Spatio-spectral patterns based on Stein kernel for EEG signal classification

STUDENT: STEVEN GALINDO NOREÑA
SUPERVISOR: DAVID AUGUSTO CÁRDENAS PEÑA, PHD

Universidad Tecnológica de Pereira
Faculty of Engineering
Master in Electrical Engineering
Automatics Research Group
Pereira, Risaralda, Colombia
April 9, 2021

Abstract

Attention-Deficit/Hyperactivity Disorder (ADHD) is a childhood-onset neurological disorder that can persist in adolescence and adult life, reducing concentration, memory, and productivity. The main drawback with mental health abnormalities of this type is the traditional diagnostic technique. Since this is based exclusively on a symptomatological description without considering any biological data, leading to high overdiagnosis rates. To address the above problem, clinical researchers are attempting to extract ADHD biomarkers from recorded electroencephalographic (EEG) signals. Among the most common biomarkers are Theta/Beta Ratio and P300, of which recent studies have shown a lack of significance on the differences between ADHD and control subjects. Besides, another great challenge in EEG processing is given by the sensitivity of the signals, since they can be easily affected by background noise, muscle artifacts, head movements and flickering that greatly impair their quality, which limits its introduction into real world applications. This work proposes an EEG signal representation methodology for identifying subject-wise discrepancies of inhibitory responses, decoding the data structure, and supporting diagnosis of mental disorders. For this, first we develop a feature extraction approach based on the common spatial patterns (CSP) from EEG signals to support the ADHD diagnosis as show in Chapter 3. Then, we develop a methodology for the representation of EEG signals that uses the similarity between time series through their covariance matrices in the Riemannian manifold of positive semidefinite matrices (PSD), using the logdet-divergence of Jensen Bregman, the Stein kernel, and Centered Kernel Alignment (CKA) as a cost function to perform a spatial filters optimization. Finally, in chapter 5 we present a methodology for the diagnostic support of ADHD. The proposal involves the use of the optimal spatial patterns developed in Chapter 4, a decomposition in brain rhythms, and the discriminative decoding of Chapter 3. The resulting subject-wise features fed a linear discriminant analysis as the supported-diagnosis tool. Achieved 93% accuracy rate proves that the discriminative index based on the stein spatial patterns outperforms conventional biomarkers in the ADHD diagnosis.

Aknowledgments

I would first like to thank my supervisor, Professor David Cárdenas, whose experience was invaluable in developing this work. Also, my sincere thanks to my colleagues and friends in the research group for their wonderful collaboration. Furthermore, I would like to thank my family for their wise advice and understanding during the development of my studies. Finally, I want to sincerely thank the Automatics Research Group at UTP for allowing me to develop and fund my research.

This work was supported by the research project “Herramienta de apoyo al diagnóstico del TDAH en niños a partir de múltiples características de actividad eléctrica cerebral desde registros EEG” with code number 111080763051 funded by MinCiencias and by the Vice-rectory for research from "Universidad Tecnológica de Pereira" by the project with code E6-20-3.

Contents

List of Tables

List of Figures

Chapter 1

List of Symbols and Abbreviations

1.1 Symbols

Symbol	Definition
$\mathcal{X}$	Input space
$\boldsymbol{A}$	Matrix $\boldsymbol{A}$
$\boldsymbol{b}$	Vector $\boldsymbol{b}$
$\mathbb{E}$	Expected value
∇F	Gradient of the differentiable function $F(\cdot)$
$\mathrm{tr}\{\cdot\}$	Trace operator
$\langle , \rangle_F$	Frobenius inner product
k	Kernel function
$\delta(\cdot,\cdot)$	Delta Dirac function
$\|\|\cdot\|\|_F$	Frobenius norm
$\|\cdot\|$	Absolute value

1.2 Abbrevations

The following abbreviations are used in this manuscript:

ADHD	Attention-Deficit Hyperactivity Disorder
BCI	Brain–Computer Interface
CKA	Centered Kernel Alignment
CSP	Common Spatial Patterns
DC	Decreasing Condition
EEG	Electroencephalography
ERN	Error-Related Negativity
ERP	Event-Related Potential
FIR	Finite Impulse Response
IC	Increasing Condition
MI	Motor Imagery
MKL	Multiple Kernel Learning
MKSSP	Multi-Kernel Stein Spatial Patterns
RSST	Reward Stop Signal Task
SPD	Symmetric Positive Definite
SSP	Stein Spatial Pattern
TBR	Theta/Beta Ratio

Chapter 2

Introduction

2.1 Problem statement

Attention deficit hyperactivity disorder (ADHD) is a neurological disorder of childhood-onset. It is the most common behavioral problem of the school period, affecting the child's daily life and learning due its impulsivity and low self-esteem symptoms [2]. ADHD can persist in adolescence and adult life (15-65%), manifesting itself as a lower ability to concentrate, lower memory capacity, and lower productivity [3]. Nonetheless, the current examination relies exclusively on a symptomatological description without considering any biological data, yielding high overdiagnosis rates [4]. To address the above issue, clinical researches attempt extracting ADHD biomarkers from the electroencephalographic (EEG) signals recorded while performing demanded tasks. For instance, the power ratio of theta and beta bands (TBR) compares the power of slow waves (4-7 Hz) against the fast waves (13-30 Hz) from a subject in resting state, expecting that ADHD subjects evoke higher power at slow waves than in the fast ones. The US Food and Drug Administration agency approved TBR for clinical diagnosis in 2013. However, recent studies state that there is not enough evidence on the TBR robustness as a diagnostic tool [5]. Another widely studied biomarker corresponds to a positive deflection on the event-related potentials (ERPs) approximately 300 milliseconds after a stimulus occurs, termed P300. Although some studies evidenced a latency increase and an amplitude decrease in ADHD patients, other clinical researches suggest that the differences between ADHD and controls are not significant. [6]. Another considered biomarker consists of a negative ERP deflection between 50-100 milliseconds at the frontal-central region when incorrect responses occur, termed error-related negativity (ERN) wave [7]. Despite the reported decreased amplitudes in ADHD children [8], ERN needs to be further explored as an ADHD biomarker [9].

Some of the problems that feature extraction techniques from EEG must deal with are the presence of background noise, muscle artifacts, head movements, and blinks. This greatly impairs the performance of the techniques, limiting their introduction into real-world applications. [10, 11]. One way to deal with the above issues comprises

the design of robust feature sets. Among the wide variety of feature extraction approaches, the Common Spatial Patterns (CSP) stands out for capturing different mental states activity within the EEG. CSP projects signals from two classes into a space with maximum variance for one class and minimum for the other, providing discriminant features [12]. In this sense, the relative power in the CSP space, computed from the channel-wise covariance matrix, characterizes a band-passed multichannel EEG trial. Nonetheless, the interclass variability within subjects hampers the system effectiveness for some applications [13].

As an initial solution, multiple CSP-based variants optimize the spectral filtering for yielding discriminative features under varying conditions. For instance, the Common Spatio-Spectral Patterns (CSSP) approach enhances a finite impulse response (FIR) filter by incorporating a time delay for filtering to improve CSP performance proposed [14]. The Common Sparse Spectral-Spatial Patterns modified the CSSP by simultaneously tuning the FIR and CSP filters [15]. Later, Novi et al. introduced the CSP features from multiple sub-bands for EEG classification with a score fusion strategy [16]. In the same approach, the Filter-Bank CSP (FBCSP) exploited the potential correlation between CSP characteristics extracted from different bands to improve the signal discrimination [13]. In general, filter-banked feature extraction approaches outperform conventional multichannel time-series representations, including traditional CSP, in supervised learning schemes [17]. Nonetheless, spectral variants of CSP hardly decode mental states that activate spatially close regions, no matter the frequency differences [18].

Other approaches decode the trial variability through Riemannian manifolds composed of Symmetric Positive Definite (SPD) matrices, i.e., trial covariance estimations. Due to preserving the geometric structure and behaving like a matrix Hilbert space, the Riemannian manifolds devote conventional pattern recognition machines to time-series classification [19]. However, the covariance matrices as features suffer from the curse of dimensionality, since the usual matrix dimensions compare to the number of training samples [20]. For coping with high-dimensionality, various approaches map data from the Riemannian manifold into lower-dimensional vector spaces. For instance, the extension of three nonlinear dimension reduction (DR) approaches, namely, Local Linear Embedding (LLE), Hessian LLE, and Laplacian Eigenmaps, to the Riemannian geometry allowed the motion and image segmentation from a clustering point of view [21]. Nonetheless, such nonlinear DR algorithms lack a parametric mapping to the low-dimensional space, depending on an interpolation stage [22]. In addition, Principal Geodesic Analysis (PGA) emerges as a principal component analysis generalization for Riemannian manifolds by finding a tangent space with maximized variance [23]. Two subsequent tangent space approaches introduce kernel-based mappings [21, 24]. However, tangent space projections distort data structure, with larger distortions at regions far from the space origin [25]. Another DR alternative maps from high to low-dimensional manifolds, where the resulting output manifold serves as input for existing SPD-based algorithms [20]. As an example, a linear mapping takes advantage of provided labels to maximize the geodesic distance among samples from different classes while minimizing distances among equally-labeled samples [22]. Despite favoring the

supervised tasks, the linear combination of distances as a cost function underperforms at inherently nonlinear distributed classes.

Therefore, there is a need for extracting features from EEG signals to suitable identify evoked activity by a known paradigm, decode differences in mental states between and within subjects, and enhance the supervised learning in supported diagnosis applications.

2.2 Justification

Attention deficit hyperactivity disorder (ADHD) is a childhood-onset neurological disorder characterized by symptoms such as inattention, hyperactivity, and impulsivity. This disorder is the most common behavior problem of the school period, affecting the child's daily life and learning. ADHD can persist into adolescence and in some cases into adult life, where it manifests itself in a reduced ability to concentrate, reduced memory capacity, low productivity, among others [26]. This condition can be classified according to the DSM-V (Diagnostic and Statistical Manual of Mental Disorders) into three subtypes: predominant attention deficit, predominant, or combined hyperactivity-impulsivity. The prevalence of this disorder in the population ranges between 8% and 20%, this percentage varies according to the diagnostic method used, the age of the patients, geographic location, and level of education [27]. In Colombia, according to the 2015 national mental health survey, 3% of children between 7 and 12 years old suffer from ADHD. Prevalence studies in Antioquia and Caldas have made it possible to establish a global prevalence in the population of 15% to 17%, and in Bogotá with school populations shows that 5.7% of the children in the schools evaluated have ADHD [28]. The prevalence in Colombian adults is not known, but it is believed that the condition persists between 2% and 5% of affected children [29].

Currently, there are problems that affect the diagnosis of ADHD, since it is a process of clinical observation. Such problems include discrepancies in the information provided by parents and teachers, as well as the overlap of attentional and behavioral symptoms with other disorders [30]. In view of the above, strategies have been proposed to integrate biomarkers with the regular clinical evaluation of ADHD to help the specialist in a specific and sensitive diagnosis of the disease or to determine confounding factors with possible comorbidities or different etiologies [31]. Studies such as the one presented in [32], present a methodology for the assisted diagnosis of ADHD integrating the TBR biomarker and a multidisciplinary team (psychiatrist, psychologist, and neurodevelopmental pediatrician), where 275 children and adolescents with suspected of ADHD with the traditional diagnosis and with the proposed methodology, finding that with the first one there was a 34% overdiagnosis. Therefore, it is important to implement tools that help in the diagnosis of ADHD. Considering that the Automatics research group of the Universidad Tecnologica de Pereira has carried out several bioengineering projects using signals from this type, the development of a feature extraction methodology that allows discriminating control subjects from patients

with the disorder can be considered viable. This research is framed in the project entitled “Herramienta de apoyo al diagnóstico del TDAH en niños a partir de múltiples características de actividad cerebral desde registros EEG” with code 1110-807-63051, of MinCiencias.

2.3 State of the art

In recent years, the diagnosis of ADHD has been made using observation by a doctor and the reports that parents and teachers gave regarding the behaviors and activities carried out by children. However, the information provided was subject to discrepancies and confusion with other disorders [33]. For this reason, to provide effective diagnoses without being confused with possible related disorders, strategies have been proposed that involve the use of biomarkers in the clinical evaluation of ADHD [32].

The biomarkers currently used are based on the electroencephalographic activity of patients diagnosed with ADHD [34]. Emerging, for example, endophenotypic markers, which are obtained with electrophysiological measurements, and genes such as DAT1 and DRD4 that encode dopamine-dependent neurotransmission [35]. Likewise, the p300 wave has been considered a marker of an attentional process. This wave appears as a positive deflection 300 milliseconds after a stimulus is presented, which generates great interest to evaluate its morphological characteristics through the analysis of Cognitive Evoked Potentials [36]. Another biomarker used is the Teta/Beta spectral ratio (TBR). This ratio was approved by the Food and Drugs Administration (FDA) of the United States in June 2013 under the name of Neuropsychiatric evaluation EEG-Bassed Assessment Aid (NEBA).

Likewise, various features extracted from EEG signals have been proposed in an attempt to discriminate subjects with the disorder. Among which are the frontal lobe alterations (excess theta dominant frontal activity or excess alpha frequency activity). Parietal alpha variant (alpha frequency greater than 12Hz in the posterior cortex, with normal to high amplitudes). Alpha-independent diffuse slow activity (increased delta and theta (1-7 Hz) with or without posterior dominant slow rhythm). Mixed fast and slow activity (activity increases below 8 Hz. absence of alpha and increased beta frequency), among others [37, 38, 39, 40, 41]. Another study identifies that in ADHD patients the Error related negativity wave is low or even absent in comparison with the control patients and can change according to the emotional involvement in inhibition tasks [42].

Regarding the EEG-based feature extraction methodologies, several solutions have been proposed attempting to capture the electrical activity patterns resulting from different mental states. Among these, the Common Spatial Patterns algorithm (CSP) presented in [12] stands out. The CSP algorithm projects the signals of two classes into a space of maximum variance for one class and minimum for the other, providing discriminative features. Besides, many variants of the original CSP algorithm can be found, which try to increase the effectiveness, as presented in [43]. In this article the

authors present the Common Spatio-Spectral Patterns (CSSP) approach, where a finite impulse response (FIR) filter is improved by incorporating a time delay for filtering. Allowing for individually tuned frequency filters at each electrode position resulting in an improved and more robust machine learning procedure. Later in [14] a variation of CSSP is proposed where, in addition to optimizing the spectral filtering, an optimization of the spatial filters is performed. This variation is called Common Sparse Spectral-Spatial Patterns. Later, in [15] another approach is presented where the extraction of CSP features is performed using frequency sub-bands, where the best bands are chosen using a Linear Discriminant Analysis. From the same perspective of frequency bands, in [13] a variation of CSP using filter banks called Filter Bank Common Spatial Pattern (FBCSP) is proposed. In this work, the specific frequency band for each subject is optimized for the feature extraction. Finding the specific frequency bands of each subject that better encode the different mental states. In general, it is observed that feature extraction approaches that take into account the spectral relationship of brain activity and different mental states outperform conventional multi-channel time series representations, including traditional CSP, in supervised learning schemes.

2.4 Objectives

2.4.1 General objective

To develop an EEG signal representation methodology for identifying subject-wise discrepancies of inhibitory responses, decoding the data structure, and supporting diagnosis of mental disorders.

2.4.2 Specific objectives

- To develop a methodology for feature extraction from EEG signals registered in inhibitory control tasks that decodes the discrepancies between different mental states.
- To develop a methodology for the representation of EEG signals based on second-order statistics that favors the class separability and decodes the non-linear relationships.
- To develop a methodology to support the diagnosis of ADHD from EEG signals, which takes into account spatio-spectral information and mental states within an inhibitory control paradigm using second-order statistics.

Chapter 3

CSP-based discriminative capacity index from EEG

This chapter presents the development of a methodology for the feature extraction from EEG signals registered in inhibitory control tasks, decoding the discrepancies between different mental states based on the Common spatial patterns (CSP) algorithm. To this end, CSP projects EEG signals into a decorrelated space so that successful and failed inhibitions differ in their variance. Then, the obtained spatial patterns and their eigenvalues allow us to extract a channel-wise feature accounting for the capacity to discriminate inhibitions. Validation on synthetic EEG simulating two visual stimuli evidence that the discriminative capacity index suitably identifies evoked sources. On real EEG, results prove that the proposed features outperform behavior and ERN biomarkers in the supported ADHD diagnosis using a Linear Discriminant Analysis (LDA) classifier. The work presented in this chapter was accepted and published at the 28th European Signal Processing Conference (EUSIPCO) in 2020 [44].

