Multi-class Fuzzy Time-delay Common Spatio-Spectral Patterns with Fuzzy Information Theoretic optimization for EEG based Regression Problems in Brain Computer Interface (BCI)

Tharun Kumar Reddy, Vipul Arora, Laxmidhar Behera, *Senior Member, IEEE*, Yu-kai Wang, *Member, IEEE* and Chin-Teng Lin, *Fellow, IEEE*

Abstract*—Electroencephalogram (EEG) signals are one of the most widely used non-invasive signals in Brain Computer Interfaces (BCI). Large dimensional EEG recordings suffer from poor SNR (Signal to Noise Ratio). These signals are very much prone to artifacts and noise, so sufficient preprocessing is done on raw EEG signals before using them for classification or regression. Properly selected spatial filters enhance the signal quality and subsequently improve the rate and accuracy of classifiers, but their applicability to solve regression problems is quite an unexplored objective. This paper extends Common Spatial Patterns (CSP) to EEG state-space using fuzzy time delay and thereby proposes a novel approach for spatial filtering. The approach also employs a novel fuzzy information theoretic framework for filter selection. Experimental performance on EEG based reaction time prediction from a Lane-keeping task data from 12 subjects, demonstrated that the proposed spatial filters can significantly increase the EEG signal quality. A comparison based on Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and correlation to true responses is made for all the subjects. In comparison to the baseline fuzzy CSPROVR (Common Spatial Patterns Regression One Versus Rest), the proposed Fuzzy Time-delay Common Spatio-Spectral (FTDCSSP) filters reduced the RMSE on an average by 9.94%, increased the correlation to true reaction time on an average by 7.38% and reduced the MAPE by 7.09%.

***Index Terms*—Fuzzy Mutual Information (FMI), Common Spectro-Spatial Patterns (CSSP), Brain Computer Interface (BCI), Reaction Time (RT), Fuzzy Logic**

I. Introduction

A widely used approach for classification of EEG signals is spatial filtering. In EEG signal analysis, the signals recorded from the surface of scalp constitute a composite signal obtained from a combination of signal sources hidden inside skull [1]. Addressing volume conduction effects (also known as the smearing effect of the skull and brain), the spatial filter methods estimate sources, whose signals are more discriminative between classes than signals obtained purely at the surface or sensor level [2]. Spatial filtering broadly consists of defining a small number of new set of channels which are a linear combination of the existing ones [3],

$$\overline{\mathbf{x}} = \sum_{i} w_i \mathbf{x}_i \tag{1}$$

where, $\overline{\mathbf{x}}$ is the EEG signal after spatial filtering and $\mathbf{x}_i$ is the row vector containing signal from EEG channel 'i' before spatial filtering. Spatial filters are broadly categorized into two classes: fixed (or constant) spatial filters and data-adaptive filters. Among the fixed spatial filters, one can distinctly mention the Common Average Referencing (CAR) [4], Bipolar and Laplacian [5] local spatial filters that try to locally reduce the volume conduction effect and the background noise [6]. The next category of spatial filters are the data-adaptive filters which are optimized based on the training data. Similar to any data driven algorithm, the spatial filter weights w_i can be learned in an unsupervised way (without any utilization of class labels in training data) or in a supervised way, with each training sample being designated with its class label. Among the unsupervised spatial filters, the prominent ones are Principal Component Analysis (PCA) and Independent Component Analysis (ICA). PCA finds the spatial filters that account for most of the variance of the data. Independent Component Analysis (ICA) finds spatial filters whose resulting signals are independent from each other [7]. The latter approach is shown to be useful to design spatial filters to remove or dampen the effect of artifacts (EOG, EMG, etc.) on EEG signals [8]. Alternatively, spatial filters can be optimized in a supervised way, i.e., the weights will be defined in order to optimize some measure of classification performance.

In [9], a supervised spatial filter, called CSP has been designed for oscillatory activity based EEG BCI (OA-BCI). Several variants of CSP are proposed in the literature to deal with the issues of noise [10] and non-stationarities [11]. This algorithm (CSP) has tremendously contributed to the improvement in the performance of OA-BCI. CSP based features are used to enhance the binary classification performance of EEG data in comparison to fixed filters. The basic idea is to separate the EEG signal into several additive components which display significant differences in variance between the two classes. Further, Dornhege et al., [12] extended the traditional CSP from binary classification to multiple classes (CSP (OVR) can be found in the section II of [13]).

Tharun Kumar Reddy, Vipul Arora and Laxmidhar Behera are with the Department of Electrical Engineering, Indian Institute of Technology, Kanpur, Kanpur 208016, India (e-mails: tharun@iitk.ac.in, vipular@iitk.ac.in, lbehera@iitk.ac.in)

Y.-K. Wang and C.-T. Lin are with the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mails: yukai.wang@uts.edu.au, chintenglin@gmail.com)

CSP uses covariance matrices to magnify the class specific disparities on the spatial scale, but it neglects the frequency information which is quite necessary for modeling oscillatory activities [14]. In order to deal with the frequency information, spectral filters are applied to the raw EEG signal before the spatial filtering. However, it is observed that the bands of frequencies used in the spectral filters are subject-specific. Manual selection of frequency band for each subject is inefficient, hence a joint optimization of spatial and spectral filters is suggested. Common Spatial Spectral Patterns (CSSP) [15] was developed to address the above limitations of CSP.

A large portion of the spatial filters proposed so far are for the EEG classification problems, but not all BCIs are based on classification problems. Several brain signal decoding tasks need regression approaches to estimate continuous brain states. For instance, BCIs are used to estimate and monitor continuous cognitive workload levels [16] and user reaction time from oscillatory activity [17, 18]. In addition to classification techniques, regression models can also significantly gain from the deployment of spatial filters. Thus, Dahne et al. [19] proposed the Source Power Comodulation (SPoC) technique, which is looked at as an extension of CSP to regression problems. Infact, SPoC intends to find spatial filters such that the power of the filtered EEG signals maximally covaries with a continuous target prediction variable. Later in [2], authors introduced trace normalized and Tikhonov regularized SPoC variant (NTR-SPoC) which outperformed the standard SPoC for most of the individual subjects. Recently, the concepts of fuzzy logic and one versus rest approach are employed to extend the CSP filters for regression-one versus rest (CSPROVR) [13].

In the present work, authors extend the jointly optimized spatial and spectral filters (CSSP) approach for regression while incorporating the fuzzy time delay variable in the extended state space model. Also, a novel non-heuristic criterion in the form of fuzzy mutual information is formulated to select the spatial filters for feature extraction. The approach is validated on a novel dataset collected from 12 subjects for an EEG based lane keeping task experiments. This is one of the very few works which contributes to the domain of Regression based OA-BCIs.

