minimum warping path $\mathbf{P}^*$. These time instances are the instances that will make $\mathbf{X}_s^i$ to be as much similar to $\bar{\mathbf{X}}_{\mathbf{t_c}}$ as possible given the above constraints. Subsequently the covariance matrix of $\mathbf{X}^i_{s_{aligned}}$ is calculated as follows:

$$\boldsymbol{\Sigma}^i_{\mathrm{s_{aligned}}} = \frac{(\mathbf{X}^i_{s_{aligned}})^\top \mathbf{X}^i_{s_{aligned}}}{\mathbf{tr}((\mathbf{X}^i_{s_{aligned}})^\top \mathbf{X}^i_{s_{aligned}})}. \tag{10}$$

Finally, the proposed DTW-based transferred average covariance matrix of the aligned trials from previous subjects/sessions for each class c is computed as

$$\boldsymbol{\Sigma}_{\mathrm{DTW_c}} = (1/n_c) \sum_{i=1}^{n_c} \boldsymbol{\Sigma}^i_{\mathrm{s_{aligned}}}, \tag{11}$$

where n_c is the total number of trials of class c from other subjects/sessions.

2.3. Regularization Parameter Selection

Typically, regularization parameter is selected from a set of predefined values by applying cross-validation on the training data [31]. However, cross-validation becomes ineffective and in some cases impossible when we have very few training trials available from the target subject. Moreover, conventional optimization methods such as iterative optimization methods, or heuristic methods such as evolutionary algorithms could also be used to select the regularization parameter. However, a main drawback of using these techniques is that they require extensive computational time [32]. In this paper, we address the above-mentioned challenge by selecting the best regularization value using the classifier scores (i.e confidence scores) rather than the accuracy.

Algorithm 1: Offline method

Input: $\mathbf{\Sigma}_{\mathrm{DTW_c}}$, $\mathbf{\Sigma}_{\mathrm{SS_c}}$ for each class c, A predefined values of r, K cross-validation folds, and n_{eva} evaluation trials from the target subject

Output: Regularization parameter r^*

for $r = r_1$ *to* r_A **do**
  for $k = 1 : K$ **do**
    for c **do**
      calculate $\mathbf{\Sigma}_{\mathrm{TLR_c}}$ using (1)
    calculate the corresponding DTW-RCSP features using (3)
    train the classifier
    for $tr = 1 : n_{eva}$ **do**
      calculate the classifier score CS for each tr
      $score_{tr} = CS_{tr} * label_{tr}$
    $score_k = \sum_{tr=1}^{n_{eva}} score_{tr}$
  $score_r = \sum_{k=1}^{K} score_k$
Score* = **arg max** $score_r$
Return: r^* assigned to the highest **Score***

Figure 1. The proposed offline method to select the regularization parameter based on the confidence scores of the classifier on the training trials from the target subject

We propose using the classification scores to select the best regularization value in two different ways, namely referred to as offline and online. The offline method is applicable if we have sufficiently enough training trials available from the new target subject. The offline method applies cross-validation on the training trials and selects the regularization value that yields the highest summation of classification scores multiplied by the true class labels of the corresponding evaluation target trials over the 10-fold validations. Please see our algorithm in Fig. 1 for more details.

In online method, the few upcoming testing trials with known labels will be used for selecting regularization value. Thus, among a set of predefined values, the selected regularization value is the one which yields the highest summation of the classification scores multiplied by the true classification labels of the upcoming few testing trials. Fig. 2 provides more details on the proposed online regularization parameter selection method. The proposed online method can be used for any available number of training trials, while the proposed offline method is not applicable if less than K training trials are available from the new target subject where K is the number of cross-validation folds.