3.1 Methods

3.1.1 Common Spatial Patterns

Let a set of N labeled multichannel EEG time series (trials), acquired from a single subject $\mathcal{X}=\{\boldsymbol{x}_n(t)\in\mathbb{R}^C, y_n\in\{+,-\}\}_{n=1}^N$, where C stands for the number of channels, $t\in[1,T]$ indexes the time instants, and y_n labels the n-th time series $\boldsymbol{x}_n(t)$ as either, positive $(+)$ or negative $(-)$ class. The Common Spatial Patterns (CSP) technique maps trials using a matrix $\boldsymbol{W}\in\mathbb{R}^{C\times D}$ into a space of uncorrelated sources as $\boldsymbol{z}_n(t)=\boldsymbol{W}^\top\boldsymbol{x}_n(t)$, being $\boldsymbol{z}_n(t)\in\mathbb{R}^D$ the projected time series, and $D\leq C$ the number of components. Taking into account that trials are band-pass filtered, the covariance matrix of $\boldsymbol{z}_n(t)$, assessing the source correlation, is computed as in Equation (3.1) where $\mathbb{E}_t\{\cdot\}$ stands for the expected value over t, and $\boldsymbol{\Sigma}_n\in\mathbb{R}^{C\times C}$ denotes the covariance of

$\boldsymbol{x}_n(t)$.

$$\begin{aligned} \boldsymbol{S}_n &= \mathbb{E}_t\left\{\boldsymbol{z}_n(t)\boldsymbol{z}_n^\top(t)\right\} = \boldsymbol{W}^\top \mathbb{E}_t\left\{\boldsymbol{x}_n(t)\boldsymbol{x}_n^\top(t)\right\}\boldsymbol{W} \\ \boldsymbol{S}_n &= \boldsymbol{W}^\top \boldsymbol{\Sigma}_n \boldsymbol{W} \end{aligned} \tag{3.1}$$

Therefore, the covariance of the i-th class in the CSP space $\boldsymbol{S}_i \in \mathbb{R}^{D\times D}$ results from averaging the trials so that $y_n{=}i$:

$$\begin{aligned} \boldsymbol{S}_i &= \mathbb{E}_n\left\{\boldsymbol{S}_n{:}y_n{=}i\right\} = \boldsymbol{W}^\top \mathbb{E}_n\left\{\boldsymbol{\Sigma}_n{:}y_n{=}i\right\}\boldsymbol{W} \\ \boldsymbol{S}_i &= \boldsymbol{W}^\top \boldsymbol{\Sigma}_i \boldsymbol{W} \end{aligned} \tag{3.2}$$

Since CSP aims at discriminating trials in terms of their projected variance, $\boldsymbol{W}$ must maximize $\boldsymbol{S}_+$ and minimize $\boldsymbol{S}_-$, while diagonalizing them to guarantee the source decorrelation [12]. These three goals are achieved by solving the following problem generalized eigenvalue problem:

$$\boldsymbol{w}^\top \boldsymbol{\Sigma}_+ = \lambda \boldsymbol{w}^\top (\boldsymbol{\Sigma}_+ + \boldsymbol{\Sigma}_-), \tag{3.3}$$

where $\boldsymbol{w}$ is an eigenvector and λ its associated eigenvalue. The set of D eigenvectors $\{\boldsymbol{w}_d\}$ solving Equation (3.3) correspond to the columns of $\boldsymbol{W}$, resulting in an orthonormal projection matrix. Besides, $\boldsymbol{W}$ produces unitary total variances at each CSP axis as the solution of Equation (3.3) yields $\boldsymbol{S}_+ + \boldsymbol{S}_-{=}\boldsymbol{I}$. Hence, the vectors $\boldsymbol{w}_d$, also termed spatial filters, must maximize the variance of trials in one class and minimizes the variance of the others to meet above constraint.

3.1.2 Discriminative decoding of CSP

Given the projected trial $\boldsymbol{z}_n(t)$, the input time series can be re-obtained using a second linear mapping, $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{z}(t)$. Since CSP provides the spatial filters, such a matrix is computed as $\boldsymbol{A} = \boldsymbol{W}^{-\top}$. The columns of $\boldsymbol{A} \in \mathbb{R}^{C\times D}$, called spatial patterns, explain how a presumed source is read by the channels.

Further, unitary total covariance constraint in Equation (3.3) results in eigenvalues bounded to the range $\lambda \in [0,1]$. If $\lambda_d{=}1$, variances in the $\boldsymbol{w}_d$ axis are one for positive trials, and zero for negatives. While $\lambda_d{=}0$ indicates that $\boldsymbol{w}_d^\top \boldsymbol{\Sigma}_n \boldsymbol{w}_d{=}0$ for positive trials and $\boldsymbol{w}_d^\top \boldsymbol{\Sigma}_n \boldsymbol{w}_d{=}1$ for negative ones. Consequently, spatial patterns with associated eigenvalues $\lambda{=}1/2$ lack discriminative information. To account for the discriminative capacity of spatial pattern, this work corrects the eigenvalues according to Equation (3.4), so that $\lambda'_d{=}0$ comes from an axis without class separability and $\lambda'_d{=}1$ enhances the variance of either class.

$$\lambda'_d = 2\left|\lambda_d - \frac{1}{2}\right| \tag{3.4}$$

Finally, the weighted average of $\boldsymbol{A}$ along the columns provides a scalar value indicating the capacity of each channel to discriminative between the two classes:

$$\rho_c = \sum_{d=1}^{D} \lambda'_d |a_{cd}|, \tag{3.5}$$

with c and d indexing rows and columns of $\boldsymbol{A}$, as well as the channels and patterns from CSP, respectively. As a result, the corrected eigenvalues favor those patterns maximizing the class separation in terms of the trial variance, and the vector $\boldsymbol{\rho} \in \mathbb{R}^C$, in the channel space, describes how well CSP solves the discrimination problem for $\mathcal{X}$.

3.2 Datasets

3.2.1 Synthetic EEG records

To assess the discriminative capacity of the proposed index, two sets of EEG records were simulated using the SEREEGA toolbox [45]. SEREEGA generates synthetic event-related EEG signals by describing the electrical brain activity.

Dataset SEREEGA-1

This dataset simulates two different classes evoking potentials at different electrical activity sources, named A and B. Those sources contra-laterally located, A frontal and B occipital, as shown in Figure 3.1. For each source, 300 trials were simulated with a 200 ms pre-stimulus and 800 ms post-stimulus, at a sample rate of 250 Hz. Table 3.1 summarizes the parameters of the negative (ERP-N) and positive (ERP-P) evoked potentials composing the trials, as well as the additive white noise. Therefore, SEREEGA-1 holds M=600 trials, each of them with C=30 channels distributed all over the scalp, and lasting T=250 time instants. Trials with source A evoked are labeled as positive (y_n=+), while the ones with source B are negative (y_n=−).

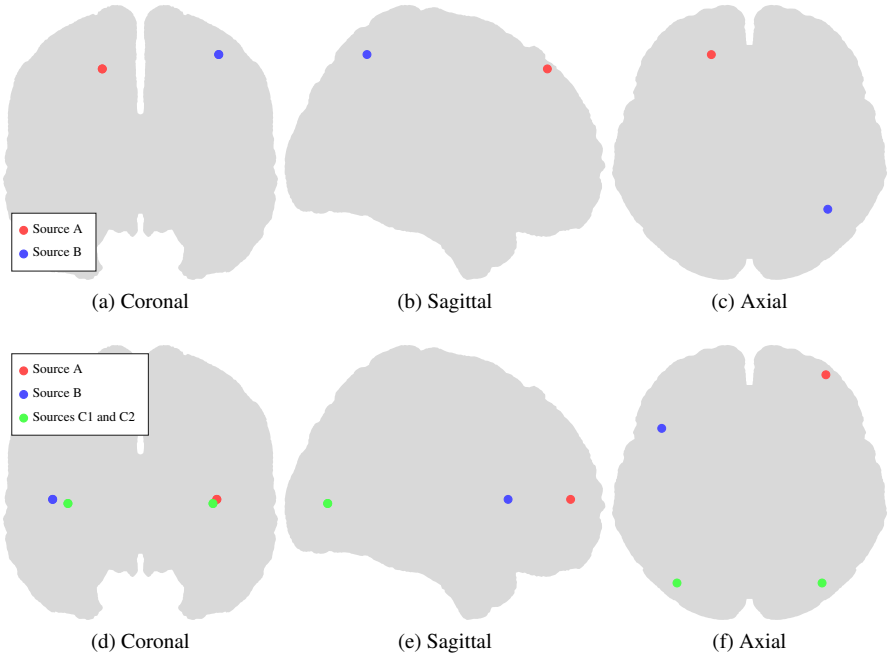
Signal type	**Amplitude** (mV)	**Latency** (ms)	**Width** (ms)	**Probability**
ERP-N	-0.6±0.1	350±20	120±10	80%
ERP-P	0.8±0.1	470±30	200±20	80%
Noise	0.55±0.1			100%

Table 3.1. Simulation parameters for the dataset SEREEGA-1.

Figure 3.2 plots the evoked potentials for each class over time. Each curve depicts a single EEG channel, from frontal to occipital brain regions. Although the generated evoked potentials simultaneously occur for both stimuli ($\sim$ 160 ms for ERP-N and $\sim$ 240 ms for ERP-P), their spatial location changes from class to class, representing different cognitive processes.

SEREEGA-2

This dataset simulates two stimuli evoking common and different electrical activity sources, named A, B, C1, and C2, with parameters and locations described in Table 3.2 and bottom of Figure 3.1. A and B are contra-lateral frontal sources, both holding one positive and one negative potential, while C1 and C2 emerge with a positive potential as the shared visual activity [46]. Therefore, the first stimulus (class) evokes sources A, C1, and C2 on 300 trials. On the contrary, the second stimulus evokes B, C1,



(a) Coronal (b) Sagittal (c) Axial

(d) Coronal (e) Sagittal (f) Axial

Figure 3.1. Location of sources for the datasets SEREEGA-1 (top) and SEREEGA-2 (bottom).

and C2 on 300 different trials. As a result, SEREEGA-2 holds M=600 trials with C=62 channels and T=250 time instants. Bottom of Figure 3.2 illustrates the evoked

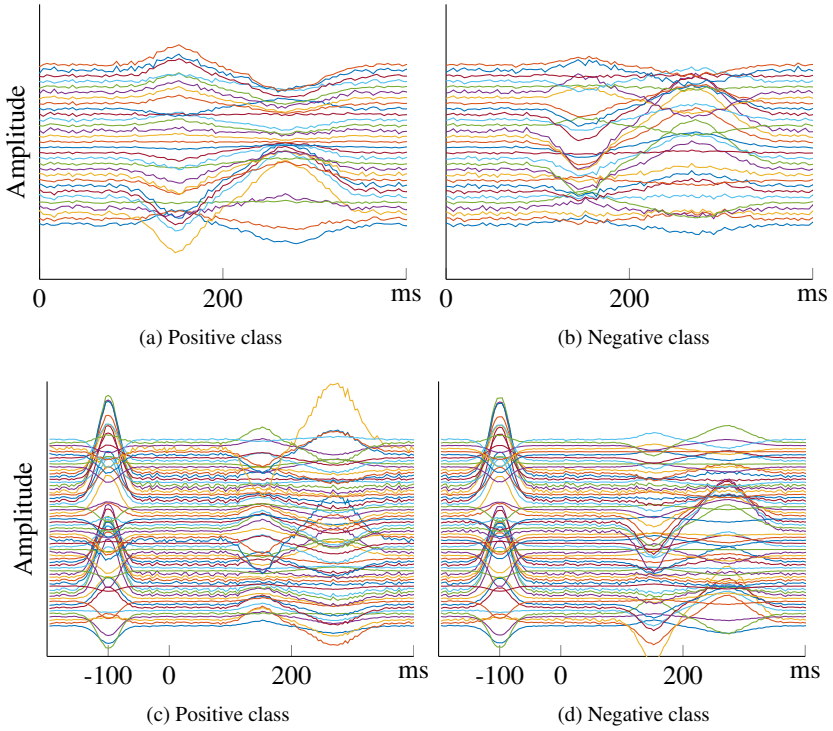
Signal type	**Amplitude** (mV)	**Latency** (ms)	**Width** (ms)	**Source**	**Probability**
ERP-N	–0.8±0.1	350±40	120±20	A,B	80%
ERP-P	1±0.1	470±30	200±30	A,B	80%
ERP-P	1±0.1	100±10	100±20	C1,C2	100%
Noise	0.2 ± 0.1			A,B,C1,C2	100%

Table 3.2. Simulation parameters for the dataset SEREEGA-2.

response potentials over time and channels. Sources C1 and C2, emerging at the visual cortex around -100 ms, behave similarly for both classes, as they emulate sensorial activation before the stimulus processing. In turn, A and B occur at the same time and different spatial location, representing stimulus processing at different areas.

3.2.2 Real EEG records

The real EEG dataset, termed ADHD-1, holds recordings from 69 children, aging between 7 and 12 y/o, labeled as either healthy control or attention deficit hyperactivity disorder (ADHD) patient. During EEG recording, subjects performed *Reward Stop-Signal Tasks* (RSST) by pressing a key each time they face a frequent stimulus called



(a) Positive class (b) Negative class

(c) Positive class (d) Negative class

Figure 3.2. Time courses for both classes on dataset SEREEGA-1 (top) and SEREEGA-2 (bottom).

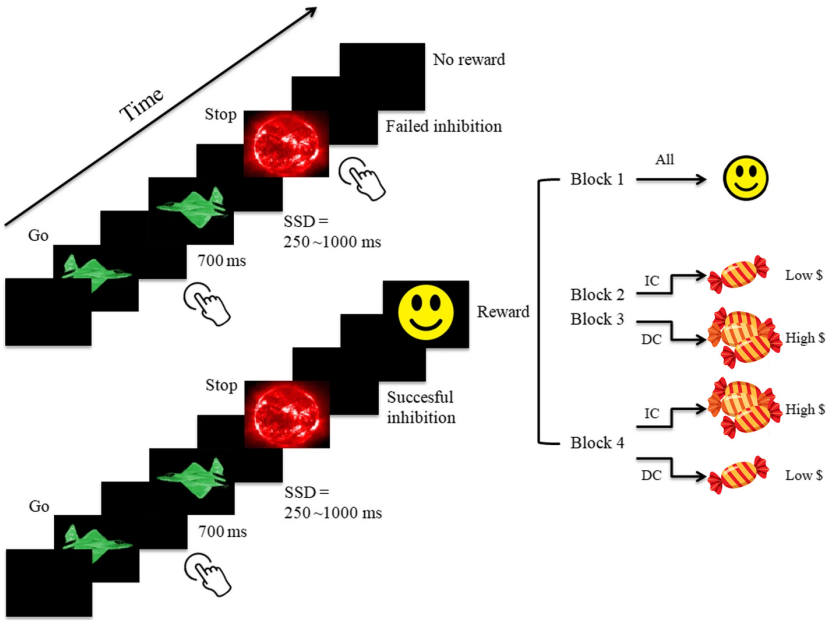


Figure 3.3. Reward Stop Signal Task (RSST) paradigm [1]

Go, unless being followed by an unfrequent *Stop* signal [1]. Each subject performed about 300 trials, triggered by the *Go*, distributed in four four-minutes blocks. Between each block, there is a lag period long enough for the subject to rest, but short to avoid the resting state. RSST rewards the subjects with a *Smiley* sticker, a *Low* amount of candies, or a *High* amount of candies if they successfully inhibit from pressing the key. When the monetary reward decreases from block to block, the child belongs to the decreasing condition group (DC). Otherwise, they compose the increasing condition group (IC). Table 3.3 summarizes the average number of successful and failed trials according the reward, condition and diagnosis. Regarding the time series details, EEG signals were recorded at 250 Hz and C=32 channels distributed over the scalp. Firstly, each trial is trimmed 200 ms before and 800 ms after the *Stop* stimulus, producing sequences of T=250 time instants.