The rest of this paper is organized as follows. Section II introduces our proposed spatial filters for the supervised BCI regression problem. Section III describes the experimental setup, data collection, EEG data preprocessing approaches, and the metrics to assess the performance of the obtained spatial filters. Section IV presents a comparison of the results of the proposed filters with several baseline studies and a parameter reactivity analysis for the proposed spatial filters. Section V discusses the limitations of the proposed techniques and provides guidelines for possible future research. Finally, Section VI draws conclusion.

II. Proposed Method

Notations: Matrices and vectors are denoted by bold faced letters in this paper. The bold text in tables represent results of proposed algorithms.

Let $\mathbf{X}^k \in \mathbb{R}^{C\times T}, k \in \{0, 1, .., N\}$ be the k^{th} EEG trial, C and T refer to number of channels and time samples respectively. Firstly, the interval $\big[0, 100\big]$ is divided into $K+1$ adjoint regions with the parting points named, p_k, where

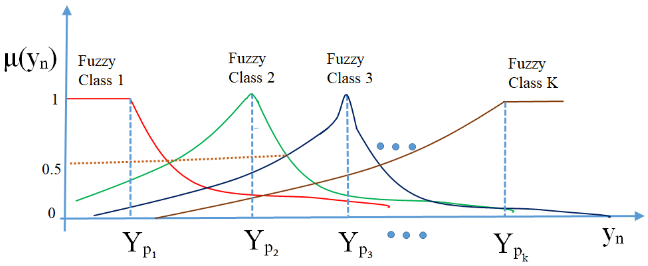


Fig. 1: M-fuzzy classes for EEG reaction time values

$$p_k = \frac{100k}{K+1}, k \in \{1, 2, ...K\} \tag{2}$$

For each p_k, we associate its ${p_k}^{th}$ percentile point Y_{p_k}in the training set y_n (cf. Fig. 1). From, these points, K regions are defined as the fuzzy sets. One can segregate all training y_n into K fuzzy classes. Each y_n belongs to a fuzzy class with a corresponding membership value $\in \big[0, 1\big]$.

A. Fuzzy Multiclass Common Spatial Patterns

In literature, there are three approaches for computing common spatial filters for multiclass problems. Two of them are one versus one (computes CSP's for every two class combinations) and one versus rest (computes CSP's for every class against rest of the classes considered jointly) approaches. In addition, simultaneous diagonalization using Joint Approximate Diagonalization (JAD)[20] can also be used. In this section, we first illustrate the CSP optimization for multiple classes problem using JAD.

1) Joint approximate Diagonalization approach: Consider a Multi-class problem with M fuzzy classes. Let $\mathbf{X}^k = \mathbf{X}^k{}_{(c,t)}$, $c = 1, 2, ..., C$, $t = 1, ..., T$ represent the pre-processed (post Band-pass filtered) EEG recording of the k^{th} trial, where C is the number of EEG channels used. In addition, let $Y_k \in \{0, 1, 2, ..., M-1\}$ corresponds to the fuzzy class-label of the k^{th} trial.

Further, an estimate of the class-wise fuzzy covariance matrix is obtained below. A fuzzy averaged trial matrix $\overline{\mathbf{X}}_i$ is obtained for each fuzzy class i. Then, the corresponding fuzzy class covariance matrix $\overline{\mathbf{\Sigma}}_i$ is computed.

$$\overline{\mathbf{X}}_i = \frac{\sum_{k=1}^{N_i} \mu_{k,i}\mathbf{X}^k}{N_i} \; i \in \{1, 2, \cdots M\}\,(k : Y_k = i-1) \tag{3}$$

In (3), N_i represents the number of trials in the i^{th} fuzzy class.

$$\overline{\mathbf{\Sigma}}_\mathbf{i} \;=\; \overline{\mathbf{X}}^i\overline{\mathbf{X}}^{i\intercal} \; i \in \{1, 2, ..., M\} \tag{4}$$

The obtained class covariance matrices are further normalized using

$$\mathbf{\Sigma}_i = \frac{\overline{\mathbf{\Sigma}}_i}{\mathrm{Tr}(\overline{\mathbf{\Sigma}}_i)} \;\forall i \in \{1, 2, \cdots M\} \tag{5}$$

Given the averaged and normalized covariance matrices for all M classes, the goal of JAD is to find a transformation matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ which simultaneously diagonalizes them. In other words, $\mathbf{W}$ must satisfy

$$\mathbf{W}\overline{\mathbf{\Sigma}}_i\mathbf{W}^{\intercal} = \mathbf{D}_i \ \forall i \in \{1, \cdots M\} \tag{6}$$

$$\sum_{i=1}^{M} \mathbf{D}_i = \mathbf{I}_{N \times N} \tag{7}$$

This decomposition can be done exactly for $N = 2$ but approximate solutions can be obtained for $N > 2$. Using this weight matrix $\mathbf{W}$, one can obtain the projected EEG trial matrix:

$$\mathbf{Z} = \mathbf{W}\mathbf{X}^k \tag{8}$$

The rows of matrix $\mathbf{W}$ are the spatial filters.

B. Fuzzy Mutual Information based Filter Selection

The rows of the matrix $\mathbf{W}$ are to be properly selected using a quantitative approach. Fuzzy Mutual Information can be used to select the optimal spatial filters.

Consider a universal set of matrices with its members $\mathbf{X}_l$ represented in input space as $X = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \cdot\cdot\cdot\mathbf{X}_r\}$, here $l = 1, 2, \cdot\cdot, r$ with 'r' denoting the size of training set. Consider 'M' fuzzy classes, where membership function can be defined for the l^{th} data matrix belonging to the i^{th} fuzzy class as follows:

$$\mu_{il} = \mu_i(\mathbf{X}_l) \in [0, 1] \tag{9}$$

Analogously, one can define average data sample matrix which belongs to fuzzy class 'i' as $\overline{\mathbf{X}}_i$ and the bound on the size of the set as 'L'

$$L = \max \|\overline{\mathbf{X}}_i - \mathbf{X}_l\|_\sigma \tag{10}$$

Thus a fuzzy membership μ_{il} can be computed as follows

$$\mu_{il} = \left(\frac{\|\overline{\mathbf{X}}_i - \mathbf{X}_l\|_\sigma}{L + \delta}\right)^{\frac{-2}{\nu-1}} \tag{11}$$

where, ν is the fuzzy membership parameter ($\nu = 1.152$), and $\delta > 0$ is a small number to avoid NaN values. σ is the standard deviation value involved in distance calculation. The sum of the membership values of each sample matrix over all the fuzzy classes is scaled to 1, i.e. $\sum_{i=1}^{M} \mu_{il} = 1$.