Algorithm 2: Online method

Input: $\mathbf{\Sigma}_{\mathrm{DTW_c}}$, $\mathbf{\Sigma}_{\mathrm{SS_c}}$ for each class c, A predefined values of r, and T upcoming labelled test trials from the target subject

Output: Regularization parameter r^*

for $r = r_1$ *to* r_A **do**
  for c **do**
    calculate$\mathbf{\Sigma}_{\mathrm{TLR_c}}$using (1)
  calculate the corresponding DTW-RCSP features using (3)
  train the classifier
  for $tr = 1 : T$ **do**
    calculate the classifier score CS for each tr
    $score_{tr}$= $CS_{tr} * label_{tr}$
  $score_r$=$\sum_{t=1}^{T} score_{tr}$
Score*= **arg max** $score_r$
Return: r^* assigned to the highest **Score***

Figure 2. The proposed online method to select the regularization parameter based on the classifier confidence scores of the upcoming few labelled testing trials

3. Experiments

3.1. Data Description

In order to evaluate the proposed transfer learning framework, two datasets with 9 and 17 subjects were used.

1) Dataset 2a from BCI Competition IV (medium dataset) [33]: This dataset includes 9 subjects' EEG data recorded using 22 electrodes. Each subject attended two sessions of data recording on two different days. A total number of 288 trials were recorded from each subject per session. Subjects were instructed to perform 4 motor imagery tasks. In this paper, we used only trials recorded for right and left-hand motor imagery (i.e. 144 trials). Moreover, to imitate a real life situation where the training and the testing trials of a new BCI user are recorded at the same session we used only data from the second session.

2) Dataset from [34] (large dataset): This dataset includes 19 healthy subjects' EEG data recorded using 27 electrodes. Two sessions at two separate days were recorded for each subject without feedback. In this dataset , subjects performed hand motor imagery, either left or right, versus rest condition. Each recorded sessions contained two runs, each run consisted of 80 trials without feedback, half of the trials is MI and the other half is rest condition. In this paper, only data from subjects who performed right hand motor imagery (17 subjects) were included. We did that to ensure the data used for transfer learning were neurologically relevant. Again, to fulfill the real life situation mentioned before, only data recorded in the first session are used.

3.2. Data Processing

A single zero-phase elliptic bandpass filter ranging from 8 to 30 Hz was used for EEG data filtration, since the range of frequencies that are mainly associated with performing motor imagery are included in this single frequency band. Then, the first and the last three spatial filters of CSP/CCSP/DTW-RCSP are used to obtain the spatially filtered signals as recommended in [35]. Thereafter, features are computed as the normalized log band power of the spatially filtered signals. Finally, Linear support vector machine (SVM) was used as the classifier.

For each subject, the investigated trials were divided into 3 sets, namely training, validation , and testing. The testing set consisted of the last 50 trials for the medium dataset, and the last 70 trials for the large dataset. For both datasets, the validation trials are the 10 trials immediately before testing trials, and the training set consisted of the remaining trials. Validation trials will be used in the proposed online method for regularization parameter selection. To assess the performance of the proposed DTW-RCSP framework, different scenarios have been considered when different numbers of training trials from new target subjects were available. Moreover, the DTW-based transfer learning covariance matrix is estimated using all the available training trials of the other subjects from the same dataset, except the target subject in each case. The optimum regularization parameter was selected from the predefined set of $r \in \{0, 0.1, \cdots, 1\}$.

The three proposed transfer learning-based regularized CSP algorithms (namely DTW-RCSP-CV, DTW-RCSP-Off, and DTW-RCSP-On) were evaluated. These algorithms are different on how the regularization parameter is selected. For DTW-RCSP-CV, the optimum regularization parameter is selected using 10 fold cross-validation on training data of the target subject based on the classification accuracy. For DTW-RCSP-Off and DTW-RCSP-On, the regularization parameter is selected using the proposed offline and online methods respectively. The results compares the proposed algorithms against two baseline algorithms, i.e. the commonly used subject-specific CSP algorithm, and CCSP (the first method proposed in [13]). The regularization parameter in CCSP is selected using cross-validation on the available training data of the target subject. In fact, if DTW alignment is omitted from the proposed DTW-RCSP-CV, it gets identical with CCSP.