Condition	**Diagnosis**	**N**	**Smiley**		**Low**		**High**	
			Succ	**Failed**	**Succ**	**Failed**	**Succ**	**Failed**
Decreasing	Control	15	20.28	14.90	13.46	18.76	27.05	40.42
	ADHD	15	15.76	17.02	13.39	22.72	25.99	43.12
Increasing	Control	21	13.81	19.64	26.45	42.44	13.29	20.73
	ADHD	18	12.65	20.92	24.09	45.66	13.59	20.96

Table 3.3. Average number of successful and failed inhibitions. **N** stands for the number of subjects in the diagnostic group.

3.2.3 Proposed scheme for feature extraction

Figure 3.4 presents the proposed scheme for the extraction of static discriminative features from EEG signals using discriminative decoding according to CSP for an RSST paradigm. The methodology begins by separating the trials into groups according to the reward within the paradigm. Second, CSP is applied within each group assuming successful and unsuccessful inhibitions as the two classes, resulting in a matrix for each reward. Finally, the discriminative capacity of each channel in each group is calculated using the resulting eigenvalues and spatial patterns according to Equation (3.5) to obtain the vector of characteristics that represent the subject.

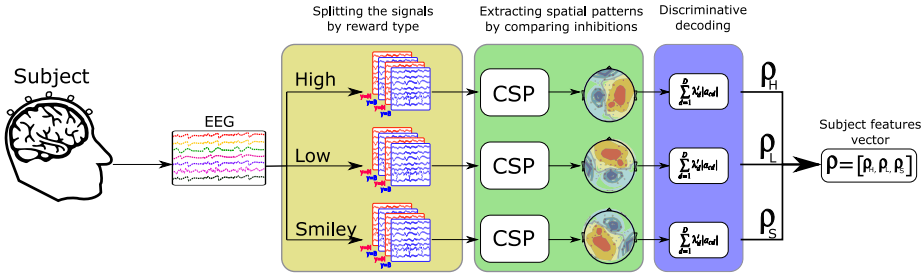
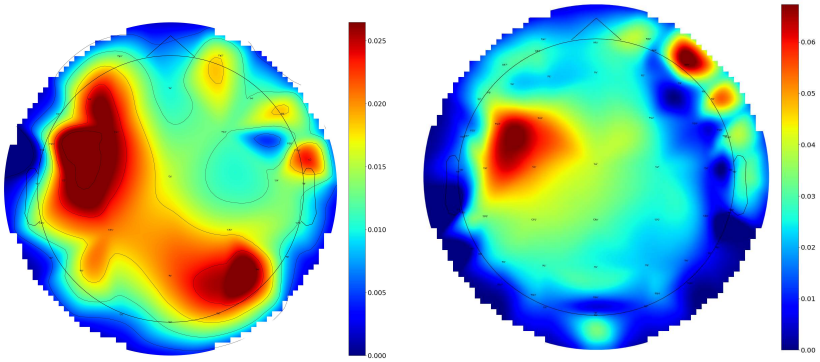


Figure 3.4. Proposed methodology for the feature extraction based on discriminative decoding of csp

3.3 Results

3.3.1 Discriminative decoding on simulated data

Figure 3.5a illustrates the topographic maps resulting from the proposed discriminative decoding for the SEREEGA-1 dataset. Red regions, denoting a larger ρ_c, are located over the simulated activity sources shown in Figure 3.1c. On the other hand, channels within blue and green regions, which are far from the activity sources, result in a lower discriminative index, indicating that those channels hardly distinguish both classes. Therefore, the proposed discriminative index highlights channels related to the underlying brain activity. On the other hand, the discriminative decoding for the SEREEGA-2, displayed in Figure 3.5b, highlights in red two contra-lateral frontal regions that are related to A and B sources. It is worth noting that C1 and C2 sources, over the visual cortex, hold low ρ_c values since they arise as a common activity in both classes. Such a fact is because the most discriminative spatial patterns (shown at the top of Figure 3.6) are associated with either source A or B. On the contrary, patterns explaining common activity hold eigenvalues close to 0.5 (bottom of Figure 3.6), meaning that they are not biased towards any class. As a result, the proposed discriminative decoding associates channels with the difference between two cognitive states, while ignoring the shared activity.



(a) SEREEGA-1 (b) SEREEGA-2

Figure 3.5. Topographic maps of the proposed discriminative coding applied on the synthetic EEG records.

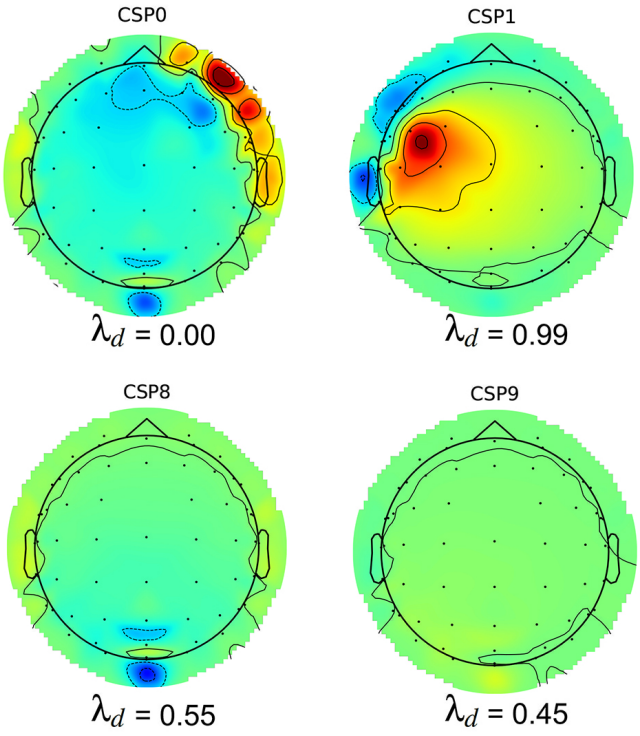


Figure 3.6. Four spatial patterns from the SEREEGA-2 dataset along their corresponding eigenvalues.

3.3.2 Feature extraction by discriminative decoding

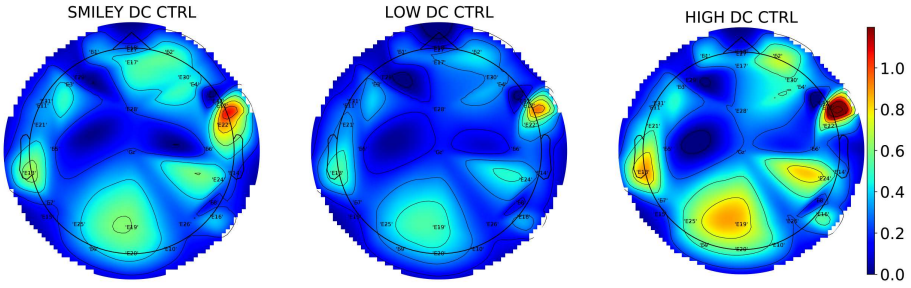
In real EEG, we applied the proposed approach subject-wise using successful and failed inhibitions as the two classes, taking into account that the ERN responses differ between ADHD and control children [8]. In this sense, the discriminative decoding provides a single feature vector at each reward for each subject, as the four patients in Figure 3.7 illustrate. Regarding the decreasing condition in Figures 3.7a and 3.7b, the ADHD subject evidence larger ρ_c values than the control in the frontal-central region, where studies found that the disorder affects the most the error-related wave [47]. On the other hand, applying the discriminative decoding on the increasing condition (Figures 3.7c and 3.7d) results in fewer differences between both subjects, indicating that a reward increment similarly stimulates both groups. Consequently, the proposed approach highlights the differences between cognitive states evoked by the paradigm condition.

3.3.3 Diagnostic support of ADHD

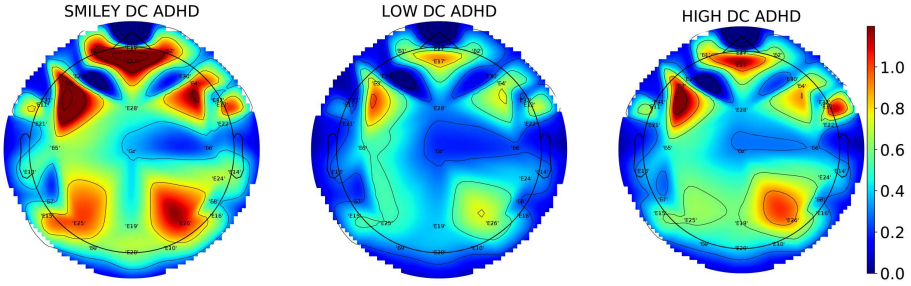
We validate the proposed approach as an ADHD biomarker by performing the classification of control versus ADHD subjects using our features from Section 3.3.2, ERN descriptors, and behavioral measures. For the former, we concatenate the 32 discriminative capacity indexes (one per channel) across the rewards, yielding 96 features per subject. For the second, we measure, at each reward, the amplitude and latency of the ERN wave at channels located over the medial frontal, left frontal, ventromedial orbitofrontal, and prefrontal cortices, known to evoke error-related responses, producing 54 features [48]. As behavioral measures, we consider the Mean Reaction time, the Stop Signal Delay, and the Stop Signal Reaction Time per block for each subject because they are related to ADHD [49], achieving 12 features. Subsequently, we apply a Principal Component Analysis (PCA) as dimension reduction, followed by a Linear Discriminant Analysis (LDA) as the classification machine, for each of the three feature sets.

Figure 3.8 illustrates the average classification accuracy as a function of the number of components (M) for each feature set and both IC and DC conditions. Regarding the decreasing condition, PCA increases the accuracy in comparison with no dimension reduction, independently on the feature set. Notably, the proposed discriminative decoding better discriminates the diagnostic groups by achieving an 87% accuracy. As a contrast, classification score little varies over the number of components when increasing the reward, implying that the dimension reduction lacks a significant effect on the diagnostic performance. As a result, under the proper condition, the proposed approach provides features improving the ADHD diagnostic compared to conventional approaches.

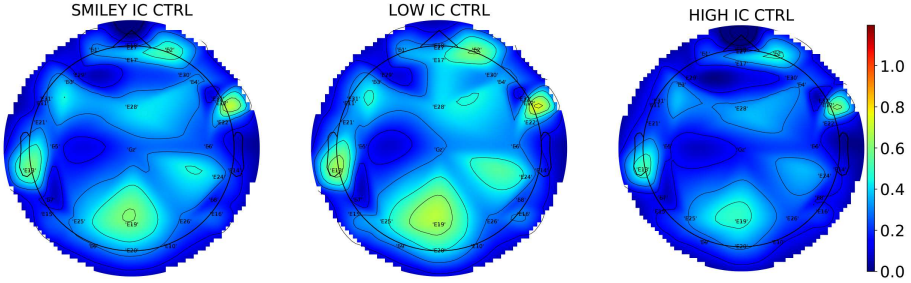
Table 3.4 presents the average performance achieved by each feature extraction approach at the optimal number of components in both conditions using a 10-fold cross-validation scheme. Also, we report the fold-paired t-test p-value comparing validation scores of the proposed decoding against behavior and ERN features at accuracy, sensitivity, specificity and F1-score metrics. Overall, DC achieves higher accuracy rates than IC because the former allows CSP to intensify differences between patterns from both



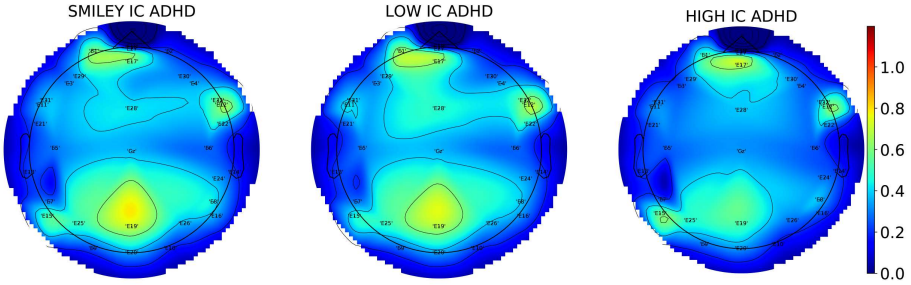
(a) Control DC



(b) ADHD DC

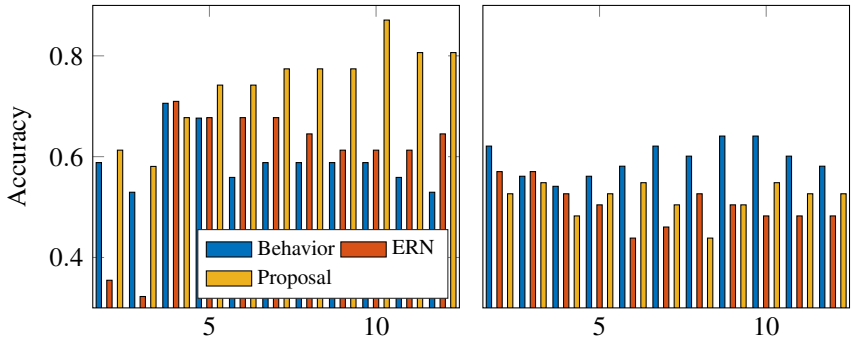


(c) Control IC



(d) ADHD IC

Figure 3.7. Topographic plots of the discriminative decoding on four real subjects.



(a) DC (b) IC

Figure 3.8. Tuning of the number of components for each condition reward and considered feature sets.

inhibitions. Particularly for decreasing condition, the discriminative decoding outperforms both conventional features sets, reaching a significant difference according to the hypothesis test (p-value <5%). However, p-values at IC are large enough to prove that the three features sets are statistically similar. Such findings agree with Figures 3.7c and 3.7d, where the difference between both subjects is hardly perceivable. Therefore, the discriminative decoding finds changes in the error-related response from control to ADHD children, highlighted by the decreasing condition, at clinically interpretable brain regions [7, 47].

		M	**Accuracy**	**Sensitivity**	**Specificity**	**F1-score**
DC	ERN	4	0.70±0.19*	0.70±0.40	**0.85±0.23**	0.65±0.35
	Behavior	4	0.71±0.15**	0.77±0.40	0.25±0.34**	0.67±0.37
	Proposal	10	**0.87±0.20**	**0.85±0.23**	**0.85±0.32**	**0.84±0.17**
IC	ERN	2	0.60±0.28	0.63±0.40	**0.70±0.40**	0.60±0.33
	Behavior	10	**0.72±0.13**	0.70±0.33	0.47±0.48	**0.68±0.32**
	Proposal	13	0.66±0.22	**0.75±0.33**	0.62±0.37	0.64±0.25

Table 3.4. Average performance for 10-fold cross-validation. Significant differences are marked for p-value $< 5\%$ (*) and $< 1\%$ (**). In bold highest score values

Chapter 4

Multiple Kernel Stein Spatial Patterns

This chapter proposes a methodology for the representation of EEG signals based on second-order statistics that favors the class separability and decodes non-linear relationships. For this, the proposal uses the similarity between time series through their covariance matrices in the Riemannian manifold of positive semidefinite matrices (PSD), using the logdet-divergence of Jensen Bregman, the Stein kernel, and Centered Kernel Alignment (CKA) as a cost function. To validate the representation, a methodology was designed for the classification of EEG signals termed Multiple Kernel Stein Spatial Patterns (MKSSP) dealing with noise, raveled brain activity, and subject variability issues. Experimental evaluations in the well-known four-class MI dataset 2a BCI competition IV proves that the methodology significantly improves state-of-the-art approaches. Further, the proposal is interpretable in terms of data distribution, spectral relevance, and spatial patterns. Such interpretability demonstrates that MKSSP encodes features from different spectral bands into a single representation improving the discrimination of mental tasks. The results of this chapter were presented as a paper. Evaluated and published in the journal Applied Sciences, by MDPI in December 2020 [50].

4.1 Methods

4.1.1 EEG Decomposition

Let a set of N labeled multichannel EEG time series (trials), acquired from a single subject $\mathcal{X}=\left\{\boldsymbol{x}_n(t)\in\mathbb{R}^C, y_n\in\mathcal{L}\right\}_{n=1}^{N}$, where C stands for the number of channels, $t\in[1,T]$ indexes the time instants, and y_n labels the n-th time series $\boldsymbol{x}_n(t)$. $\mathcal{L}$ defines the set of possible classes, usually related to mental states.