C. Fuzzy Entropy and Mutual Information

Entropy, as a measure of uncertainty, is in general used to denote Mutual Information (MI) between each input feature ($\mathbf{X}$) and the output class label (Y) according to:

$$I(\mathbf{X}, Y) = H(\mathbf{X}) + H(Y) - H(\mathbf{X}, Y) \tag{12}$$

$$= H(\mathbf{X}) - H(\mathbf{X}|Y) \tag{13}$$

Here, H($\mathbf{X}$) and H(Y) are the marginal entropies of $\mathbf{X}$ and Y, respectively, and H($\mathbf{X}$,Y) and H($\mathbf{X}|Y$) are the joint and conditional entropies of X and Y, respectively. According to the Fano's inequality [20], maximizing the MI between input features and class labels leads to a lowest probability of error. Various approaches based on kernel density predictors and histogram based estimators are cited in [21] for estimating probability density. Estimation of probability densities through histogram method carries a high computational complexity. This is because a huge number of input data points are to be processed to capture the density correctly. Therefore, this paper proposes a fuzzy membership based entropy for evaluating the mutual information. It has also been shown that usage of fuzzy entropy and MI can help to trim down the number of decision zones to be processed for classification, thereby decreasing the width of search space and computational time complexity [22].

Let $X = \{\mathbf{X}_1, \mathbf{X}_2, \cdot\cdot\cdot, \mathbf{X}_n\}$ be a discrete random variable containing 'n' symbols. Let $\mu_B(\mathbf{X}_i)$ denote the membership level of element $(\mathbf{X}_i)$ to fuzzy set B, and R be a set to point mapping $R : \mathbf{F}(2^{\mathbf{X}}) \rightarrow [0, 1]$. Here, the power set $2^{\mathbf{X}}$ of $\mathbf{X}$ is the set of all $\mathbf{X}$'s subsets. By $\mathbf{F}(2^{\mathbf{X}})$, we denote the fuzzy power set of $\mathbf{X}$, the crisp set of all fuzzy subsets of $\mathbf{X}$ [23]. Here, R is a function defined on fuzzy sets. In the seminal paper [24], Luca and Termini introduced fuzzy entropy in a way similar to Shannon's entropy. Authors suggested a set of postulates to be satisfied by fuzzy entropy function R.

1) $R(A) = 0$ iff $A \in 2^{\mathbf{X}}$, where $2^{\mathbf{X}}$ indicates powerset of $\mathbf{X}$ and A is a non-fuzzy set.
2) $R(A) = 1$ iff $\mu_A(X_i) = 0.5 \ \forall \ i \in \{1, 2, ..., n\}$.
3) $R(A) \leq R(B)$ if A is less fuzzy than B, i.e., $\mu_A(\mathbf{X}_i \leq \mu_B(\mathbf{X}_i))$ when $\mu_B(\mathbf{X}_i) \leq 0.5$ and $\mu_A(\mathbf{X}_i) \geq \mu_B(\mathbf{X}_i)$ when $\mu_B(\mathbf{X}_i) \geq 0.5$.
4) $R(A) = R(A^{\complement})$

Shannon's Entropy is defined for a discrete random variable X with a probability mass function $p(x_k)$ and it satisfies all the above postulates. It is given by:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) \tag{14}$$

Using (14), one can define an analogous fuzzy entropy using the fuzzy membership in (11). Assume, the input feature samples are broken into 'M' fuzzy classes and a fuzzy equivalent of a joint entropy can be defined for a particular class 'j' and an input feature 'FE' here as

$$P(FE, M_j) = \frac{\sum_{k \in B_j} \mu_{jk}}{N_p} \tag{15}$$

Here, $P(FE, M_j)$ is to be interpreted as the importance function of the samples belonging to a fuzzy class j, B_j is the set of index values of the training samples in class j and N_p is the total number of training samples. The joint fuzzy entropy of patterns in class j is indicated as $H(FE, M_j)$, which is equal to

$$H(FE, M_j) = -P_{FE, M_j} \log_2 P_{FE, M_j} \tag{16}$$

Entropy in (16) is summed up over the global set to obtain the unified joint fuzzy entropy $H(FE, \mathcal{M})$

$$H(FE, \mathcal{M}) = \sum_{j=1}^{M} H(FE, M_j) \tag{17}$$

One can check that, $H(FE, \mathcal{M})$ defined in (17) satisfies Luca and Termini postulates.

In a similar manner, marginal class entropy and marginal feature entropies can be defined. In order to compute marginal entropy of each feature, all the membership contributions are added corresponding to the samples in the M-fuzzy sets M_i as follows:

$$P(FE_{M_i}) = \frac{\sum_k \mu_{ik}}{N_p} \tag{18}$$

Marginal entropy is obtained by

$$H(FE) = -P_{FE_{M_i}} \log P_{FE_{M_i}} \tag{19}$$

To construct class marginal entropy $H_{\mathcal{M}}$, a corresponding fuzzy element of class probability is obtained by summing up fuzzy memberships on all the constructed fuzzy sets.

$$P(M_i) = \frac{\sum_{k \in B_i, \forall S} \mu_{ik}}{N_p} \tag{20}$$

Thus the marginal class entropy is given by

$$H(\mathcal{M}) = -P_{M_i} \log P_{M_i} \tag{21}$$

Using (12) one can obtain the fuzzy mutual information as:

$$\begin{aligned} I(FE, \mathcal{M}) &= H(FE) + H(\mathcal{M}) - H(FE, \mathcal{M}) \\ &= -P_{FE_{M_i}} \log P_{FE_{M_i}} - P_{M_i} \log P_{M_i} \\ &+ \sum_{j=1}^{M} P_{FE,M_j} \log_2 P_{FE,M_j} \end{aligned} \tag{22}$$

D. Fuzzy Time-delay CSSP

In [15], robust invariant CSP features are extracted by extending the state space by a delay co-ordinate. In the present work, a fuzzified time delay is incorporated in the extended state space model. We extend the model in (8) with the inclusion of a fuzzy time delay variable. We propose a generalized, extended fuzzy state space model:

$$\mathbf{Z^k} = \int_{\tau} \mu_(\tau) \mathbf{W}^{(\tau)} * (\delta^{\tau} \mathbf{X^k}) d\tau \tag{23}$$

where $\delta^{(\tau)}$ is the delay operator over the signal state space, μ_{τ} is the fuzzy membership value for the variable τ and $\mathbf{W}^{(\tau)}$ is the optimized fuzzy CSSP weights matrix.

$$\delta^{(\tau)}(\mathbf{X}^k) = \mathbf{X}^{(k-\tau)} \tag{24}$$

Assume that the time delay variable τ follows a exponential membership function $e^{-\tau}$. This is justified because generally in system dynamics, the contribution from higher order delays to the integral become insignificant after a particular delay threshold.