4. Results and Discussion

The first part of this section presents the results when 5 or more trials per class were available from the target subject. Thus 10-fold cross-validation and our proposed offline method could be used to select the regularization parameter using the available training trials from the target subject. Fig. 3 compares the average classification accuracies of the baseline algorithms (CSP, and CCSP) with the results of the proposed DTW-RCSP-CV, DTW-RCSP-Off and Best-DTW-RCSP. Best-DTW-RCSP represents the

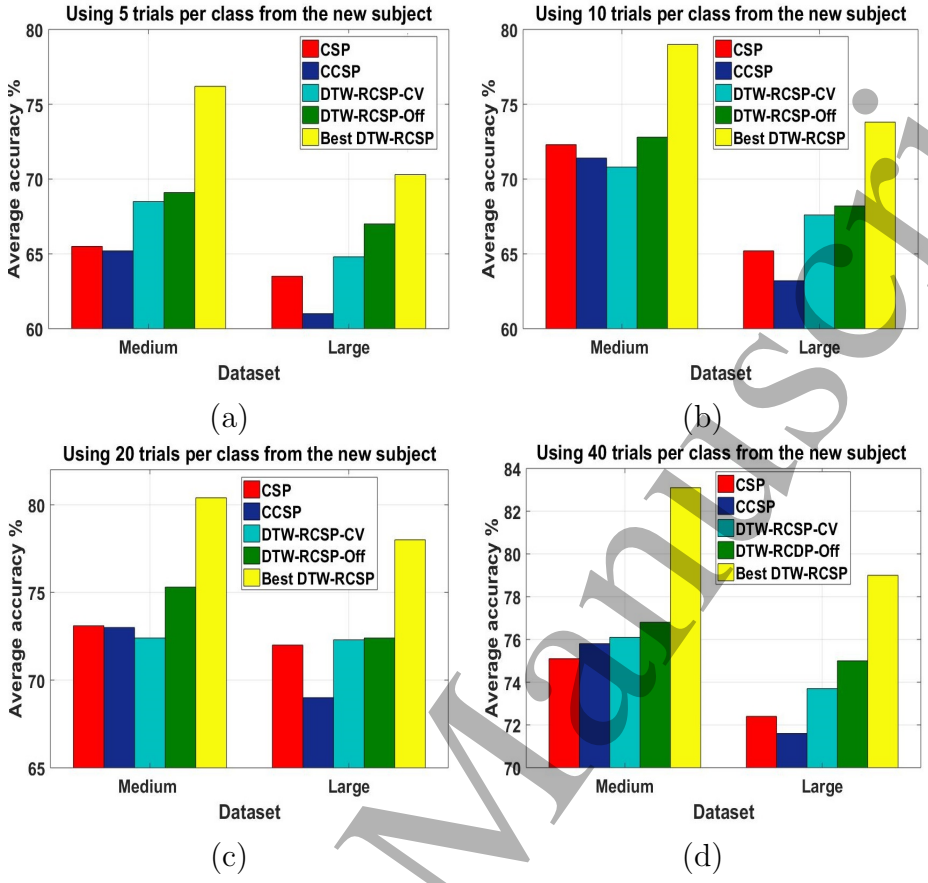


Figure 3. Comparison of the average classification results between the baseline algorithms (CSP, and CCSP), the proposed DTW-RCSP-CV, and DTW-RCSP-Off algorithms, and the DTW-RCSP results if the best regularization parameter yielding the highest test classification accuracy was selected (i.e. best DTW-RCSP). The classification results were calculated for different number of trials available for training from the new target subject.

classification accuracy if the best regularization parameter yielding the highest test accuracy could have been selected from $\{0, 0.1, \ldots, 1\}$. As shown in Fig. 3, for both datasets the proposed DTW-RCSP-Off algorithm outperformed the CSP and CCSP algorithms using most number of training trials. Interestingly, DTW-RCSP-Off was more successful than DTW-RCSP-CV in selecting regularization parameters yielding a higher average test classification accuracy.