To take advantage of the spectral content of the EEG signals, each trial is band-passed through a set of B filters, achieving the filter-banked time series representation

of EEG data described in Equation (4.1):

$$\mathcal{X}_B = \left\{ \boldsymbol{x}_{nb}(t) = [x_{nc}(t) * h_b(t)]_{c=1}^{C}, b \in [1,B] \right\}_{n=1}^{N} \tag{4.1}$$

where $x_{nc}(t) \in \mathbb{R}$ stands for the c-th channel from the n-th trial, $h_b(t) \in \mathbb{R}$ corresponds to the impulse response of the b-th linear phase FIR filter, and $*$ denotes the convolution operator. Afterwards, a band-wise spatial filtering linearly mixes the input channels into components at each time instant resulting in a new set of band-filtered time series as follows:

$$\mathcal{X}_W = \left\{ \boldsymbol{z}_{nb}(t) = \boldsymbol{W}_b^\top \boldsymbol{x}_{nb}(t) : b \in [1,B] \right\}_{n=1}^{N} \tag{4.2}$$

with $\boldsymbol{W}_b \in \mathbb{R}^{C \times Q}$ as the linearly mixing matrix of spatial filters for the band b, and $\boldsymbol{z}_{nb}(t) \in \mathbb{R}^Q$ stands for the spatially-filtered trial with Q components.

4.1.2 Time-Series Similarity through the Stein Kernel for PSD Matrices

Since each trial in the component space is band-pass filtered, the expected value of $\boldsymbol{z}_{nb}(t)$ becomes zero. Hence, the band-wise covariance is computed as

$$\boldsymbol{S}_{nb} = \boldsymbol{z}_{nb}(t)\boldsymbol{z}_{nb}(t)^\top \tag{4.3}$$

$$\boldsymbol{S}_{nb} = \boldsymbol{W}_b^\top \boldsymbol{x}_{nb}(t)\boldsymbol{x}_{nb}(t)^\top \boldsymbol{W}_b \tag{4.4}$$

with $\boldsymbol{S}_{nb} \in \mathbb{R}^{Q \times Q}$. The set of covariance matrices from the dataset $\mathcal{X}$ holds the linear relationships between component pairs:

$$\mathcal{S} = \left\{ \boldsymbol{S}_{nb} = \boldsymbol{W}_b^\top \boldsymbol{\Sigma}_{nb} \boldsymbol{W}_b : b \in [1,B] \right\}_{n=1}^{N} \tag{4.5}$$

being $\boldsymbol{\Sigma}_{nb} = \boldsymbol{x}_{nb}(t)\boldsymbol{x}_{nb}(t)^\top \in \mathbb{R}^{C \times C}$ the trial covariance for the b-th band in the input channel space. Given that each matrix in the set $\mathcal{S}$ satisfies that $\langle \boldsymbol{A}, \boldsymbol{S}_{nb}\boldsymbol{A} \rangle \leq 0$ for any $\boldsymbol{A} \neq \boldsymbol{0}$, then $\boldsymbol{S}_{nb} \in \mathcal{S}$ is positive semidefinite (PSD). As a set of PSD matrices, $\mathcal{S}$ belongs to the Riemannian manifold $\mathbb{P}_Q$, which is differentiable and a canonical higher-rank symmetric space within the real symmetric matrix space $\mathbb{S}_Q$ [51]. Therefore, there exists a PSD matrix representing each time-series in a Hilbert space of matrices endowed with a metric allowing to compare two trials.

For such a manifold, the Jensen-Bregman divergence $\delta_F(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) \in \mathbb{R}^+$ measures the disimilarity between two matrix elements $\boldsymbol{S}_{nb}$ and $\boldsymbol{S}_{mb}$ as:

$$\delta(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) = \frac{B_F(\boldsymbol{S}_{nb}, \bar{\boldsymbol{S}}_b) + B_F(\bar{\boldsymbol{S}}_b, \boldsymbol{S}_{mb})}{2} \tag{4.6}$$

$$B_F(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) = F(\boldsymbol{S}_{nb}) - F(\boldsymbol{S}_{mb}) - \langle \boldsymbol{S}_{nb} - \boldsymbol{S}_{mb}, \nabla F(\boldsymbol{S}_{mb}) \rangle_F \tag{4.7}$$

$$\bar{\boldsymbol{S}}_b = \frac{\boldsymbol{S}_{nb} + \boldsymbol{S}_{mb}}{2} \tag{4.8}$$

where ∇F denotes the gradient of the strictly convex and differentiable function $F(\cdot)$, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \operatorname{tr}\left\{ \boldsymbol{A}^\top \boldsymbol{B} \right\}$ the Frobenius inner product in the PSD matrix space, and $\operatorname{tr}\{\cdot\}$

the trace operator. Equation (4.7) defines the Bregman divergence as the positive tail of the first-order Taylor expansion of $F(\cdot)$. Then, Equation (4.6) corresponds to the symmetrized version of B_F and holds the properties of a distance function in $\mathbb{P}_Q$ [52]. In particular, using as the convex function $F(\boldsymbol{S}_{nb}) = -\log|\boldsymbol{S}_{nb}|$ yields the Jensen-Bregman LogDet divergence defined in Equation (4.9), with $|\cdot|$ as the determinant operator.

$$D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) = \log\left|\frac{\boldsymbol{S}_{nb} + \boldsymbol{S}_{mb}}{2}\right| - \frac{\log|\boldsymbol{S}_{nb}\boldsymbol{S}_{mb}|}{2} \tag{4.9}$$

Moreover, Equation (4.9) is invariant to affine transformations, that is, the LogDet divergence at the component space is the same in the channel space if the number of components equals the number of channels ($Q = C$). The following procedure proves the affine-invariance, assuming $\boldsymbol{W} \in \mathbb{R}^{C\times C}$ square:

$$\begin{aligned}
D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) &= \log\left|\frac{\boldsymbol{S}_{nb} + \boldsymbol{S}_{mb}}{2}\right| - \frac{\log|\boldsymbol{S}_{nb}\boldsymbol{S}_{mb}|}{2} \\
D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) &= \log\left|\frac{\boldsymbol{W}^\top\boldsymbol{\Sigma}_{nb}\boldsymbol{W} + \boldsymbol{W}^\top\boldsymbol{\Sigma}_{mb}\boldsymbol{W}}{2}\right| - \frac{\log\left|\boldsymbol{W}^\top\boldsymbol{\Sigma}_{nb}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{\Sigma}_{mb}\boldsymbol{W}\right|}{2} \\
D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) &= \log\left|\boldsymbol{W}^\top\frac{\boldsymbol{\Sigma}_{nb} + \boldsymbol{\Sigma}_{mb}}{2}\boldsymbol{W}\right| - \frac{\log\left|\boldsymbol{W}^\top\boldsymbol{\Sigma}_{nb}\boldsymbol{W}\right|}{2} - \frac{\log\left|\boldsymbol{W}^\top\boldsymbol{\Sigma}_{mb}\boldsymbol{W}\right|}{2} \\
D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) &= \log\left|\boldsymbol{W}^\top\right| + \log\left|\frac{\boldsymbol{\Sigma}_{nb} + \boldsymbol{\Sigma}_{mb}}{2}\right| + \log|\boldsymbol{W}| \\
&\quad - \frac{\log\left|\boldsymbol{W}^\top\right|}{2} - \frac{\log|\boldsymbol{\Sigma}_{nb}|}{2} - \frac{\log|\boldsymbol{W}|}{2} \\
&\quad - \frac{\log\left|\boldsymbol{W}^\top\right|}{2} - \frac{\log|\boldsymbol{\Sigma}_{mb}|}{2} - \frac{\log|\boldsymbol{W}|}{2} \\
D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) &= \log\left|\frac{\boldsymbol{\Sigma}_{nb} + \boldsymbol{\Sigma}_{mb}}{2}\right| - \frac{\log|\boldsymbol{\Sigma}_{nb}\boldsymbol{\Sigma}_{mb}|}{2} \\
D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}) &= D(\boldsymbol{\Sigma}_{nb}, \boldsymbol{\Sigma}_{mb})
\end{aligned} \tag{4.10}$$

Thanks to the distance property in Equation (4.6), the LogDet divergence parameterizes a radial basis function to build a similarity measure between two trials in $\mathcal{X}_W$ through their corresponding covariances in the component space, that is:

$$k(\boldsymbol{x}_{nb}(t), \boldsymbol{x}_{mb}(t)|\boldsymbol{W}_b) = e^{-\gamma_b D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}|\boldsymbol{W}_b)} \tag{4.11}$$

being $\gamma_b \in \mathbb{R}^+$ the scale parameter for the kernel from the b-th frequency band. Equation (4.11) is known as the Stein kernel for PSD matrices. Since the linear mapping $\boldsymbol{W}_b$ determines the component space and the Stein kernel is affine-invariant, tuning of $\boldsymbol{W}_b$ demands an optimization procedure for $Q < C$ to enhance class separability at each band.

4.1.3 Spatial Filter Optimization Using Centered Kernel Alignment

Given the kernel function in Equation (4.11) and the vector of target labels $\boldsymbol{y} = \{y_n\}_{n=1}^{N}$, optimization of $\boldsymbol{W}_b$ is carried out by minimizing the negative logarithm of the Centered Kernel Alignment (CKA) cost function, assessing the similarity between two random variables through the inner product of their representing kernel matrices as defined by Equation (4.12).

$$L(\mathcal{X}_b, \boldsymbol{y}|\boldsymbol{W}_b) = -\log \frac{\left\langle \overline{\boldsymbol{K}}_b(\boldsymbol{W}_b), \overline{\boldsymbol{K}}_y \right\rangle_F}{||\overline{\boldsymbol{K}}_b(\boldsymbol{W}_b)||_F ||\overline{\boldsymbol{K}}_y||_F} \tag{4.12}$$

where kernel matrices $\boldsymbol{K}_b \in \mathbb{R}^{N\times N}$, with elements $k_{nm} = k(\boldsymbol{x}_{nb}(t), \boldsymbol{x}_{mb}(t)|\boldsymbol{W}_b)$, and $\boldsymbol{K}_y = \{\delta(y_n, y_m) : n, m = [1, N]\} \in \{0,1\}^{N\times N}$ hold every pair-wise trial similarity in the component Riemmanian manifold and in the label space, respectively. $||\boldsymbol{A}||_F = \sqrt{\langle \boldsymbol{A}, \boldsymbol{A} \rangle_F}$ corresponds to the Frobenius norm and $\delta(\cdot,\cdot)$ denotes the delta Dirac function. $\overline{\boldsymbol{K}}$ stands for the centered kernel matrix asociated to $\boldsymbol{K}$ and computed as in Equation (4.13), being $\boldsymbol{U}_N$ the centering matrix, $\boldsymbol{I}_N$ the N-order identity matrix, and $\boldsymbol{1}_N$ an all ones column vector. The centering operation translates all samples in the Hilbert space reproduced by the kernel $k(\cdot,\cdot)$ near the convex hull of the samples, so coping with ill-conditioning due to biased pair-wise inner products [53].

$$\overline{\boldsymbol{K}} = \boldsymbol{U}_N \boldsymbol{K} \boldsymbol{U}_N \tag{4.13}$$

$$\boldsymbol{U}_N = \boldsymbol{I}_N - \frac{1}{N} \boldsymbol{1}_N \boldsymbol{1}^{\top}{}_N \tag{4.14}$$

Aiming at minimizing the negative logarithm CKA between $\boldsymbol{K}_b$ and the label matrix $\boldsymbol{K}_y$, we consider a stochastic gradient descent approach with respect to spatial filter matrix $\boldsymbol{W}_b$. Using the chain rule, the gradient of the cost function w.r.t. $\boldsymbol{W}_b$ is expressed as:

$$\nabla L(\boldsymbol{W}_b) = \sum_{n=1}^{N} \sum_{m=1}^{N} \frac{\partial L}{\partial k_{nm}} \frac{\partial k_{nm}}{\partial D_{nm}} \frac{\partial D_{nm}}{\partial \boldsymbol{W}_b} \tag{4.15}$$

$$\frac{\partial L}{\partial \boldsymbol{K}_b} = -\frac{\overline{\boldsymbol{K}}_y}{\left\langle \overline{\boldsymbol{K}}_b, \overline{\boldsymbol{K}}_y \right\rangle_F} + 2\frac{\overline{\boldsymbol{K}}_b}{\left\langle \overline{\boldsymbol{K}}_b, \overline{\boldsymbol{K}}_b \right\rangle_F} \tag{4.16}$$

$$\frac{\partial k_{nm}}{\partial D_{nm}} = -\beta e^{-\gamma_b D(\boldsymbol{S}_{nb}, \boldsymbol{S}_{mb}|\boldsymbol{W}_b)} = -\gamma_b k_{nm} \tag{4.17}$$

$$\begin{aligned} \frac{\partial D_{nm}}{\partial \boldsymbol{W}_b} &= \boldsymbol{\Sigma}_{nb} \boldsymbol{W}_b [2(\boldsymbol{W}_b^{\top}(\boldsymbol{\Sigma}_{nb} + \boldsymbol{\Sigma}_{mb})\boldsymbol{W}_b)^{-1} - (\boldsymbol{W}_b^{\top} \boldsymbol{\Sigma}_{nb} \boldsymbol{W}_b)^{-1}] \\ &+ \boldsymbol{\Sigma}_{mb} \boldsymbol{W}_b [2(\boldsymbol{W}_b^{\top}(\boldsymbol{\Sigma}_{nb} + \boldsymbol{\Sigma}_{mb})\boldsymbol{W}_b)^{-1} - (\boldsymbol{W}_b^{\top} \boldsymbol{\Sigma}_{mb} \boldsymbol{W}_b)^{-1}] \end{aligned} \tag{4.18}$$

Therefore, maximizing the alignment between samples and their label kernels yields a matrix $\boldsymbol{W}_b$ rotating the channel data so that the Riemmanian space of component covariances better discriminates the given classes in the frequency band b.

4.1.4 Assembling of Multiple Kernel Representations

Given the kernel function in Equation (4.11) providing a metric for a frequency band, the knowledge from each band can be merged into a single metric using a multiple

kernel learning (MKL) approach. For the set of matrices $\{\boldsymbol{K}_b \in \mathbb{R}^{N\times N}\}_{b=1}^{B}$, the linear combination in Equation (4.19) corresponds to a weighted concatenation of the Hilbert space features reproduced by each kernel [54].

$$\boldsymbol{K}_{\mu} = \sum_{b=1}^{B} \mu_b \boldsymbol{K}_b \tag{4.19}$$

$$k_{\boldsymbol{\mu}}(\boldsymbol{x}_n(t), \boldsymbol{x}_m(t)) = \sum_{b=1}^{B} \mu_b k(\boldsymbol{x}_{nb}(t), \boldsymbol{x}_{mb}(t) | \boldsymbol{W}_b) \tag{4.20}$$

Aiming to preserve the positive definiteness of a kernel function, the vector of weights $\boldsymbol{\mu} = \{\mu_b\}_{b=1}^{B}$ must belong to subset of positive one-valued norm vectors, that is, $\boldsymbol{\mu} \in \{\boldsymbol{\mu} \in \mathbb{R}^B : ||\boldsymbol{\mu}||_2 = 1, \boldsymbol{\mu} \geq \mathbf{0}\}$. Finding the optimal $\boldsymbol{\mu}$ is posed as the maximization of the CKA cost between the kernel matrices $\boldsymbol{K}_{\boldsymbol{\mu}} = \{k_{\boldsymbol{\mu}}(\boldsymbol{x}_n(t), \boldsymbol{x}_m(t))\}_{nm=1}^{N}$ and $\boldsymbol{K}_y$ while constraining the weight values as follows:

$$\begin{aligned} \max_{\boldsymbol{\mu}\in\mathbb{R}^B} & \frac{\langle \overline{\boldsymbol{K}}_{\boldsymbol{\mu}}, \overline{\boldsymbol{K}}_y \rangle_F}{||\overline{\boldsymbol{K}}_{\boldsymbol{\mu}}||_F ||\overline{\boldsymbol{K}}_y||_F} \\ s.t. & ||\boldsymbol{\mu}||_2 = 1 \\ & \boldsymbol{\mu} \geq \mathbf{0} \end{aligned} \tag{4.21}$$

The optimization problem in Equation (4.21) is rewritten as a straightforward quadratic optimization problem with linear constraints [54]:

$$\begin{aligned} \min_{\mathbf{v}\in\mathbb{R}^B} & \mathbf{v}^\top \boldsymbol{M} \mathbf{v} - 2\mathbf{v}^\top \boldsymbol{a} \\ s.t. & \mathbf{v} \geq \mathbf{0} \end{aligned} \tag{4.22}$$

where vector $\boldsymbol{a} = \left\{\left\langle \overline{\boldsymbol{K}}_b, \overline{\boldsymbol{K}}_y \right\rangle_F\right\}_{b=1}^{B} \in \mathbb{R}^B$ and matrix $\boldsymbol{M} = \left\{\left\langle \overline{\boldsymbol{K}}_b, \overline{\boldsymbol{K}}_{b'} \right\rangle_F\right\}_{bb'=1}^{B} \in \mathbb{R}^{B\times B}$ account for the alignment with the supervised information and between the input kernels, respectively. Lastly, the solution weighting vector is achieved by normalizing $\mathbf{v}$ as:

$$\boldsymbol{\mu} = \frac{\mathbf{v}}{||\mathbf{v}||_2} \tag{4.23}$$

Therefore, the weights given by Equation (4.23) rank each frequency band according its similarity with the provided labels, so that the smaller the value, the leaser the contribution to build the kernel in Equation (4.19). Further, the introduced MKL gathers the band-wise covariance matrices from different Riemannian manifolds into a single supervised reproduced kernel Hilbert space favoring separability of mental states.