For the sake of practical implementation, (23) is approximated below:

$$\mathbf{Z^k} \approx \sum_{\tau=0}^{2} \mu_{(\tau)} \mathbf{W}^{(\tau)} * (\delta^{\tau} \mathbf{X^k}) \tag{25}$$

Further, the terms in (25) can be simplified to obtain

$$\mathbf{Z}^k = [\mathbf{W^{(0)}}\, \mathbf{W^{(1)}}\, \mathbf{W^{(2)}}] \begin{bmatrix} \mu_0 \mathbf{X}^{(k)} \\ \mu_1 \mathbf{X}^{(k-1)} \\ \mu_2 \mathbf{X}^{(k-2)} \end{bmatrix} \tag{26}$$

The concatenated vector $\begin{bmatrix} \mu_0 \mathbf{X}^{(k)} \\ \mu_1 \mathbf{X}^{(k-1)} \\ \mu_2 \mathbf{X}^{(k-2)} \end{bmatrix}$ is denoted as the equivalent EEG trial $\mathbf{X}^k$ in (8). Further, the optimization criteria is designed similar to (4) using the fuzzy multi-class covariance matrices $\overline{\mathbf{\Sigma}}_\mathbf{1}, \overline{\mathbf{\Sigma}}_\mathbf{2}, \overline{\mathbf{\Sigma}}_\mathbf{3}$ obtained from $\begin{bmatrix} \mu_0 \mathbf{X}^{(k)} \\ \mu_1 \mathbf{X}^{(k-1)} \\ \mu_2 \mathbf{X}^{(k-2)} \end{bmatrix}$. Following the approach outlined in steps (4)–(6) provides a solution to the optimization problem, a composite weight matrix, $\begin{bmatrix} \mathbf{W^{(0)}} & \mathbf{W^{(1)}} & \mathbf{W^{(2)}} \end{bmatrix}$. Each of the matrices $\mathbf{W^{(0)}}$, $\mathbf{W^{(1)}}$ and $\mathbf{W^{(2)}}$ apply to $\mu_0 \mathbf{X}^{(k)}$, $\mu_1 \mathbf{X}^{(k-1)}$ and $\mu_2 \mathbf{X}^{(k-2)}$ respectively. The fuzzy CSSP filters, which are the rows in the matrices $\mathbf{W^{(0)}}$, $\mathbf{W^{(1)}}$ and $\mathbf{W^{(2)}}$ are those filters which maximize the fuzzy mutual information criterion (22). In the standard CSP formulation [10], it is suggested to choose atleast two filters (corresponding to maximum and minimum variance directions) for each class. Following this guideline, $F = 2M = 6$ is chosen in the experiments for $M = 3$. In (26), calculating three weight matrices comprises calculating $3 \times 2M = 18$ row vectors.

In summary, the complete procedure is described in the form of Algorithm 1 .

Algorithm 1 Fuzzy time delay Common Spatio-Spectral Filters (FTDCSSP)

```
Input: EEG training set $(X^{(k)}, Y^{(k)})$  $k \in \{1, 2, ..N\}$
    $\mathbf{X}^{(k)} \in \mathbb{R}^{C \times T}, n \in \{1, 2, .., N\}$
    '$L_i$', number of spatial filters for each fuzzy class '$i$' (let $L_i = F$)
    '$M$' is the number of fuzzy classes
Output: Spatial filter matrices $\begin{bmatrix} \mathbf{W^{(0)}} & \mathbf{W^{(1)}} & \mathbf{W^{(2)}} \end{bmatrix}$
    (PREP) Preprocessed EEG trials are bandpass filtered to remove mean;
    Compute the thresholds $Y_{p_k}$ for the Gaussian membership functions for each of the fuzzy classes (Fig. 1);
    Compute $\overline{\mathbf{X}}_i$ using (3);
    Compute Covariance matrix $\overline{\mathbf{\Sigma}}_\mathbf{i}$ for each fuzzy class '$i$' using (4);
    Normalize the covariance matrices for each class as per (5);
    Compute the Spatial filters matrix $\mathbf{W}$ obtained after applying Joint Approximate Diagonalization (JAD) on the normalized covariance matrices as per (6);
    Extract '$L_i$' filters per each fuzzy class using maximization of Fuzzy Mutual Information criterion (22);
    Spatial filter matrix $\begin{bmatrix} \mathbf{W^{(0)}} & \mathbf{W^{(1)}} & \mathbf{W^{(2)}} \end{bmatrix}$ is obtained, which contains $\sum_{i=1}^{M} L_i$ number of rows;
    return $\begin{bmatrix} \mathbf{W^{(0)}} & \mathbf{W^{(1)}} & \mathbf{W^{(2)}} \end{bmatrix}$
```

E. Regression Metrics

RMSE, Correlation Co-efficient (CC) (also known as the Pearson's Linear or rank correlation) and MAPE are the metrics in use for judging the regression performance. Assume, there are N training points, y_{d_i} represents the true reaction time value of the i^{th} data point and y_i represents the predicted reaction time value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{d_i} - y_i)^2}{N}} \tag{27}$$

$$MAPE = \frac{100}{N}\frac{\sum_{i=1}^{N}(|y_{d_i} - y_i|)}{y_{d_i}} \tag{28}$$

III. Experiments and Data

This section introduces a lane-keeping experimental task that was used to evaluate the performance of the proposed spatial filtering algorithms, the corresponding Reaction Times (RTs), EEG data preprocessing procedures, and the feature extraction procedures.