Statistical paired t-tests revealed that for the large dataset using DTW-RCSP-Off was significantly better than CSP when 10 trials were available for training from the target subject ($P = 0.04$) and tended to be significantly better when 5 trials were available ($P = 0.09$). Besides, DTW-RCSP-Off was significantly better than CCSP when 5 trials were available with P value equal to 0.015. Moreover, DTW-RCSP-CV was significantly better than CCSP when 10 and 20 trials were available with P values equal to 0.04 and 0.017 respectively. These statistical results suggested that our proposed transfer learning algorithms performed significantly better than the baseline

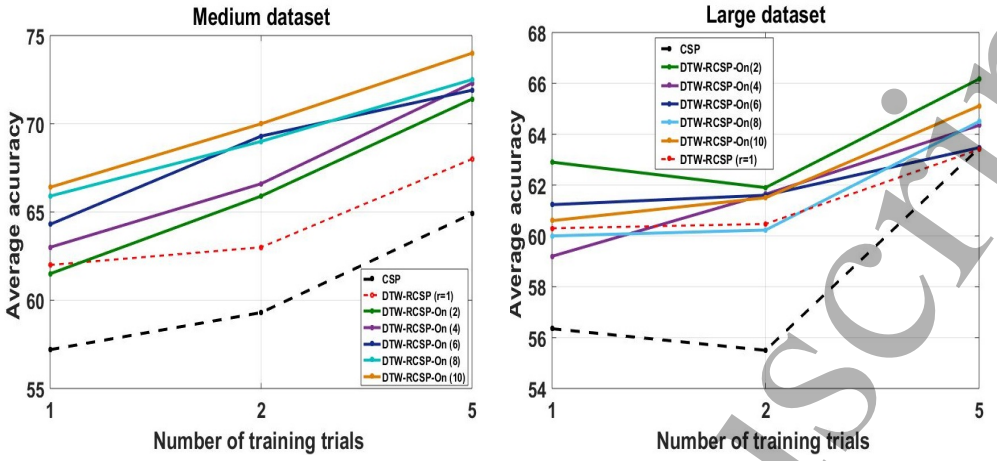


Figure 4. Comparing average classification results of the proposed DTW-RCSP-On(v), where v is the number of the validation trials used to select the regularization parameter and can be 2,4,6,8,or 10, with those of DTW-RCSP with (r=1) and CSP when 1,2, and 5 trials per class were available for training from the new target subject.

algorithms if a large number of previously recorded data from other subjects were available. Nevertheless, comparing the Best-DTW-RCSP results with those obtained by DTW-RCSP-CV and DTW-RCSP-Off revealed that if better regularization parameters could have been selected, the proposed DTW-RCSP algorithm could yield much higher significant improvements.

Although the proposed DTW-RCSP-Off algorithm improved the average classification accuracy, the Best-DTW-RCSP results showed that there was still room for improvement. Moreover, DTW-RCSP-Off with 10-fold cross validation for selecting the regularization parameter could not be viable if the number of the available training trials from the target subject is less than 5 trials per class. Therefore, in such cases our proposed DTW-RCSP-On could be used where the first few testing trials (referred to as the validation set in this study) were employed to select the regularization parameter. Apart from the benefits mentioned above, using the first few testing trials for selecting the regularization parameter could possibly reduce the negative impact of non-stationarity between the training and testing trials.