4.2 Experimental Setup

4.2.1 Dataset IIa from BCI Competition IV (BCICIV2a)

To test our proposal, we use the dataset IIa of the BCI Competition IV, provided by the BCI Laboratory at the Graz University of Technology (`http://www.bbci.de/`

competition/iv/). The dataset BCICIV2a comprises EEG trials from nine subjects while executing four motor imagery tasks, namely, left hand, right hand, foot, and tongue. The participants imagined each movement 72 times following a visual cue, resulting in $N = 288$ trials. For each trial, twenty-two EEG electrodes ($C = 22$) distributed over the scalp (as seen in Figure 4.1a) recorded the brain activity at a sampling frequency of 250 Hz during six seconds. The trial started with an auditory signal, followed by a black fixation cross, that warned the subject for the upcoming cue. An arrow pointing towards the movement during 1.25 s indicated the task to perform. The trial ended four seconds after the fixation cross disappeared as Figure 4.1b shows. A break period of about 1.5 s allows the participant to rest before starting the next trial. Taking into account the cue onset and the short-lasting imagined movement, the methodology classifies each recording only using the period between 2.5 and 4.5 s, that is, $T = 500$ samples.

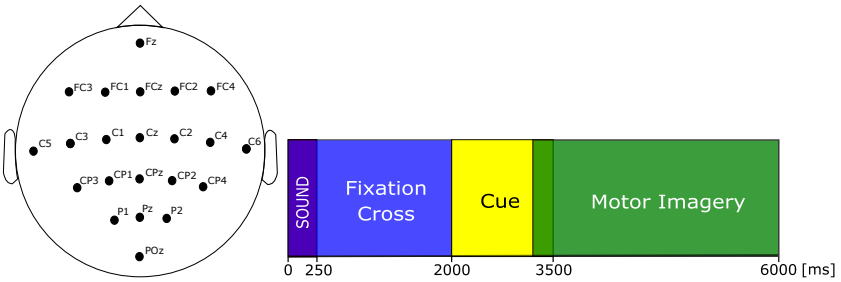


Figure 4.1. BCI competition IV acquisition setup. (**a**) EEG montage. (**b**) Paradigm time scheme.

4.2.2 Proposed BCI Methodology

Figure 4.2 illustrates the five stages of MKSSP. Firstly, $B = 17$ band-pass filters decompose each trial using five-ordered Butterworth filters of 4 Hz bandwidth and overlapping 2 Hz within [4,40] Hz [55]. Secondly, the spatial filters, optimized in Section 4.1.3, project the band-wise covariance matrix into the lower-dimensional Riemannian manifold $\mathbb{P}_Q$. Thirdly, the Stein kernel assesses the similarity between the projected test and training covariances, relying on Equation (4.11). Further, the multikernel linearly combines the 17 similarities for a test trial into a single kernel value. Lastly, a support vector machine, fed by the learned kernel, labels the recordings into one of the four classes according to a one-vs-one scheme.

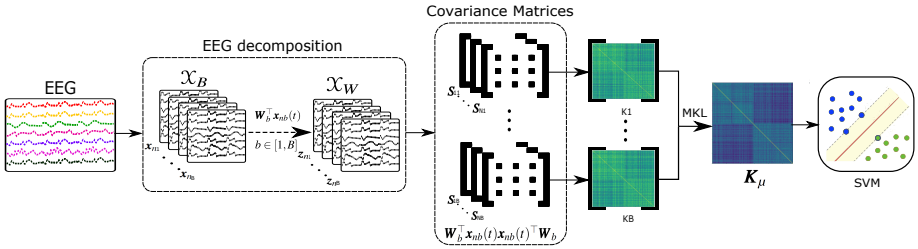


Figure 4.2. Proposed EEG classification methodology based on filter-banked Stein-kernel trial similarities.

The proposed methodology contains three sets of parameters, namely, the band-wise spatial filters, the kernel bandwidth, and the kernel mixing weights. For the former two, the gradient descent algorithm, updating the parameter according to Equation (4.18), finds the channel-to-component space projections and gamma values enhancing class separability at each band. For the latter, the solution to the quadratic programming problem in Equation (4.22) defines the contribution of each spectral component to the supervised discrimination task. Besides, the single hyperparameter of the proposal corresponds to Q, defining the number of columns for matrices $\boldsymbol{W}_b$, that is, the number of spatial filters. A cross-validated grid-search approach fixes the optimal Q by maximizing the classification performance within the range $[2, 10]$. Specifically, we assess the performance through Cohen's kappa score defined in Equation (4.24), where p_o corresponds to the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e determines the expected agreement at classification by chance.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.24}$$

4.3 Results

4.3.1 Performance Results

Figure 4.3 presents the subject-wise grid search results for tuning the number of spatial filters. Tuning curves evidence that performance decreases for either a very low or very high number of components. In the first case, the resulting model misses class separability patterns, due to the low flexibility in the lowest dimensional manifolds. In the second case, the methodology overfits the training data, since the component space becomes sparse, then drops the validation performance. Since four to eight components improve the classification performance in comparison to the input channel space, the proposed methodology reduces Riemmannian manifold dimension while enhancing MI discrimination.

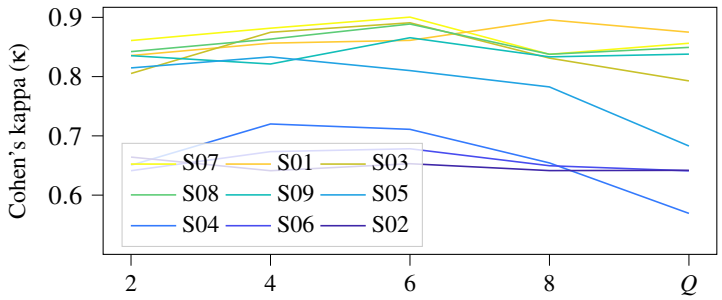


Figure 4.3. Cross-validated classification performance along the number of components.

Table 4.1 presents the mean kappa value for a 5 fold cross-validation scheme using different methods tested in the BCICIV2a dataset, with the highest kappa value in bold for each subject. The last column holds the p-value for a paired t-test between the proposed and each compared approaches. Subjects are sorted according the score attained by the challenge winner. Note that MKSSP outperforms other approaches at the worst performing subjects (namely S06, S05, S02, and S01). Specifically, the proposal favors the most the subject S05, as the kappa score raises between 5% and 57%. Despite MKSSP performs at subjects S09, S03, S08, and S07, the best approach on such subjects are considerably biased towards them. Furthermore, MKSSP reaches the highest average kappa score of 0.82, with significant differences in most of the comparisons. Hence, the proposed methodology collects and combines discriminative features from different spaces in a single representation, enhancing the separability of different mental tasks.

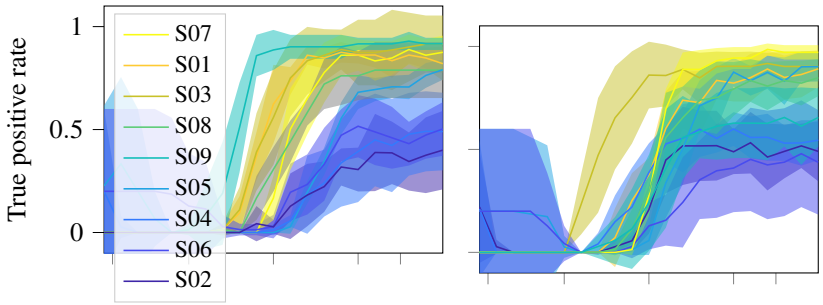
Table 4.1. Mean kappa scores attained by compared approaches at each subject from the BCICIV2a dataset. The last two columns present the average kappa and the t-test p-value between Multi-Kernel Stein Spatial Patterns (MKSSP) and the corresponding approach. In bold highest kappa values, in italic p-values $< 5\%$.

Approach	**S06**	**S05**	**S02**	**S04**	**S01**	**S09**	**S03**	**S08**	**S07**	κ	***p*-Value**
Challenge winner [56]	0.27	0.40	0.42	0.48	0.68	0.61	0.75	0.75	0.77	0.57 ± 0.17	*0.0002*
SUSS-SRKDA [57]	0.35	0.56	0.51	0.68	0.83	0.75	0.88	0.84	0.90	0.70 ± 0.18	*0.0179*
CBN [58]	0.42	0.78	0.51	**0.85**	0.69	0.45	0.87	**0.97**	0.54	0.68 ± 0.19	0.0577
KPCA with CILK [59]	0.37	0.26	0.46	0.44	0.71	0.61	0.76	0.75	0.79	0.57 ± 0.18	*0.0009*
PSO [60]	0.53	0.62	0.62	0.77	0.87	0.76	**0.90**	0.82	0.80	0.74 ± 0.12	*0.0282*
CSP-FLS [61]	0.37	0.35	0.54	0.52	0.74	0.80	**0.90**	0.86	0.82	0.66 ± 0.20	*0.0146*
EMD+Riemann [62]	0.34	0.36	0.24	0.68	0.86	0.82	0.70	0.75	0.66	0.60 ± 0.21	*0.0050*
CSP/AM-BA-SVM [63]	0.41	0.58	0.55	0.60	0.87	0.80	0.89	0.84	0.88	0.71 ± 0.17	*0.0147*
Dempster–Shafer [64]	0.57	0.67	0.59	0.72	0.78	0.88	0.85	0.86	0.81	0.75 ± 0.11	*0.0084*
Functional brain [65]	0.61	0.63	0.54	0.70	0.77	0.86	0.84	0.84	0.77	0.73 ± 0.11	*0.0036*
CNN-LSTM [66]	0.66	0.77	0.54	0.78	0.85	**0.90**	0.87	0.83	**0.95**	0.80 ± 0.12	0.3313
sDPLM [20]	0.36	0.34	0.49	0.49	0.75	0.76	0.76	0.76	0.68	0.60 ± 0.17	*0.0008*
uDPLM [20]	0.36	0.30	0.49	0.47	0.76	0.76	0.76	0.76	0.69	0.59 ± 0.18	*0.0012*
Proposed MKSSP	**0.68**	**0.83**	**0.66**	0.72	**0.90**	0.87	0.89	0.89	0.90	$\mathbf{0.82 \pm 0.09}$	–

To evaluate the proposal sensitivity to noise, we performed a nested cross-validation test by adding several noise levels to the EEG signals. For each subject, 70% of the trials were considered to learn the model parameters at the optimal number of components using a five fold inner cross-validation. The remaining 30% was firstly contaminated

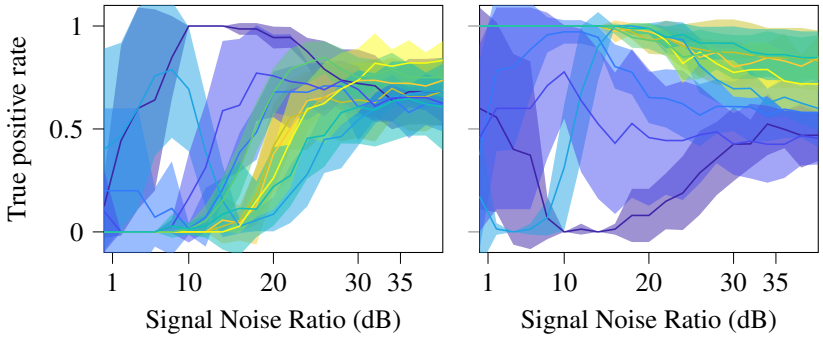
with additive Gaussian noise from 0 dB to 40 dB Signal-to-Noise Ratios (SNR). Then, the class-wise performance was computed from the contaminated trials while varying the SNR. Figure 4.4 illustrates the resulting class-wise true positive ratio (TPR) with the subjects ordered from best to worst performance, as in Figure 4.3.

Overall, the proposed MKSSP reaches a stable performance from 30 dB. Within 20 to 30 dB, most of the subjects and classes increase their TPR to reach the optimal behavior. However, the overall performance drops for SNR worse than 20 dB. At the subject level, the best performing ones yield the same behavior for all classes, that is, a low TPR for SNR $<$ 20 dB and the optimal TPR for SNR $>$ 30 dB. On the contrary, the worse performing subjects exhibit a TPR biased towards Foot and Tongue when the SNR $<$ 20 dB. For SNR $>$ 30 dB, the hand-related classes evidenced higher TPR values for the best performing subjects than for the worse ones. Such findings suggest that the additive noise mostly hinders the discriminative patterns related to Left and Right. In turn, the Foot class presents a similar performance for SNR $>$ 30 dB among all the subjects, implying patterns that are harder to reveal. Therefore, the proposed classification methodology reliably operates at SNR levels better than 30 dB with less biased results.



(a) (b)

Figure 4.4



(c) (d)

Figure 4.4. Noise sensitivity test per class for additive Gaussian noise using the optimal parameter set. (**a**) Left class. (**b**) Right class. (**c**) Foot class. (**d**) Tongue class.

4.3.2 Model Interpretability

We evaluate the model interpretability from three points of view: the data, spectral, and spatial representations. For analyzing the data representation, Figures 4.5 and 4.6 illustrate the three-dimensional mapping of K_μ using Kernel Principal Component Analysis (KPCA) for the best- (S07) and worst-performing (S02) subjects, respectively. At the first two KPCA components of S07, MKSSP discriminated Left and Right from Foot and Tongue classes. Those two components also allow classifying between Left and Right, whereas the third one separated Foot and Tongue. Despite being less evident, MKSSP identified three trial groups from Subject S02, namely Foot, Tongue, and hands. The first two KPCA components better separated Foot from the remaining classes, while the first and third ones enhanced Tongue discrimination.

As Table 4.1 presented, subject S05 benefited the most from the proposed MKSSP, in comparison with the challenge winner. For such a subject, Figure 4.7 compares the trial data distribution resulting from four approaches: CSP, with a single 8-to-30 Hz band-pass filter, and Filter-Banked CSP (FBCSP), with filters as in Section 4.2.2, representing the baseline spatial-filtering and spectro-spatial-filtering techniques. MKSSP without spatial filtering (i.e., $\boldsymbol{W}$ is the $C \times C$ identity matrix $\boldsymbol{I}_C$) and with the optimal number of components $Q^* = 4$ contrast the former approaches as spectral and spectro-spatial representations. For the sake of visualization, PCA and KPCA map trial features to two dimensions for CSP-bases and MKSSP-based representations, respectively. It is worth noting that CSP features without spectral filtering (top-left) and MKSSP without spatial filtering (bottom left) lacked any class separability. In turn, FBCSP separated Left and Right but mixed Foot and Tongue within the four classes. Lastly, optimal MKSSP not only increased Left and Right distance but also unraveled Foot and Tongue.