A. *Experimental Setup*

The EEG signals are recorded from 30 sintered Ag/AgCl EEG active electrode sites (all of these are referred to linked mastoids). All the EEG electrodes referring to the right ear lobe were placed in accordance with a modified International 10–20 system of electrode placement. During the driving task, Lane Deviation (LD) was kept constant and RT of each subject was recorded. The RT stands for the time period between onset of deviation and onset of response and is used as an objective measure of the drowsiness (DS) level during each lane departure event [25]. In order to assure that each subject had a homogeneous DS level during the entire period, DS levels of each subject are normalized to the range $[0, 1]$. The normalized DS levels are presented as time varying and are employed as the desired output of the proposed system in this study. Before each stage of the experiment, the subjects answered a questionnaire regarding their sleep patterns to ensure that they were not insomniacs or taking any medication that can effect their cognitive states. To properly figure out their driving performance, the participants showed up at a pre-test session to corroborate that none of the participants were stricken with simulator queasiness. The Institutional Review Board of the Veterans General Hospital, Taipei, Taiwan, approved the study. A total of 12 university students (average age 22.4, standard deviation 1.6) from the National Chiao Tung University (NCTU) in Taiwan volunteered to support the data collection efforts over a five months period to study EEG correlation of attention and performance changes under specific conditions of real world drowsiness [26]. The voluntary and fully informed consent of the participants of this research was obtained as required by federal and army regulations [27], [28]. The Institutional Review Board of NCTU approved the experimental protocol. Simulated driving experiments were conducted on a virtual reality (VR) based dynamic driving simulator. A real car frame was mounted on a six degree-of-freedom Stewart motion platform which moved in sync with the driving scene during ‘motion’ sessions (filename ending with ‘m’). The motion platform was inactive during ‘motionless’ sessions (filename ending with ‘n’). The VR driving scene simulated nighttime cruising (100 km/h) on a straight highway (two lanes in each direction) without other traffic. The computer program generated a random perturbation (deviation onset), and the car started to drift to the left and right of the cruising lane with equal probability. Following each deviation, subjects were required to steer the car back to the cruising lane as quickly as possible using the steering wheel (response onset), and hold on the wheel after the car returned to the approximate center of the cruising lane (response offset). A lane departure trial is defined as consisting of three events, deviation onset, response onset, and response offset. The next lane departure trial randomly occurs about 5 to 10 sec after response offset in the current trial. If the subject does not respond promptly within 2.5 (1.5) sec, the vehicle will hit the left (right) roadside without a crash and continue to move forward against the curb event, during which the subject completely ceases to respond. No intervention was made when the subject fell asleep and stopped responding. After reaching the lapse period, subjects resumed the task voluntarily and steered the car back to the cruising position at the earliest.

The goal is to predict reaction time using a 5s EEG trial immediately before it.

B. *Performance Evaluation Process*

Following procedure is used to assess the performance of various spatial filters and feature extraction methods.

1) EEG Data PreProcessing:

- At first, raw EEG data was passed through standard pre-processing pipeline (PREP) of EEGLAB to increase the signal to noise ratio. It comprises mainly of three operations [29].
  - Removing line noise.
  - Determining and removing robust reference signal.
  - Interpolating the bad channels.
- Further, the data was downsampled to 250 Hz.
- Thus, the data was epoched to 5 sec trials, i.e. if the event is starting at time ‘t’ then the EEG data from $[t-5, t]$ is used to predict the RT. Each EEG trial is of size 30×1250.
- Outliers in the RT values are removed by ignoring the EEG trials with RT values greater than sum of mean and three times the standard deviation.
- Then, the obtained trials are filtered by a $[1, 20]$ Hz finite impulse response band-pass filter to make the channel zero mean and to remove the irrelevant high frequency components.
- The obtained data is then fed through the appropriate spatial filters.

2) Reaction Time PreProcessing: The crude RTs for two subjects are shown in Fig. 2. Data plotted in Fig. 2(b) is from a typical subject, whose RT values were mostly shorter than 2 second. Data plotted in Fig. 2(a) is from a subject with possible data recording issues, because lots of RTs were longer than 5 seconds, which are highly absurd in practice. So, the subject in Fig. 2(a) is removed from consideration in this paper, and only data from the remaining 11 subjects is used. The final distribution of RTs obtained after pre-processing are shown in Fig. 3.

As shown in Fig. 3, the RTs are highly non-stationary (indicative of noise), and there were obvious anomalies. It is very important to nullify the outliers and noise power so that the performances of different algorithms can be more accurately compared and contrasted. Therefore, following procedure for RT preprocessing is used in this paper.

a) Mean Based Outlier Gating: This method aims to suppress abnormally large RTs. For each subject a threshold of $\mu + 3\sigma$ is computed, where μ is the mean RT and σ is the standard deviation for the RTs of a subject over all the sessions. For each subject, RT values larger than this threshold are treated as outliers. The thresholds are different for different subjects. The final distribution of the RTs after the above preprocessing, can be found in Fig. 3.

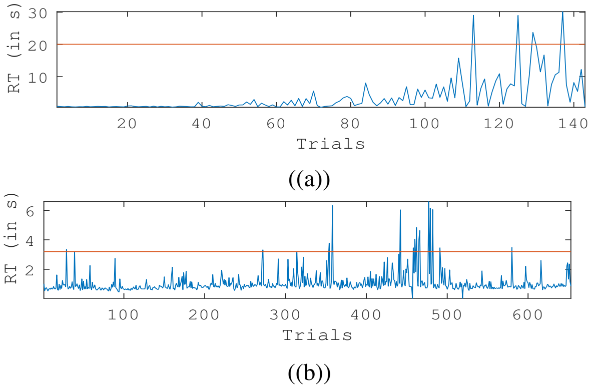
3) Feature Sets: 8-fold Cross-Validation (CV) is used to compute the regression performance for every possible fusion of feature set and regression method. Following feature sets are extracted for each EEG trial. PCA is used for dimensionality reduction and the components explain 90% of the variance.

- Theta and Alpha powerband features (PSD) [1] are extracted from the EEG trials which are initially spatial filtered by Common Average Referencing (CAR)[6]. A vector of size $(30 \times 2) = 60$ (band power in decibels) constitutes the feature vector. It is further reduced using PCA to a 18 dimensional vector (Final feature vector is denoted by '*FS1*').
- Theta and Alpha powerband features are extracted from the EEG trials which are initially spatial filtered by fuzzy CSPROVR (CSP-regression One versus Rest). A vector of size $(6 \times 2) = 12$ (band power in decibels) constitutes the feature vector (Final feature vector is denoted by '*FS2*').
- Theta and Alpha powerband features are extracted from the EEG trials which are initially spatial filtered by FTDCSSP. A vector of size $(18 \times 2) = 36$ (band powers in decibels) constitutes the feature vector. It is further reduced using PCA to a 14 dimensional vector (Final feature vector is denoted by '*FS3*').
- RG features[2][17] are extracted from the EEG trials which are initially spatial filtered by Common Average Referencing (CAR). RG features are extracted from $Z^k \in \mathbb{R}^{90\times1250}$ in (26) and a RG feature vector of size $\frac{90\cdot91}{2} = 4095$ is obtained. It is further reduced using PCA to a 18 dimensional vector (Final feature vector is denoted by '*FS4*').
- RG features are extracted from the EEG trials which are initially spatial filtered by fuzzy CSPROVR (CSP-Regression One versus Rest). RG features are extracted from $Z^k \in \mathbb{R}^{MF\times1250}$ in (26) and a RG feature vector of size $\frac{(MF)\cdot(MF+1)}{2} = \frac{6\cdot7}{2} = 21$ is obtained (Final feature vector is denoted by '*FS5*').
- RG features are extracted from the EEG trials which are initially spatial filtered by FTDCSSP. RG features are extracted from $Z^k \in \mathbb{R}^{3MF\times1250}$ in (26) and a RG feature vector of size $\frac{(3MF)\cdot(3MF+1)}{2} = \frac{18\cdot19}{2} = 171$ is obtained. It is further reduced using PCA to a 15 dimensional vector (Final feature vector is denoted by '*FS6*').