Fig. 4 shows the results of DTW-RCSP-On. The average classification accuracy across all subjects from each dataset was reported when the subject-specific training trials were as few as 1, 2, and 5 trials per class. The proposed DTW-RCSP-On, when different number of testing trials were used to select the regularization parameter, was compared to CSP and DTW-RCSP with ($r=1$) (i.e. only Σ_{DTW} was used for obtaining features). It is shown that using the proposed DTW-RCSP-On algorithm greatly improved the average classification accuracy. Impressively, when only 1 subject-specific trial per class was available for training, the proposed DTW-RCSP-On outperformed CSP by an average 3.7%, 5.2%, 6.4%, 8.1%, and 8.7% for dataset 1, and 8.1%, 2.9%, 4.9%, 3.7%, and 4.2% for dataset 2 when using 2, 4, 6, 8, and 10 validation trials for selecting the regularization parameter respectively. Moreover, in case of having only

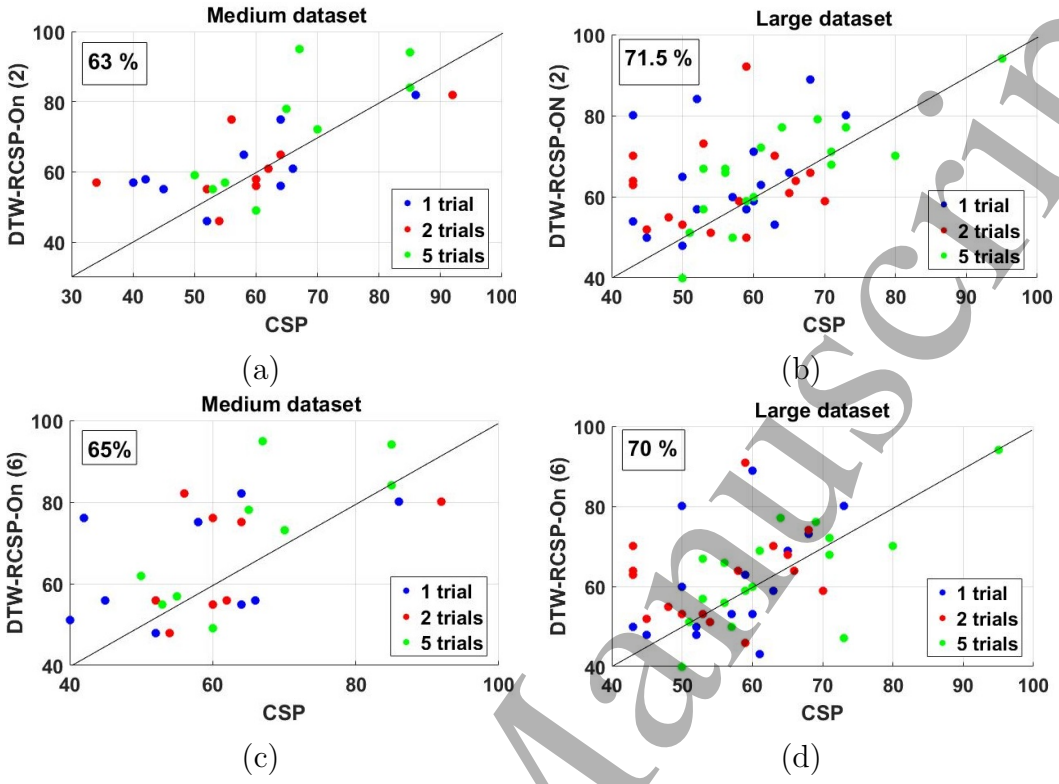
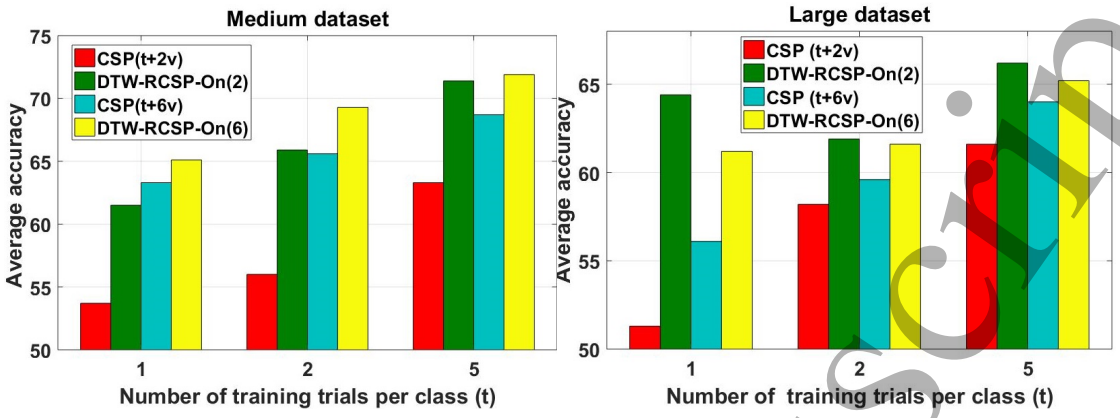