Regarding the spectral representation, Figure 4.8 illustrates the MKL weights of each band per subject, descending-ordered on the Y-axis (from best to worst based on the kappa score). A visual inspection of spectral weights segment subjects into

two groups: Subjects S07, S01, S03, S08, and S09; and subjects S05, S04, S06, and S02. The first group, holding the best performing subjects, concentrates weights at frequencies lower than 26Hz. Besides, the reduced number of highlighted spectral bands translates into a computational burden reduction. The second group, with kappa lower than 0.83, widely spreads all over the signal bandwidth, without accentuating any spectral filter. Accordingly, the weight distribution marks the MKSSP performance on the subject.

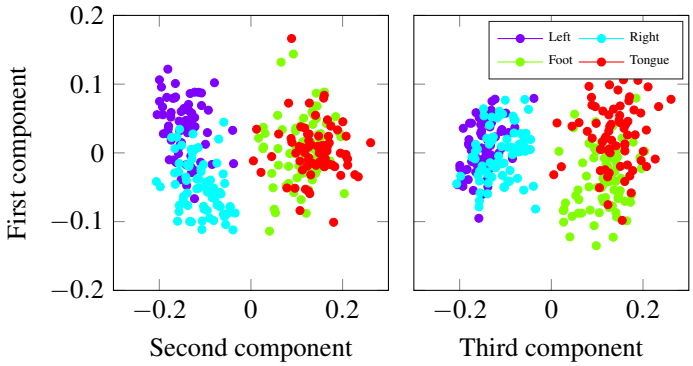


Figure 4.5. Resulting MKSSP kernel projected into three Kernel Principal Component Analysis (KPCA) components for the best performing subject (S07).

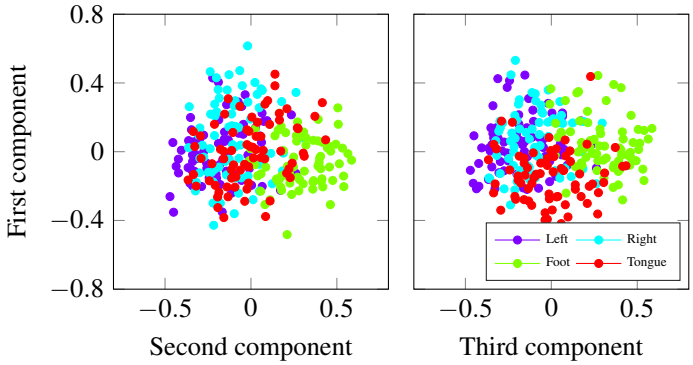


Figure 4.6. Resulting MKSSP kernel projected into three KPCA components for the worst-performing subject (S02).

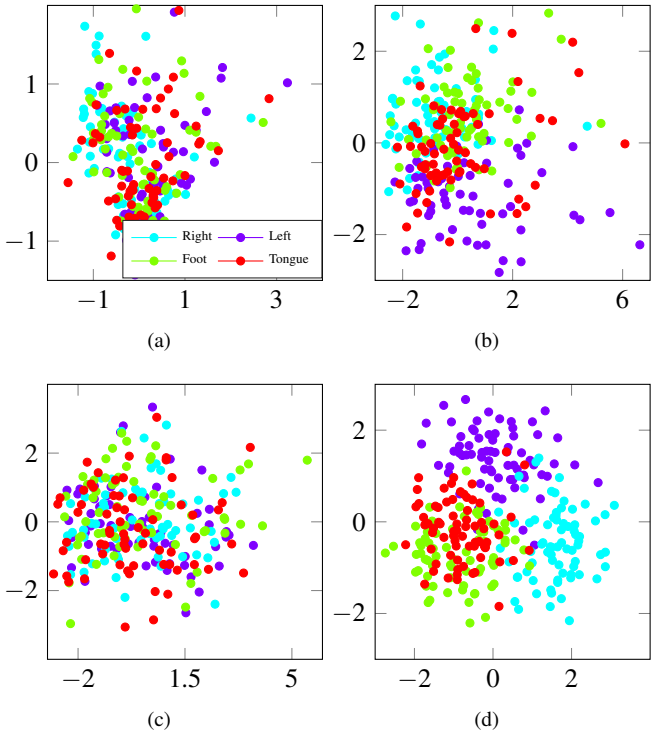


Figure 4.7. Projections of subject S05 trials using four spatial patterns approaches. Either PCA or KPCA maps to a 2D space features from (FB) CSP or MKSSP, respectively. (**a**) CSP. (**b**) FBCSP. (**c**) MKSSP ($\boldsymbol{W} = \boldsymbol{I}_C$). (**d**) MKSSP ($Q^* = 4$).

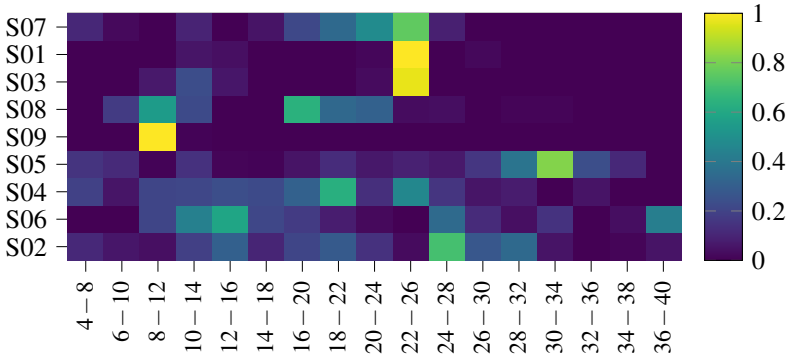


Figure 4.8. Multiple kernel learning (MKL) kernel weights per frequency band. Subjects are sorted according to the kappa score from the best (top) to the worst (bottom) in Y-axis. X-axis indexes each band-pass filter in the bank.

For the spatial representation, Figure 4.9 depicts the spatial patterns of the best and worst performing subjects for MKSSP (S07 and S02, respectively) and the most

favored one (S05), since such topographic maps represent the projection of estimated brain activity sources to electrodes [12]. For each subject, the top and bottom row hold the least and most weighted frequency band, respectively, while columns sort the first four components. For subject S07, the first two columns highlight a contralateral activity localized over the motor cortex region at the most relevant frequency band, which may be associated with the Right and Left classes [43]. On the contrary, the subject S02 lacks any activity over such cortex, yielding patterns without MI interpretation. In addition, the bottom frequency band of subject S05 relates the more to imagined movements than the top one, since the first two components better localize brain activity over hand movement regions. Lastly, the activity concentrated over the center vertex stands out in the third component and bottom band of all the subjects and explains Foot movement imagination [43].

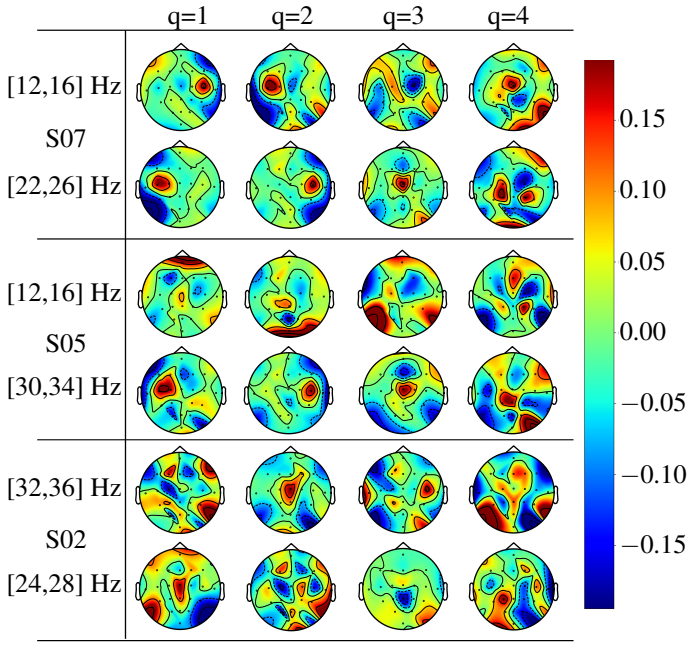


Figure 4.9. First four spatial patterns of the best performing (S07), worst performing (S02), and most improved (S05) subjects computed from MKSSP. For each subject, the top and bottom row hold the least and most weighted frequency band, respectively.

Chapter 5

SSP-based discriminative capacity index from EEG supporting ADHD diagnosis

This chapter presents a methodology to support the diagnosis of ADHD from EEG signals, which takes into account spectral information and mental states within the RSST paradigm using second-order statistics. For this, the proposal uses the index of discriminative capacity proposed in Chapter 3, the optimized spatial filter methodology proposed in Section 4.1.3, and an EEG decomposition in brain rhythms. The resulting features quantify the discriminative capacity of each channel in different brain rhythms using the Stein spatial patterns and eigenvalues. To validate the subject-wise features, the database from Section 3.2.2 was used, refining the number of spatial filters and feeding a linear discriminant analysis as a diagnostic support tool.

Figure 5.1 illustrates the five stages scheme for the feature extraction from EEG signals under the RSST paradigm. Firstly, the subject's EEG signals are split by the reward type. Secondly, an EEG decomposition is performed in the brain rhythms presented in Section 5.1.1, obtaining a total of 5 sets of frequency band signals for each type of reward. Thirdly, SSP is applied using the inhibition type (Good or Bad inhibition) as a class label to obtain a set of spatial filters for each frequency band and type of reward. Then, the SSP results are used to apply the discriminative decoding of Section 5.1.3, obtaining 32 features for each band, one for each channel, and a total of 160 characteristics for the type of reward. Finally, the 480 resulting features for each subject are concatenated in a single vector.

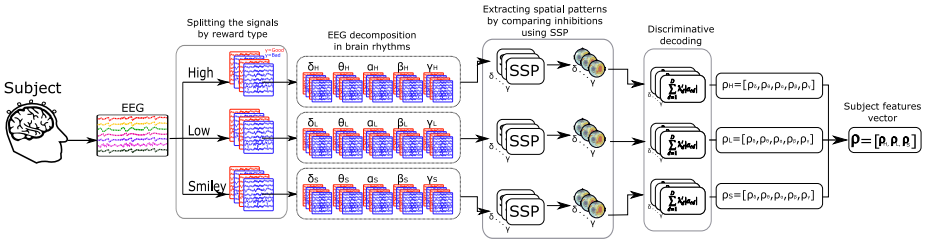


Figure 5.1. Proposed EEG feature extraction methodology based on Stein spatial patterns

In this chapter, the database described in Section 3.2.2 was used. A real EEG dataset, termed ADHD-1, holds recordings from 69 children, aging between 7 and 12 y/o, labeled as either healthy control or attention deficit hyperactivity disorder (ADHD) patient.

5.1 Methods

5.1.1 Brain rhythms EEG decomposition

Similar to the frequency decomposition presented in Section 4.1.1, this chapter performs a decomposition of the set of signals $\mathcal{X}=\{\boldsymbol{x}_n(t)\in\mathbb{R}^C, y_n\in\mathcal{L}\}_{n=1}^{N}$, where C stands for the number of channels, $t\in[1,T]$ indexes the time instants, and y_n labels the n-th time series $\boldsymbol{x}_n(t)$. $\mathcal{L}$ defines the set of possible classes, in this case, the type of inhibition within the paradigm.

Since in the state of the art some proposals explore specific brain rhythms to characterize ADHD, as observed in the theta/beta ratio [32], or the exploration of spectral power values in different brain rhythms such as alpha, beta, theta and delta [67]. Our proposal tries to exploit the spectral content of the signals filtering each trial using a set of 5 filters as described in Equation (5.1) and Table 5.1. Where $x_{nc}(t)\in\mathbb{R}$ stands for the c-th channel from the n-th trial, $h_b(t)\in\mathbb{R}$ corresponds to the impulse response of the b-th linear phase FIR filter, and $*$ denotes the convolution operator.

B	Rythm	Low Cut Frequency [Hz]	High Cut Frequency [Hz]
1	δ	1	4
2	θ	4	8
3	α	8	12
4	β	13	30
5	γ	30	70

Table 5.1. Frequency of brain rhythms

$$\mathcal{X}_B=\{\boldsymbol{x}_{nb}(t)=[x_{nc}(t)*h_b(t)]_{c=1}^{C}, b\in[1,B]\}_{n=1}^{N} \tag{5.1}$$

5.1.2 Stein Spatial Patterns (SSP)

Given the cost function of section 4.1.3, where the optimization of the $\boldsymbol{W}_b$ matrix is proposed by minimizing the negative logarithm of the CKA function of Equation (4.12), where the kernel matrices $\boldsymbol{K}_b \in \mathbb{R}^{N\times N}$, with elements $k_{nm} = k(\boldsymbol{x}_{nb}(t), \boldsymbol{x}_{mb}(t)|\boldsymbol{W}_b)$, that correspond to the pair-wise trial Stein kernel, and $\boldsymbol{K}_y = \{\delta(y_n, y_m) : n,m = [1,N]\} \in \{0,1\}^{N\times N}$ hold every trial similarity in the component Riemmanian manifold in the label space. Solving the optimization problem using the stochastic gradient descent approach of Equation (4.15), yields a matrix $\boldsymbol{W}_b$, where their columns correspond to spatial filters that rote the channel data to an uncorrelated component space for the signal labels.

$$L(\mathcal{X}_b, \boldsymbol{y}|\boldsymbol{W}_b) = -\log \frac{\left\langle \overline{\boldsymbol{K}}_b(\boldsymbol{W}_b), \overline{\boldsymbol{K}}_y \right\rangle_F}{||\overline{\boldsymbol{K}}_b(\boldsymbol{W}_b)||_F ||\overline{\boldsymbol{K}}_y||_F}$$

The spatial filters eigenvalues are obtained from the diagonal of the projected mean covariance matrix of the positive class, as shown in Eq. (5.2). Where C_+ is the mean covariance matrix of the positive class in the original space, S_+ is the projection of the matrix C_+ using the spatial filters, and λ_+ the associated eigenvalues.

$$C_+ = \frac{\sum_{n=1}^{N} cov(x_{nb})/tr\{x_{nb}\}}{N} | y_n = + \tag{5.2}$$

$$S_+ = \boldsymbol{W}_b C_+ \boldsymbol{W}_b^\top \tag{5.3}$$

$$\lambda_+ = diag(S_+) \tag{5.4}$$

$$\tag{5.5}$$

5.1.3 Discriminative decoding of SSP

Analogously to the decoding presented in section 3.1.2 for CSP, given a projected trial $\boldsymbol{z}_{nb}(t) = \boldsymbol{W}_b^\top \boldsymbol{x}_{nb}(t)$, the input time series can be re-obtained using a second linear mapping, $\boldsymbol{x}_{nb}(t) = \boldsymbol{A}_b \boldsymbol{z}_{nb}(t)$. Since SSP (same as CSP) provides the spatial filters, such a matrix is computed in the same way as $\boldsymbol{A}_b = \boldsymbol{W}_b^{-\top}$. The columns of $\boldsymbol{A}_b \in \mathbb{R}^{C\times D}$, are then the Stein spatial patterns for the band b.

Further, the resulting eigenvalues in Equation (5.2) are bounded to the range $\lambda \in [0,1]$. If λ_d=1, variances in the $\boldsymbol{w}_d$ axis are one for positive trials, and zero for negatives. While λ_d=0 indicates that $\boldsymbol{w}_d^\top \boldsymbol{\Sigma}_n \boldsymbol{w}_d$=0 for positive trials and $\boldsymbol{w}_d^\top \boldsymbol{\Sigma}_n \boldsymbol{w}_d$=1 for negative ones. Consequently, spatial patterns with associated eigenvalues λ=1/2 lack of discrimination.

In this chapter we use the same eigenvalues correction to account for the discriminative capacity of spatial pattern as presented in Section 3.1.2, according to Equation (3.4), so that λ'_d=0 comes from an axis without class separability and λ'_d=1 enhances the variance of either class.

$$\lambda'_d = 2\left|\lambda_d - \frac{1}{2}\right|$$

Finally, the weighted average of $\boldsymbol{A}_b$ along the columns provides a scalar value indicating the channel-wise capacity to discriminative between the two classes in the band b:

$$\rho_{cb} = \sum_{d=1}^{D} \lambda'_{db} |a_{cdb}|, \tag{5.6}$$

with c and d indexing rows and columns of $\boldsymbol{A}$, as well as the channels and patterns from CSP, respectively. As a result, the corrected eigenvalues favor those patterns maximizing the class separation in terms of the trial variance, and the vector $\boldsymbol{\rho} \in \mathbb{R}^C$, in the channel space, describes how well SSP solves the discrimination problem for $\mathcal{X}_b$.