IV. Experimental Results

A. Information revelation from the features

In this section, the authors analyze the salience of the extracted features from various channels in connection to RT prediction. At first, each subject's EEG data is partitioned into training and testing sets using 8 fold CV. Spatial filters are designed using FTDCSSP performed on the training data. The results are shown in Fig. 4. The data points on the left of the black dotted line were used for training, and the right ones are used for testing. The correlation coefficients of the data points with reaction time values are found out for both training and testing feature data points (in our case, the RG features). Each of the three curves in a subplot are the maximally correlated features (some of the good features are negatively correlated with the RT) identified from the training data. The training and testing correlation coefficients are shown on the left and right of the respective feature channel. Observe that the features from the fuzzy CSPROVR had much higher training and testing correlations to the RT than those from CAR, which confirms the knowledge that CSPROVR enhances the signal quality better than CAR. This is attributed to the Rayleigh cost function defined for fuzzy CSPROVR (which treats the current class transformed covariance in the numerator as signal and rest of the other classes transformed covariances in the denominator as noise). The proposed FTDCSSP method has the highest training and testing correlations, owing to the consideration of frequency (rhythm) specific information in the CSP algorithm with the inclusion of time delay components. Also note that in this paper, the notations CSP-OVR and CSPROVR are used synonymously and they both refer to the fuzzy CSP one versus rest algorithm.



((a))

((b))

Fig. 2: Reaction times of (a) abnormal and (b) normal subjects

B. Regression Performance Study

The RMSEs, CCs and MAPEs of LASSO using the six feature sets (explained in section III-B3) are shown in Fig. 5 for the 11 subjects. For each subject, 8-fold Cross Validation

[1] Welch's method is used to compute the Power Spectral Density (PSD) in the Theta (4-8 Hz) and Alpha bands (8-13 Hz) respectively.

[2] RG features are computed using https://goo.gl/kCyAx1

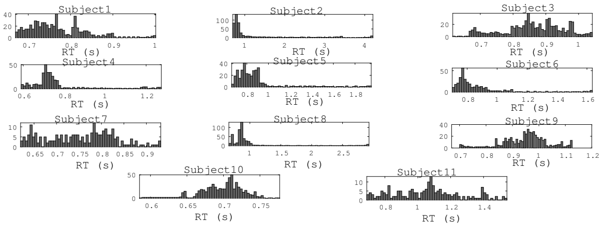


Fig. 3: Distribution of EEG reaction time values

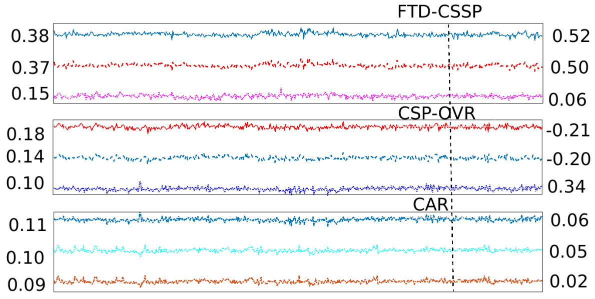
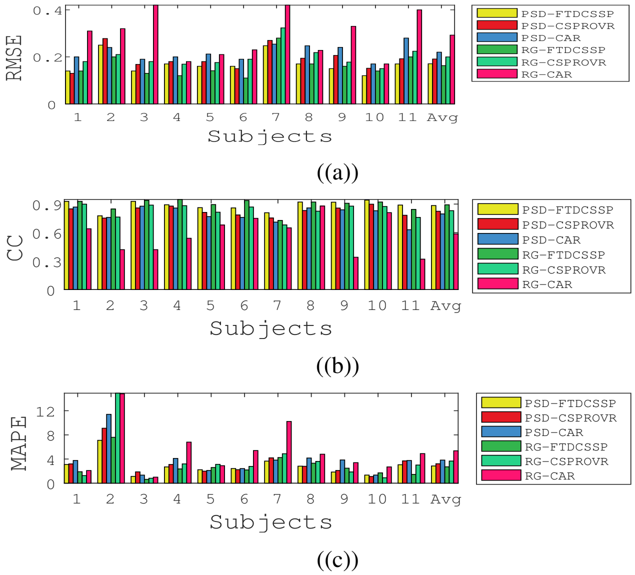


Fig. 4: Riemannian Geometry features extracted across trials (number of trials along x-axis) after passing through different spatial filters (CAR, fuzzy CSP-OVR, FTD-CSSP) and the respective training and testing CCs with the Reaction time on y-axis.

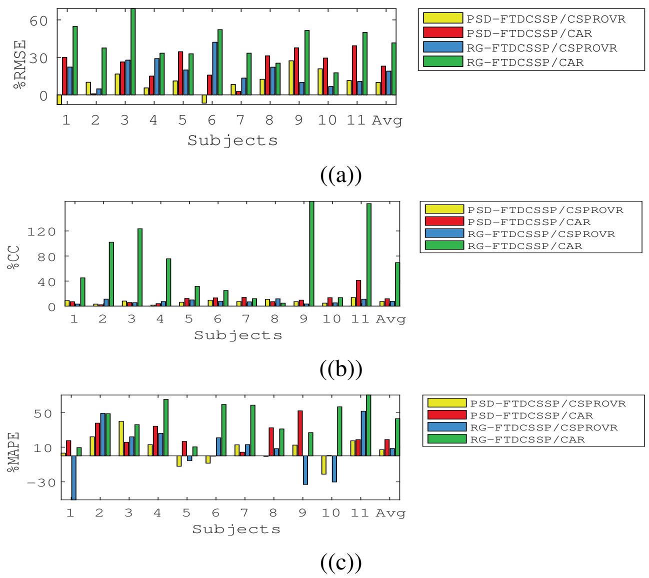


((a))

((b))

((c))