Figure 5. Classification accuracy comparison for each individual subject in both datasets when 1, 2, and 5 trials were available for training from the new target subject. (a) CSP versus DTW-RCSP-On(2) for medium dataset. (b) CSP versus DTW-RCSP-On(2) for large dataset. (c) CSP versus DTW-RCSP-On(6) for medium dataset. (d) CSP versus DTW-RCSP-On(6) for large dataset. "v" in DTW-RCSP-On(v) refers to the number of validation trials used for selecting the regularization parameter.

either 1 or 2 subject-specific trials per class, the classification results of DTW-RCSP with ($r=1$) outperformed CSP (i.e. only data from other subjects after DTW alignments were used to obtain features).

Fig. 5 provides more insight into the results of the proposed DTW-RCSP-On algorithm compared to CSP. As shown in Fig. 5, although for a few cases the use of DTW-RCSP-On led to small deterioration in the accuracy, for the majority of the subjects a considerable improvements had been achieved. Indeed, in many cases the improvement was as large as 20% to 35%.

Concerning statistical significance, A 6 (Number of trials= 1, 2, 5, 10, 20, and 40 trials per class)× 6 (Algorithms= CSP, DTW-RCSP-On (2,4,6,8,10)) repeated measure ANOVA test was performed on the results of both datasets followed by post-hoc analyses. For the large dataset, statistical results revealed that using different algorithms had a main effect on the classification accuracy ($P = 0.003$). Based on the post-hoc analysis, DTW-RCSP-On with different number of validation trials significantly outperformed CSP with P values equal to 0.001, 0.017, 0.046, 0.035, and 0.027 respectively for 2, 4, 6, 8, and 10 validation trials used to select the regularization parameter. Interestingly, using the proposed DTW-RCSP-On(2) was significantly better



(a) (b)

Figure 6. Comparison between DTW-RCSP-On(v) versus CSP trained with the available training trials(t) plus the used number the validation trials (v) when 1, 2, and 5 trials were available for training from the target subject.

than using any other number of testing trials (i.e. P values of 0.038, 0.05, 0.025, and 0.036 for 4, 6, 8, and 10 validation trials). Similarly, for the medium dataset, the statistical results revealed that using different algorithms had a main effect on the classification accuracy ($P = 0.012$). Based on the post-hoc analysis, DTW-RCSP-On with 2, 4, 6, 8, and 10 validation trials to select the regularization parameter significantly outperformed CSP with P values equal to 0.043, 0.043, 0.028, 0.022, and 0.023 respectively. However, using DTW-RCSP-On with 6, 8, or 10 testing trials to select the regularization parameter were not significantly different.

Another comparison was held to make sure that adding the validation trials used by DTW-RCSP-On for selecting the regularization parameter to the training trials of CSP would not achieve the same improvement as DTW-RCSP-On. Fig. 6 compares the average classification results of the proposed DTW-RCSP-On algorithm with the results of the CSP algorithm where the CSP was trained using the training trials plus the validation trials (i.e. CSP(t+v)). Due to limitation of the space, we limited this comparison to using 2 and 6 validation trials, and 1, 2, and 5 training trials. Fig. 6 shows that in all cases DTW-RCSP-On outperformed the corresponding CSP(t+v).