5.1.4 Generative-supervised feature relevance

To give a sense of relevance to the features given by the discriminative decoding of SSP in the different brain rhythms and mental states, a methodology is presented that combines the weights of a linear classifier and the projection matrix of a principal component analysis to assign a relevance value to each of the characteristics used in the classification tests following the procedure below:

Given a set of samples $X=\{\boldsymbol{x}, y\}$, where y contains the class label of the x sample. A class prediction model can be obtained by using Bayes' rule over a probabilistic model for the class conditional distribution of the data $P(X|y=k)$ for each class k, as:

$$P(y=k|x) = \frac{P(x|y=k)P(y=k)}{P(x)} = \frac{P(x|y=k)P(y=k)}{\sum_l P(x|y=l) \cdot P(y=l)} \tag{5.7}$$

In particular, linear discriminant analysis (LDA), models $P(x|y)$ as a multivariate Gaussian distribution with the density function:

$$P(x|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)\right) \tag{5.8}$$

More specifically, LDA assumes that the Gaussians of each class have the same covariance matrix $\Sigma_k = \Sigma$, reducing the log-posterior to:

$$\log P(y=k|x) = -\frac{1}{2}(x-\mu_k)^t \Sigma^{-1}(x-\mu_k) + \log P(y=k) + Cst. \tag{5.9}$$

Where $(x-\mu_k)^t \Sigma^{-1}(x-\mu_k)$ stands for the Mahalanobis distance between each sample x and the mean of the class μ_k. Cst is a constant term and corresponds to $P(x)$. Then LDA assigns the sample class x considering the class prior probabilities and the Mahalanobis distance to the nearest class mean. The log-posterior can also be written as:

$$\log P(y=k|x) = \omega_k^t x + \omega_{k0} + Cst. \tag{5.10}$$

where $\omega_k = \Sigma^{-1}\mu_k$ corresponds to the weights vector of the LDA decision surface and $\omega_{k0} = -\frac{1}{2}\mu_k^t\Sigma^{-1}\mu_k + \log P(y=k)$ to the intercept terms [68].

additionally, for a set of observed d-dimensional data vectors $\{\boldsymbol{t}_n\}, n \in \{1,\ldots,N\}$, the M principal axes $\boldsymbol{h}_m, m \in \{1...M\}$, are those orthonormal axes in which the variance retained in projection is maximum. The vectors $\boldsymbol{h}_m$ are given by the M eigenvectors with the largest associated eigenvalues of the sample covariance matrix $S = \Sigma_n(t_n - \bar{t})(t_n - \bar{t})^\top$ such that $Swm = \lambda_m h_m$. The vector $\boldsymbol{x}_{pca} = \boldsymbol{H}^\top(t_n - \bar{t})$, where $\boldsymbol{H}^\top = (h_1, h_2, \ldots, h_M)$. In this way, the Principal Component Analysis (PCA) technique projects the data to a set of new uncorrelated variables (components) that are ordered by the amount of original variance they describe, thus reducing the dimensionality of the data [69].

Considering the presented formulation of the LDA and PCA models, the following features relevance is proposed, Where $\boldsymbol{\omega}'$ is the normalized weight vector of the fitted LDA model, h_{ij} is an element of the PCA projection matrix in component i and the feature j, and r_j is the relevance value of the feature j.

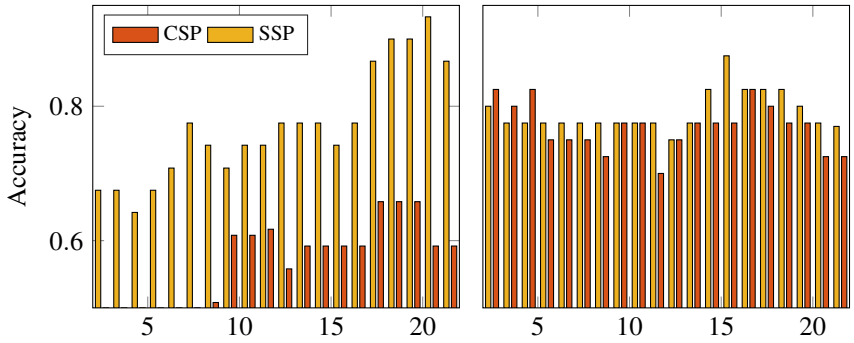
$$\omega_i' = |\omega_i| / \sum_{j=1}^{M} |\omega_j| \tag{5.11}$$

$$r_j = \sum_{i=1}^{M} \omega_i' |h_{ij}|, \tag{5.12}$$

5.2 Results

To validate the methodology proposed in Figure 5.1 as an ADHD biomarker, we performed classification tests of control subjects against ADHD using two feature sets: the discriminative decoding of SSP proposed in this chapter and the discriminative decoding of CSP proposed in chapter 3. Performing the same process of splitting the signals by rewards, the decomposition into brain rhythms, and the features concatenation for the two sets. Subsequently, we apply a Principal Components Analysis (PCA) to perform a dimension reduction, followed by a Linear Discriminant Analysis (LDA) as the classification machine for each set of features.

Figure 5.2 shows the average classification accuracy as a function of the number of PCA components (M) for each feature set. Regarding the decreasing condition (DC), PCA increases the accuracy compared to not performing a dimension reduction independently to the feature set. In particular, the discriminative decoding of SSP identifies the diagnostic groups significantly better compared to the CSP one, reaching a 93% average accuracy. In contrast, for the increasing condition (IC), the tunning of the PCA components presents a smaller variation in the average accuracy results for the two sets, however, it can be observed that the decoding of SSP presents a better result in the tuning of the number of components, reaching an average accuracy of 88%. As a result, under the adequate reward condition, the methodology proposed with the spatial



(a) DC (b) IC

Figure 5.2. Tuning of the number of components for each condition reward and considered feature sets.

patterns of SSP presents an improvement in the diagnostic support of ADHD compared to the discriminative decoding of CSP.

Table 5.2 presents the average performance achieved by each feature set at the optimal number of PCA components (M) and the optimal number of spatial filters (Q) in both reward conditions using a 10-fold cross-validation scheme. Besides, we report the fold-paired t-test p-value comparing validation scores of both feature extraction approach at accuracy, sensitivity, specificity, and F1-score metrics. Similar to the classification results of chapter 3, the decreasing condition achieves higher performance rates than the increasing one using the discriminative decoding of SSP. Particularly, for the decreasing condition, SSP achieves significant differences with CSP according to the hypothesis test (p-value <1%). However, p-values at the increasing condition are large enough to prove that the features set are statistically similar. Therefore, discriminative decoding of SSP improves ADHD diagnostic support for both reward conditions of the RSST paradigm compared to CSP decoding.

		M	Q	Accuracy	Sensitivity	Specificity	F1
DC	SSP	20	12	0.93±0.13	1.00±0	0.85±0.32	0.95±0.11
	CSP	18	14	0.66±0.23**	0.65±0.39**	0.65±0.39	0.59±0.33**
IC	SSP	15	12	0.88±0.17	1.00±0	0.75±0.34	0.91±0.12
	CSP	3	8	0.83±0.16	0.95±0.03	0.70±0.33	0.85±0.13

Table 5.2. Average performance for 10-fold cross-validation. Significant differences are marked for p-value $< 1\%$ (**).

To get a spectral relevance of the features obtained with the discriminative decoding of SSP, we apply the procedure described in Section 5.1.4. For this, we use the LDA model weights fitted with the optimal number of spatial filters, and the PCA projection matrix tunned with the number of components of Table 5.2. In total, we obtain 480 relevance values (one for each feature), where each value indicates how important that feature was to perform the classification task between control and ADHD subjects.

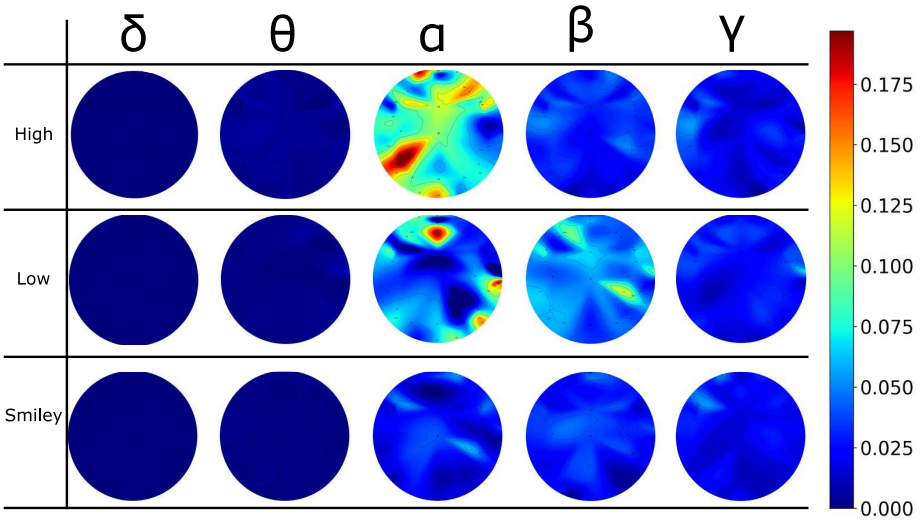


Figure 5.3. Topographic maps of the features relevance for DC by each reward and brain rhythm.

To perform a neurophysiological interpretation, we split the resulting relevancies into rewards and brain rhythms. Subsequently, using each group of relevancies, the topographic maps of each reward condition were graphed. Where Figure 5.3 contains the maps of the increasing condition and Figure 5.4 those of the decreasing one. In each graph, the rows represent a reward, and the columns the brain rhythms. In general, it can be observed that for both reward conditions, the δ, θ, and γ rhythms present low relevance in their features, indicating that these have a low contribution to the diagnostic support of the proposal. In contrast, it is observed that the α and β rhythms present a high relevance for both conditions, highlighting the "High" and "Low" rewards in α. Relevancies in α rhythm are consistent with those found in studies evaluating inhibitory control in subjects with ADHD in Go/noGo tasks, where the authors found that subjects with ADHD present abnormalities in the brain electrical activity patterns in the α rhythm [70]. Regarding the spatial location of the relevances, high values are evidenced in frontocentral regions where studies found that the disorder affects the most the error-related wave [47]. This indicates that the proposed feature extraction methodology adequately encodes the differences between mental states during inhibitory control tasks, highlighting spectral and spatial differences between subjects with the disorder and the control ones, allowing improvements in the diagnostic support of ADHD.

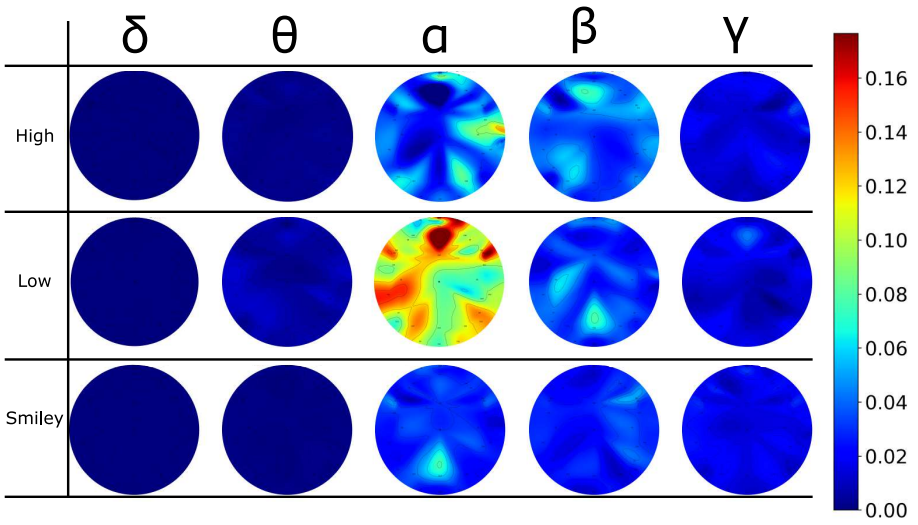


Figure 5.4. Topographic maps of the features relevance for IC by each reward and brain rhythm.

Chapter 6

Conclusions

This work proposes a feature extraction approach from EEG signals to support the ADHD diagnosis based on spatial patterns. Relying on the impulsivity symptom, we compute the patterns best discriminating successful and failed inhibitions in an RSST paradigm. Resulting patterns and eigenvalues feed a channel-wise discriminative capacity index, so that the larger the index, the easier to distinguish the inhibitions at a channel. Regarding the feature extraction based on CSP, Figure 3.5 illustrates that the proposal highlights the differences in brain activity associated with different stimuli while attenuating the activity of the common cognitive processes. Such a result is due to the eigenvalues related to the activity of interest are either biased towards one or zero, as shown in Figure 3.6. Further, the features resulting from real EEG recordings, in Figure 3.7, highlight brain regions related to the physiological response to errors [47].

Regarding the representation of EEG signals evaluation, we considered the four-class EEG signals from the Dataset IIa of the BCI Competition IV. According to Figure 4.3, illustrating the kappa score across the number of components, the dimension reduction carried out by MKSSP outperforms the kappa score of the input channel-space at all subjects, so avoiding the curse of dimensionality. Moreover, Table 4.1 evidences a significant improvement in class discrimination in comparison with state-of-the-art approaches. Specifically, MKSSP considerably increases the kappa score of usually low-performing subjects, such as S05, indicating that nonlinear relationships between channels hold discriminant MI information. While spatial filtering approaches, that map into euclidean spaces, hinder the nonlinearity, MKSSP takes decode it into a single kernel to enhance class separability.For the model interpretability, Figures 4.5 and 4.6 illustrate the kernel resulting from Equation (4.19) in a 2D KPCA-based projection for subjects S07 and S02 (the best and worst-performing, respectively). Note that MKSSP finds a space suitably separating classes for the subject S07. Despite the lowest performance of subject S02 and mixed Left and Right trials, the proposed methodology enhances the discrimination of three groups, namely, Foot, Tongue, and both hands. Such a result proves the capability of MKSSP to distinguish mental tasks despite the spatially close activity sources. Furthermore, Figure 4.7, com-

paring four feature sets for subject S05, indicates that CSP-based approaches lack separability as they miss geometric nonlinear channel relationships. Although including spectral decomposition, MKSSP without dimension reduction tangles the trials, due to the high dimension in the covariance matrices hinders discriminative patterns. On the contrary, MKSSP with the optimal number of components enhances class separation, agreeing with the test classification results, because the manifold information benefits the MI task identification. In the spectral relevance, Figure 4.8 depicts the band-wise MKL weights per subject. Regarding the spatial representation, Figure 4.9 presents the resulting Stein spatial patterns for subjects S07, S05, and S02. Only the patterns corresponding to the least (top) and most (bottom) weighted bands are displayed. On the one hand, the least weighted bands yield wider activity sources without neurophysiological interpretability. On the other hand, MKSSP mostly highlights bands with spatial patterns physiologically related to MI tasks. For instance, the contralateral activities, typical of left and right-hand movement, emerge from the first two patterns of subjects S07 and S05. In addition, the third pattern for all subjects finds activity close to the central vertex related to the foot movement. Consequently, MKSSP correctly decodes spatially interpretable patterns while spectrally highlighting the bands containing them.

Finally, we present a methodology for the diagnostic support of ADHD. The proposal involves a capacity discriminative index using Stein Spatial Patterns developed in Chapter 4, a decomposition in brain rhythms, and the discriminative decoding of Chapter 3. The performance evaluation of the proposed methodology was carried out using a PCA-LDA machine fed by two feature sets obtained using two spatial patterns techniques: CSP and SSP. The classification results show that the proposed methodology obtains better results with the SSP patterns since the CSP patterns are little benefited by the brain rhythms signal splitting, as shown in Table 5.2. Also, the results show that the proposal better identifies differences during the decreasing condition than in the increasing one, which agrees with previous studies about cognitive paradigms with motivation changes [1] and the results of Chapter 3. Besides, we perform a generative-supervise feature relevance to obtain a spatio-spectral interpretation of the features from the discriminative decoding. Where it is shown that the most important features for the machine are found in the alpha band for the High and Low rewards. Besides, high relevance of the features is observed in the frontal central regions. The foregoing coincides spatially with the regions where differences in inhibitory control are manifested in subjects with the disorder [7, 47]. Further, the results present a coincidence with the brain rhythm where in previous studies differences have been found between control subjects and ADHD in inhibitory control paradigms [70]. In conclusion, decoding the discriminative capacity of EEG channels by brain rhythms using SSP patterns allows extracting ADHD biomarkers under appropriate cognitive paradigm tests.