Fig. 5: RMSE, CC and MAPE for 11 subjects

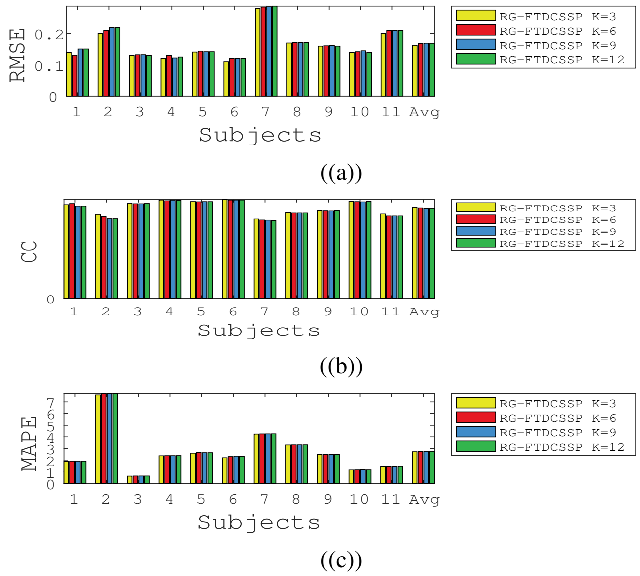


((a))

((b))

((c))

Fig. 6: Percent changes in RMSE, CC and MAPE for 11 subjects



((a))

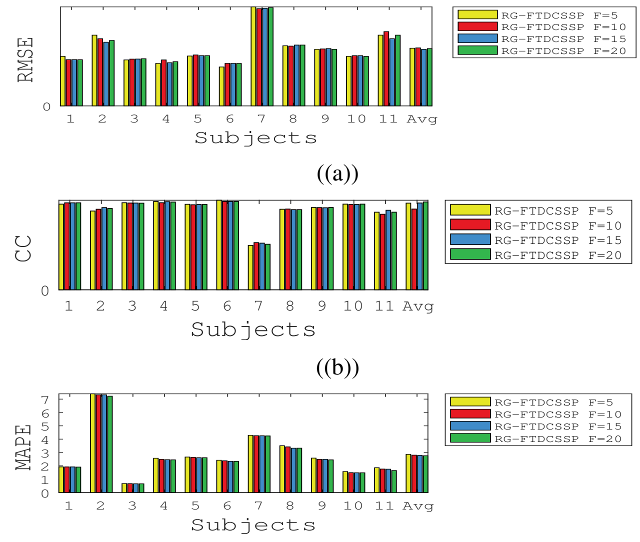
((b))

((c))

Fig. 7: RMSE, CC and MAPE with varying fuzzy classes for 11 subjects

was used to the partition the feature set into training and validation sets. The performance is averaged across all the 8-folds. Also, the performance across all the subjects is averaged to obtain the last (12^{th}) bar plot in the same plot. Here, in general FTD-CSSP recorded the best performance, and both FTD-CSSP and CSPROVR achieved much smaller RMSEs, MAPEs and much larger CCs than CAR, suggesting that our extension of CSP from supervised classification to supervised regression can indeed improve the regression performance. In conclusion, LASSO had better performance with FTD-CSSP than both CSP-OVR and CAR.

The respective percentage performance improvements of LASSO using the six feature sets are shown in Fig. 6. For instance, the terms in legend 'PSD-FTDCSSP/CSPROVR' represents the improvement in performance of FTDCSSP algorithm over the CSPROVR using the PSD features. 'RG-FTDCSSP/CAR' represents the improvements in performance of FTDCSSP algorithm over the CAR using the RG features. Using LASSO regressor with Theta and Alpha powerband features, on an average, FTDCSSP recorded a 9.94% smaller RMSE, a 7.09% smaller MAPE and a 7.38% larger CC than CSPOVR. While utilizing Riemannian Geom-



((a))

((b))

((c))

Fig. 8: RMSE, CC and MAPE with varying spatial filters for 11 subjects

TABLE I: Two-way ANOVA results

	LASSO		
	RMSE	CC	MAPE
$p-value$	0.002	0.01	**0.001**

etry (RG) features, on an average, FTDCSSP had performed with a 18.97% smaller RMSE, a 8.47% smaller MAPE and a 7.58% larger CC than CSPOVR. Also, while employing Theta and Alpha powerband features, FTDCSSP recorded a 22.96% smaller RMSE, a 18.78% smaller MAPE and a 11.82% larger CC than CAR. While using RG features, on an average, FTDCSSP recorded a 41.59% smaller RMSE, a 42.92% smaller MAPE and a 69.48% larger CC than CAR.

Further, statistical analysis is performed to test various hypotheses on MAPE, RMSE and CC's across the subjects. Firstly, a two-way Analysis of Variance (ANOVA) is performed for multiple regression algorithms with the aim to analyze if the RMSE, CC and MAPE differences amidst the feature sets are statistically significant, with the subjects set as a random factor. The results are shown in Table I, ($p-value < 0.05$) which indicated that there were statistically significant differences in RMSEs, CCs and MAPEs among different feature sets with LASSO regression.

Then, post-hoc multiple comparison tests in the form of paired t-tests are employed to find out if the difference between any couple of algorithms is statistically significant, with the $p - value$ corrected employing the False Discovery Rate method [30]. The p-values are shown in Table II, where in all the values are statistically significant. In all the couple of cases '*FS4*'(CAR) '*FS6*'(FTDCSSP), '*FS4*'(CAR) '*FS5*'(CSPOVR), '*FS6*'(FTDCSSP) '*FS5*'(CSPOVR), the results are always statistically significant. The bolded ones in Table II are extremely statistically significant[3].

TABLE II: Post-hoc multiple comparison tests based on paired t-tests

	LASSO					
	RMSE		CC		MAPE	
	'*FS4*'	'*FS5*'	'*FS4*'	'*FS5*'	'*FS4*'	'*FS5*'
'*FS5*'	**0.001**		**0.001**		**0.001**	
'*FS6*'	0.02	0.01	**0.001**	**0.001**	**0.001**	**0.001**

C. Fine-tuning parameters of FTD-CSSP

There are two tunable parameters in CSPROVR: K, the number of fuzzy classes, and F, the number of spatial filters for each fuzzy class. In this subsection, we study the variation of the regression performance in relation to these two parameters.

In Fig. 7, $K \in \{3, 6, 9, 12\}$ and F is fixed at 6. All the algorithms are repeated for 5 trials, each with a 8-fold Cross Validation and the trial averaged regression results are plotted. Averaging across all the subjects, all the values of $K \in \{3, 6, 9, 12\}$ performed similarly across MAPE, RMSE and CC regression metrics.

The number of spatial filters F lies in the set $\{5, 10, 15, 20\}$. We fix the value of K at 3 and vary F. Performance in terms of RMSE, MAPE and CC metrics as visible from Fig. 8 indicates that increasing F decreases RMSE, increases CC and decreases MAPE, on an average. But the performance saturates for a certain optimal F. Increasing the numerical values of K and F amplifies the algorithmic computations. Considering a trade-off between computational cost and regression performance $K = 3$ and $F = 6$ is chosen.