A 2 (Algorithms= CSP(t+v), and DTW-RCSP-On) $\times$ 2 (Number of validation trials= 2, and 6)$\times$ 3 (Number of training trials per class= 1, 2, and 5)) repeated measure ANOVA tests were performed on the results of both datasets followed by post-hoc analyses. For the large dataset, there was a main effect of using different number of training trials with $P = 0.024$. Moreover, the ANOVA results showed that our proposed DTW-RCSP-On tended to be significantly better than CSP(t+v) with $P = 0.059$. Posthoc analyses revealed that using 5 training trials per class were significantly better than using 1, and 2 trials with P values equal to 0.025 and 0.043 respectively. For the medium dataset using different algorithms, different training trials and different validation trials had main effects on the results with P values $0.042, 0.034$, and 0.013

respectively. Thus, we can conclude that in the medium dataset our proposed DTW-RCSP-On was significantly better than CSP(t+v) with $p = 0.042$. Posthoc analyeses showed that using 5 training trials per class were significantly better than 1, and 2 trials with P values equal to 0.016 and 0.023 respectively, and using 6 validation trials were significantly better than 2 with$P = 0.034$. In summary, our results showed that the proposed DTW-RCSP based transfer learning framework led to improved CSP features and hence improved BCI systems, particularly when a small subject-specific training data were available. The proposed framework will significantly improve future applications of BCI, such as BCI-based stroke rehabilitation, where the 20-30 minutes calibration time can be saved for real therapeutic interaction.

5. Conclusion

This paper proposed a novel DTW-based transfer learning framework on raw EEG and feature domains to improve the CSP covariance matrix estimations and hence enhance MI-based BCI systems. The proposed framework minimises the temporal variations between the EEG trials of other subjects and the few EEG trials of the target subject using DTW. Then the temporally aligned trials of other subjects are mixed with the few subject-specific trials in the CSP framework using a regularization parameter.

Our results suggested that applying the proposed framework reduced calibration time of the MI-BCI systems. Moreover, our proposed framework significantly outperformed the subject-specific CSP and CCSP algorithms in many different scenarios specially when data were available for transfer learning from a large number of subjects.

The proposed framework uses only one regularization parameter which is not computationally expensive compared to most of transfer learning-based regularized CSP algorithms that use 2 regularization parameters. Besides, the proposed online method required very slightly more computational time compared to CSP when the same number of trials are used. Thus, with these two benefits and with using only two validation trials the proposed DTW-RCSP-On could be potentially used for online applications.

Interestingly, our DTW-based transfer learning framework offered notable classification accuracy increase for majority of the participants specially when only few trials were available for training from the target subject. However, the observed improvement for some subjects with initially very low BCI performance was not pronounced. The possible reason might be having inseparable EEG signals between two classes. In future, further investigation is needed to identify these participants before transfer learning and possibly providing some human-training strategy.

In this paper the regularization parameters were selected using SVM scores. Importantly, The proposed transfer learning framework (DTW-RCSP) is not limited to the SVM classifier, and it can be applied on any classifiers. It is good to note that in the future other measurements could be used to select the regularization parameters and their performance could be compared to what we proposed.

6. Referencing

[1] E. A. Curran and M. J. Stokes, "Learning to control brain activity: a review of the production and control of EEG components for driving brain–computer interface (BCI) systems," *Brain and Cognition*, vol. 51, no. 3, pp. 326–336, 2003.

[2] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[3] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[4] Y. Guo, "Regularized discriminant analysis and its application in microarrays," *Biostatistics*, vol. 1, no. 1, pp. 1–18, 2005.

[5] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[6] C. Jeunet, B. NKaoua, S. Subramanian, M. Hachet, and F. Lotte, "Predicting mental imagery-based BCI performance from personality, cognitive profile and neurophysiological patterns," *Plos one*, vol. 10, no. 12, p. e0143962, 2015.