6.1 Future work

Regarding ADHD, for future work, we consider extending the methodology of the discriminative capacity index based on spatial patterns to be able to decode the temporal relevance of the signals using the RSST paradigm. Since some studies that consider the temporal dynamics of this type of signal have achieved favorable results performing the diagnostic support of ADHD [71].

Concerning BCI future work we considers two research directions. Firstly, we will work on the development of a BCI ability estimator based on the MKSSP methodology to evaluate the subject efficiency, since up to 30% of people underdevelop the coordination skills after training, hampering the spreading of this kind of system [72]. Secondly, we will consider a transfer learning scheme for training the MKSSP-based representation to achieve subject-independent BCI systems with reduced training times.

Bibliography

[1] Paula M Herrera, Alberto Vélez Van Meerbeke, Mario Speranza, Claudia López Cabra, Mauricio Bonilla, Michaël Canu, and Tristan A Bekinschtein. Expectation of reward differentially modulates executive inhibition. *BMC psychology*, 7(1):1–10, 2019.

[2] Russell A Barkley. *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment*. Guilford Publications, 2014.

[3] Stephen V Faraone, Joseph Biederman, and Eric Mick. The age-dependent decline of attention deficit hyperactivity disorder: a meta-analysis of follow-up studies. *Psychological medicine*, 36(2):159–165, 2006.

[4] National Health Service. Attention Deficit Hyperactivity Disorder: Symptoms. *Conditions*, page 1, 2018.

[5] Jacqueline F. Saad, Michael R. Kohn, Simon Clarke, Jim Lagopoulos, and Daniel F. Hermens. Is the Theta/Beta EEG Marker for ADHD Inherently Flawed? *Journal of Attention Disorders*, 22(9):815–826, 2018.

[6] S. Whitmont, R. Meares, E. Gordon, I. Lazzaro, and S. Clarke. The Modulation of Late Component Event Related Potentials by Pre-Stimulus EEG Theta Activity in ADHD. *International Journal of Neuroscience*, 107(3-4):247–264, 2008.

[7] Lynn Marquardt, Heike Eichele, Astri J Lundervold, Jan Haavik, and Tom Eichele. Event-related-potential (ERP) correlates of performance monitoring in adults with attention-deficit hyperactivity disorder (ADHD). *Frontiers in psychology*, 9:485, 2018.

[8] Yvonne Groen, Albertus A Wijers, Lambertus J M Mulder, Brenda Waggeveld, Ruud B Minderaa, and Monika Althaus. Error and feedback processing in children with ADHD and children with Autistic Spectrum Disorder: an EEG event-related potential study. *Clinical neurophysiology*, 119(11):2476–2493, 2008.

[9] Madeleine J Groom, Gaia Scerif, Peter F Liddle, Martin J Batty, Elizabeth B Liddle, Katherine L Roberts, John D Cahill, Mario Liotti, and Chris Hollis. Effects of motivation and medication on electrophysiological markers of response inhibition in children with attention-deficit/hyperactivity disorder. *Biological Psychiatry*, 67(7):624–631, 2010.

[10] Dean J Krusienski, Moritz Grosse-Wentrup, Ferran Galán, Damien Coyle, Kai J Miller, Elliott Forney, and Charles W Anderson. Critical issues in state-of-the-art brain–computer interface signal processing. *Journal of Neural Engineering*, 8(2):025002, 2011.

[11] Kai Keng Ang and Cuntai Guan. Eeg-based strategies to detect motor imagery for control and rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(4):392–401, 2016.

[12] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56, 2007.

[13] Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Frontiers in neuroscience*, 6:39, 2012.

[14] Steven Lemm, Benjamin Blankertz, Gabriel Curio, and K-R Muller. Spatio-spectral filters for improving the classification of single trial eeg. *IEEE transactions on biomedical engineering*, 52(9):1541–1548, 2005.

[15] Guido Dornhege, Benjamin Blankertz, Matthias Krauledat, Florian Losch, Gabriel Curio, and K-R Muller. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE transactions on biomedical engineering*, 53(11):2274–2281, 2006.

[16] Quadrianto Novi, Cuntai Guan, Tran Huy Dat, and Ping Xue. Sub-band common spatial pattern (sbcsp) for brain-computer interface. In *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pages 204–207. IEEE, 2007.

[17] Cristian Torres-Valencia, Álvaro Orozco, David Cárdenas-Peña, Andrés Álvarez-Meza, and Mauricio Álvarez. A discriminative multi-output gaussian processes scheme for brain electrical activity analysis. *Applied Sciences*, 10(19):6765, Sep 2020.

[18] Yaqi Chu, Xingang Zhao, Yijun Zou, Weiliang Xu, Guoli Song, Jianda Han, and Yiwen Zhao. Decoding multiclass motor imagery eeg from the same upper limb by combining riemannian geometry features and partial least squares regression. *Journal of Neural Engineering*, 17(4):046029, 2020.

[19] Camilo López-Montes, David Cárdenas-Peña, and Germán Castellanos-Dominguez. Supervised relevance analysis for multiple stein kernels for spatio-spectral component selection in bci discrimination tasks. In *Iberoamerican Congress on Pattern Recognition*, pages 620–628. Springer, 2019.

[20] Alireza Davoudi, Saeed Shiry Ghidary, and Khadijeh Sadatnejad. Dimensionality reduction based on distance preservation to local mean for symmetric positive definite matrices and its application in brain–computer interfaces. *Journal of neural engineering*, 14(3):036019, 2017.

[21] Alvina Goh and René Vidal. Clustering and dimensionality reduction on riemannian manifolds. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[22] Mehrtash T Harandi, Mathieu Salzmann, and Richard Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices. In *European conference on computer vision*, pages 17–32. Springer, 2014.

[23] P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

[24] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomputing*, 112:172–178, 2013.

[25] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on riemannian manifolds with gaussian rbf kernels. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2464–2477, 2015.

[26] J R Valdizán and A Izaguerri-Gracia. Trastorno por déficit de atención/hiperactividad en adultos. *REV NEUROL 2009; 48 (Supl 2): S95-9*.

[27] J W Cornejo. Prevalencia del trastorno por déficit de atención-hiperactividad en niños y adolescentes colombianos.

[28] A Velez van Meerbeke, C Talero-Gutierrez, R Gonzalez-Reyes, and M Ibañez. Attention deficit hyperactivity disorder prevalence of school students in bogotá, colombia. *Acta Neurológica de Colombia*, 24:6–12, 2008.

[29] Consuelo Vélez-Álvarez and José A Vidarte Claros. Trastorno por déficit de atención e hiperactividad (tdah), una problemática a abordar en la política pública de primera infancia en colombia. *Revista de salud pública*, 14:113–128, 2012.

[30] Joanna M Berg, Robert D Latzman, Nancy G Bliwise, and Scott O Lilienfeld. Parsing the heterogeneity of impulsivity: A meta-analytic review of the behavioral implications of the UPPS for psychopathology. *Psychological Assessment*, 27(4):1129, 2015.

[31] Steven M. Snyder, Thomas A. Rugino, Mady Hornig, and Mark A. Stein. Integration of an EEG biomarker with a clinician's ADHD evaluation. *Brain and Behavior*, 5(4):1–17, 2015.

[32] Steven M Snyder, Thomas A Rugino, Mady Hornig, and Mark A Stein. Integration of an EEG biomarker with a clinician's ADHD evaluation. *Brain and Behavior*, 5(4):e00330, 2015.

[33] American Psychiatric Association and Others. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[34] G Shahaf, T Fisher, J Aharon-Peretz, and H Pratt. Comprehensive analysis suggests simple processes underlying EEG/ERP–demonstration with the go/no-go paradigm in ADHD. *Journal of neuroscience methods*, 239:183–193, 2015.

[35] Stephen V. Faraone, Cristian Bonvicini, and Catia Scassellati. Biomarkers in the Diagnosis of ADHD – Promising Directions. *Current Psychiatry Reports*, 16(11), 2014.

[36] Terence W Picton. The P300 Wave of the Human Event-Related Potential. *Journal of Clinical Neurophysiology*.

[37] Chaitra Sridhar, Shreya Bhat, U. Rajendra Acharya, Hojjat Adeli, and G. Muralidhar Bairy. Diagnosis of attention deficit hyperactivity disorder using imaging and signal processing techniques. *Computers in Biology and Medicine*, 88(May):93–99, 2017.

[38] Frank H. Duffy, Aditi Shankardass, Gloria B. McAnulty, and Heidelise Als. A unique pattern of cortical connectivity characterizes patients with attention deficit disorders: A large electroencephalographic coherence study. *BMC Medicine*, 15(1):1–19, 2017.

[39] Robert J Chabot and Gordon Serfontein. Quantitative electroencephalographic profiles of children with attention deficit disorder. *Biological psychiatry*, 40(10):951–963, 1996.

[40] Robert J Barry, Adam R Clarke, and Stuart J Johnstone. A review of electrophysiology in attention-deficit/hyperactivity disorder: I. Qualitative and quantitative electroencephalography. *Clinical neurophysiology*, 114(2):171–183, 2003.

[41] Robert J Barry, Stuart J Johnstone, and Adam R Clarke. A review of electrophysiology in attention-deficit/hyperactivity disorder: II. Event-related potentials. *Clinical neurophysiology*, 114(2):184–198, 2003.

[42] J. R. Wiersema, J. J. Van Der Meere, and H. Roeyers. ERP correlates of impaired error monitoring in children with ADHD. *Journal of Neural Transmission*, 112(10):1417–1430, 2005.

[43] Kai Keng Ang, Cuntai Guan, Karen Sui Geok Chua, Beng Ti Ang, Christopher Kuah, Chuanchu Wang, Kok Soon Phua, Zheng Yang Chin, and Haihong Zhang. A clinical evaluation on the spatial patterns of non-invasive motor imagery-based brain-computer interface in stroke. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4174–4177. IEEE, 2008.

[44] Steven Galindo-Noreña, David Cárdenas-Peña, and Álvaro Orozco-Gutierrez. Csp-based discriminative capacity index from eeg supporting adhd diagnosis. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1343–1347. IEEE.

[45] Laurens R. Krol, Juliane Pawlitzki, Fabien Lotte, Klaus Gramann, and Thorsten O. Zander. Sereega: Simulating event-related eeg activity. *Journal of Neuroscience Methods*, 309:13 – 24, 2018.

[46] Lee Ann Remington and LA Remington. Clinical anatomy and physiology of the visual system. –st. louis, mo, 2012.

[47] Mario Liotti, Steven R Pliszka, Ricardo Perez, Delia Kothmann, and Marty G Woldorff. Abnormal brain activity related to performance monitoring and error detection in children with ADHD. *Cortex*, 41(3):377–388, 2005.

[48] K Richard Ridderinkhof, Markus Ullsperger, Eveline A Crone, and Sander Nieuwenhuis. The role of the medial frontal cortex in cognitive control. *science*, 306(5695):443–447, 2004.

[49] Magdalena Senderecka, Anna Grabowska, Jakub Szewczyk, Krzysztof Gerc, and Roman Chmylak. Response inhibition of children with ADHD in the stop-signal task: An event-related potential study. *International Journal of Psychophysiology*, 85(1):93–105, 2012.

[50] Steven Galindo-Noreña, David Cárdenas-Peña, and Álvaro Orozco-Gutierrez. Multiple kernel stein spatial patterns for the multiclass discrimination of motor imagery tasks. *Applied Sciences*, 10(23):8628, 2020.

[51] Rajendra Bhatia. Positive definite matrices, princeton ser. *Appl. Math., Princeton University Press, Princeton, NJ*, 2007.

[52] Frank Nielsen and Sylvain Boltz. The burbea-rao and bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.

[53] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

[54] Yanting Lu, Liantao Wang, Jianfeng Lu, Jingyu Yang, and Chunhua Shen. Multiple kernel clustering based on centered kernel alignment. *Pattern Recognition*, 47(11):3656–3664, 2014.

[55] Yu Zhang, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Optimizing spatial patterns with sparse filter bands for motor-imagery based brain–computer interface. *Journal of neuroscience methods*, 255:85–91, 2015.

[56] Zheng Yang Chin, Kai Keng Ang, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Multi-class filter bank common spatial pattern for four-class motor imagery bci. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 571–574. IEEE, 2009.

[57] Luis F Nicolas-Alonso, Rebeca Corralejo, Javier Gomez-Pilar, Daniel Álvarez, and Roberto Hornero. Adaptive semi-supervised classification to reduce intersession non-stationarity in multiclass motor imagery-based brain–computer interfaces. *Neurocomputing*, 159:186–196, 2015.

[58] Lianghua He, Die Hu, Meng Wan, Ying Wen, Karen M Von Deneen, and MengChu Zhou. Common bayesian network for classification of eeg-based multiclass motor imagery bci. *IEEE Transactions on Systems, man, and cybernetics: systems*, 46(6):843–854, 2015.

[59] Khadijeh Sadatnejad and Saeed Shiry Ghidary. Kernel learning over the manifold of symmetric positive definite matrices for dimensionality reduction in a bci application. *Neurocomputing*, 179:152–160, 2016.

[60] S Udhaya Kumar and H Hannah Inbarani. Pso-based feature selection and neighborhood rough set-based classification for bci multiclass motor imagery task. *Neural Computing and Applications*, 28(11):3239–3258, 2017.

[61] Thanh Nguyen, Imali Hettiarachchi, Abbas Khosravi, Syed Moshfeq Salaken, Asim Bhatti, and Saeid Nahavandi. Multiclass eeg data classification using fuzzy systems. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2017.

[62] Pramod Gaur, Ram Bilas Pachori, Hui Wang, and Girijesh Prasad. A multiclass eeg-based bci classification using multivariate empirical mode decomposition based filtering and riemannian geometry. *Expert Systems with Applications*, 95:201–211, 2018.

[63] Sahar Selim, Manal Mohsen Tantawi, Howida A Shedeed, and Amr Badr. A csp\am-ba-svm approach for motor imagery bci system. *IEEE Access*, 6:49192–49208, 2018.

[64] Sara Razi, Mohammad Reza Karami Mollaei, and Jamal Ghasemi. A novel method for classification of bci multi-class motor imagery task based on dempster–shafer theory. *Information Sciences*, 484:14–26, 2019.

[65] Qingsong Ai, Anqi Chen, Kun Chen, Quan Liu, Tichao Zhou, Sijin Xin, and Ze Ji. Feature extraction of four-class motor imagery eeg signals based on functional brain network. *Journal of neural engineering*, 16(2):026032, 2019.

[66] Ruilong Zhang, Qun Zong, Liqian Dou, and Xinyi Zhao. A novel hybrid deep learning scheme for four-class motor imagery classification. *Journal of neural engineering*, 16(6):066004, 2019.

[67] Elena I Rodríguez-Martínez, Brenda Y Angulo-Ruiz, Antonio Arjona-Valladares, Miguel Rufo, Jaime Gómez-González, and Carlos M Gómez. Frequency coupling of low and high frequencies in the eeg of adhd children and adolescents in closed and open eyes conditions. *Research in Developmental Disabilities*, 96:103520, 2020.

[68] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[69] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analysers. 1998.

[70] Marie-Pierre Deiber, Roland Hasler, Julien Colin, Alexandre Dayer, Jean-Michel Aubry, Stéphanie Baggio, Nader Perroud, and Tomas Ros. Linking alpha oscillations, attention and inhibitory control in adult adhd with eeg neurofeedback. *NeuroImage: Clinical*, 25:102145, 2020.

[71] MC Maya-Piedrahita, D Cárdenas-Peña, and AA Orozco-Gutierrez. Diagnosis of attention deficit and hyperactivity disorder (adhd) using hidden markov models. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1205–1209. IEEE, 2021.

[72] Tengjun Liu, Gan Huang, Ning Jiang, Lin Yao, and Zhiguo Zhang. Reduce brain computer interface inefficiency by combining sensory motor rhythm and movement-related cortical potential features. *Journal of Neural Engineering*, 2020.