V. Discussion and Future-work

This paper extended the CSSP algorithm for regression using fuzzy state space model. In addition, a fuzzy mutual information criterion is formulated to appropriately choose the fuzzy spatial filters for regression. 8-fold CV is employed for all the experiments. Joint Approximate Diagonalization (JAD) is used for obtaining spatial filters used in the experiments. The computational complexity analysis and a study of robustness of the proposed method is provided in a supplementary file for readers. Two interesting directions for future research are suggestible, firstly, incorporation of regularization within JAD framework for improving the generalizing ability of spatial filters and secondly, transfer learning is to be incorporated into the FTDCSSP framework to improve generalization across subjects and within subjects across sessions.

VI. Conclusion

This paper extends CSSP to EEG based reaction time prediction problem using fuzzy time delay. The approach also uses a novel fuzzy information theoretic approach for filter selection. Experimental results on EEG-based reaction time prediction from a Lane-Keeping task data collected from 12 subjects, demonstrated that the proposed spatial filters can significantly enhance the EEG signal quality. A comparison based on RMSE, MAPE and correlation to true responses (CC)

[3] $p - value \leq 0.001$

is made for all the subjects. In comparison to CSPOVR, the proposed FTDCSSP filters reduce the RMSE on an average by 9.94%, increase the correlation to true reaction time on an average by 7.38% and reduce the MAPE by 7.09% which readily suggests further testing in several scenarios to make it robust for use in drowsy OA-BCIs.

VII. Acknowledgement

This work was supported in part by the Australian Research Council (ARC) under discovery grant DP180100670 and DP180100656 and also in part by the Ministry of Human Resource Development (MHRD), Government of India under the project 'MHRD-EE-2016150'. Research was also sponsored in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022 and W911NF-10-D-0002/TO 0023. First author is supported by the project TCS/CS/2011191A. First author would like to thank Satyam, a dual degree student at IIT Kanpur for the discussions on state of the art. First author would also like to thank Prof. Dr. Benjamin Blankertz for his insights on CSSP.

References

[1] P. Von Bünau, F. C. Meinecke, S. Scholler, and K.-R. Müller, "Finding stationary brain sources in EEG data," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, Buenos Aires, Argentina, August 2010, pp. 2810–2813.

[2] A. Meinel, F. Lotte, and M. Tangermann, "Tikhonov regularization enhances EEG-based spatial filtering for single trial regression," in *International Graz Brain-Computer Interface Conference*, Graz, Austria, September 2017, pp. 1–6.

[3] F. Lotte, "A tutorial on EEG signal-processing techniques for mental-state recognition in brain–computer interfaces," in *Guide to brain-computer music interfacing*. Springer, 2014, pp. 133–161.

[4] T. D. Lagerlund, F. W. Sharbrough, and N. E. Busacker, "Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition," *Journal of clinical neurophysiology*, vol. 14, no. 1, pp. 73–82, 1997.

[5] M. Teplan *et al.*, "Fundamentals of EEG measurement," *Measurement science review*, vol. 2, no. 2, pp. 1–11, 2002.

[6] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for EEG-based communication," *Electroencephalography and clinical Neurophysiology*, vol. 103, no. 3, pp. 386–394, 1997.

[7] A. Kachenoura, L. Albera, L. Senhadji, and P. Comon, "ICA: a potential tool for BCI systems," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 57–68, 2008.

[8] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch, "EMG and EOG artifacts in brain computer interface systems: A survey," *Clinical neurophysiology*, vol. 118, no. 3, pp. 480–494, 2007.

[9] 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.

[10] 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.

[11] 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.

[12] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Muller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 993–1002, 2004.

[13] D. Wu, J.-T. King, C.-H. Chuang, C.-T. Lin, and T.-P. Jung, "Spatial filtering for EEG-based regression problems in brain-computer interface (BCI)," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 771–781, 2018.

[14] G. Pfurtscheller and F. L. Da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[15] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE transactions on biomedical engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.

[16] C. Walter, W. Rosenstiel, M. Bogdan, P. Gerjets, and M. Spüler, "Online EEG-Based Workload Adaptation of an Arithmetic Learning Environment," *Frontiers in human neuroscience*, vol. 11, p. 286, 2017.

[17] D. Wu, B. J. Lance, V. J. Lawhern, S. Gordon, T.-P. Jung, and C.-T. Lin, "EEG-based user reaction time estimation using Riemannian geometry features," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 2157–2168, 2017.

[18] T. K. Reddy, V. Arora, S. Kumar, L. Behera, Y. Wang, and C. T. Lin, "Electroencephalogram based reaction time prediction with differential phase synchrony representations using co-operative multi-task deep neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018, accepted for publication.

[19] S. Dähne, F. C. Meinecke, S. Haufe, J. Höhne, M. Tangermann, K.-R. Müller, and V. V. Nikulin, "SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters," *NeuroImage*, vol. 86, pp. 111–122, 2014.

[20] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1991–2000, 2008.

[21] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[22] H.-M. Lee, C.-M. Chen, J.-M. Chen, and Y.-L. Jou, "An

efficient fuzzy classifier with feature selection based on fuzzy entropy," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 31, no. 3, pp. 426–432, 2001.

[23] B. Kosko, "Fuzzy entropy and conditioning," *Information sciences*, vol. 40, no. 2, pp. 165–174, 1986.

[24] L. De and S. Termini, "A definition of non-probabilistic entropy in the setting of fuzzy entropy," *J Gen Syst*, vol. 5, pp. 301–312, 1972.

[25] C.-H. Chuang, L.-W. Ko, T.-P. Jung, and C.-T. Lin, "Kinesthesia in a sustained-attention driving task," *Neuroimage*, vol. 91, pp. 187–202, 2014.

[26] S. Kerick, C. Chuang, J. King, T. Jung, J. Brooks, B. Files, K. McDowell, and C. Lin, "Inter-and intra-individual variations in sleep, subjective fatigue, and vigilance task performance of students in their real-world environments over extended periods," 2016,in press.

[27] U. D. of the Army, "Use of volunteers as subjects of research," *Government Printing Office*, no. AR, pp. 70–2, 1990.

[28] U. D. of Defense Office of the Secretary of Defense, "Code of federal regulations protection of human subjects," *Government Printing Office*, vol. Government Printing Office, no. 32 CFR 19, 199, p. 19, 1999.

[29] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The PREP pipeline: standardized preprocessing for large-scale EEG analysis," *Frontiers in neuroinformatics*, vol. 9, p. 16, 2015.

[30] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.