[7] C. Jeunet, E. Jahanpour, and F. Lotte, "Why standard brain-computer interface (bci) training protocols should be changed: an experimental study," *Journal of neural engineering*, vol. 13, no. 3, p. 036024, 2016.

[8] M. Krauledat, M. Schröder, B. Blankertz, and K.-R. Müller, "Reducing calibration time for brain-computer interfaces: A clustering approach," in *Advances in Neural Information Processing Systems*, 2007, pp. 753–760.

[9] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 614–617, 2010.

[10] A. M. Azab, J. Toth, L. S. Mihaylova, and M. Arvaneh, "A review on transfer learning approaches in braincomputer interface," in *Signal Processing and Machine Learning for Brain-Machine Interfaces*. The Institution of Engineering and Technology (IET), 2018, ch. 5.

[11] I. Hossain, A. Khosravi, I. Hettiarachchi, and S. Nahavandi, "Multiclass informative instance transfer learning framework for motor imagery-based brain-computer interface," *Computational Intelligence and Neuroscience*, 2018.

[12] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama, "Application of covariate shift adaptation techniques in brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 6, pp. 1318–1324, 2010.

[13] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for subject-to-subject transfer," *IEEE Signal Processing Letters*, vol. 16, no. 8, pp. 683–686, 2009.

[14] W. Samek, F. C. Meinecke, and K.-R. Müller, "Transferring subspaces between subjects in brain–computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[15] S.-H. Park, D. Lee, and S.-G. Lee, "Filter bank regularized common spatial pattern ensemble for small sample motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 498–505, 2018.

[16] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[17] C. Vidaurre, M. Kawanabe, B. Blankertz, K. Müller *et al.*, "Toward unsupervised adaptation of LDA for brain-computer interfaces." *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 587–597, 2011.

[18] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent

mental state classification in single trials," *Neural networks*, vol. 22, no. 9, pp. 1305–1312, 2009.

[19] W. Tu and S. Sun, "A subject transfer framework for EEG classification," *Neurocomputing*, vol. 82, pp. 109–116, 2012.

[20] A. M. Azab, L. Mihaylova, K. K. Ang, and M. Arvaneh, "Weighted transfer learning for improving motor imagery-based brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1352–1359, 2019.

[21] S. J. Pan, V. W. Zheng, Q. Yang, and D. H. Hu, "Transfer learning for wifi-based indoor localization," in *Association for the advancement of artificial intelligence (AAAI) workshop*, 2008, p. 6.

[22] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 610–619, 2013.

[23] W. Holmes, *Speech synthesis and recognition*. CRC press, 2001.

[24] W. A. Chaovalitwongse, Y.-J. Fan, and R. C. Sachdeo, "On the time series k-nearest neighbor classification of abnormal brain activity," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, no. 6, pp. 1005–1016, 2007.

[25] P. Senin, "Dynamic time warping algorithm review," in *Progress in brain research*. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA., 2008, vol. 885, no. 1-23, p. 40.

[26] A. M. Azab, L. Mihaylova, H. Ahmadi, and M. Arvaneh, "Robust common spatial patterns estimation using dynamic time warping to improve bci systems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3897–3901.

[27] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[28] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[29] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM international conference on data mining*. SIAM, 2001, pp. 1–11.

[30] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.

[31] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[32] D. J. McFarland, C. W. Anderson, K.-R. Muller, A. Schlogl, and D. J. Krusienski, "Bci meeting 2005-workshop on bci signal processing: feature extraction and translation," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 14, no. 2, pp. 135–138, 2006.

[33] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008–Graz data set A."

[34] M. Arvaneh, C. Guan, K. K. Ang, T. E. Ward, K. S. Chua, C. W. K. Kuah, G. J. E. Joseph, K. S. Phua, and C. Wang, "Facilitating motor imagery-based brain–computer interface for stroke patients using passive movement," *Neural Computing and Applications*, vol. 28, no. 11, pp. 3259–3272, 2017.

[35] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.