

## Research Article

# Mining EEG with SVM for Understanding Cognitive Underpinnings of Math Problem Solving Strategies

Paul Bosch <sup>1</sup>, Mauricio Herrera <sup>1</sup>, Julio López,<sup>2</sup> and Sebastián Maldonado <sup>3</sup>

<sup>1</sup>Facultad de Ingeniería, Universidad del Desarrollo, Av. Plaza 700, Las Condes, Santiago, Chile

<sup>2</sup>Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Ejército 441, Santiago, Chile

<sup>3</sup>Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile

Correspondence should be addressed to Sebastián Maldonado; [smaldonado@uandes.cl](mailto:smaldonado@uandes.cl)

Received 4 April 2017; Accepted 24 September 2017; Published 11 January 2018

Academic Editor: Guido Rubboli

Copyright © 2018 Paul Bosch et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We have developed a new methodology for examining and extracting patterns from brain electric activity by using data mining and machine learning techniques. Data was collected from experiments focused on the study of cognitive processes that might evoke different specific strategies in the resolution of math problems. A binary classification problem was constructed using correlations and phase synchronization between different electroencephalographic channels as characteristics and, as labels or classes, the math performances of individuals participating in specially designed experiments. The proposed methodology is based on using well-established procedures of feature selection, which were used to determine a suitable brain functional network size related to math problem solving strategies and also to discover the most relevant links in this network without including noisy connections or excluding significant connections.

## 1. Introduction

Recently, there has been an outburst in the number of investigations related to the applications of data mining tools to neuroscience [1, 2]. Data mining in this domain is usually related, on one hand, to processing/analyzing three-dimensional images from different medical imaging modalities that capture structural (e.g., MRI, CT, and histology) and functional/physiological (e.g., PET, fMRI, and SPECT) information about the human brain [3, 4]. On the other hand, some tools and approaches have been specifically tailored to grasp the complexity of brain electric activity through the analysis of electroencephalographic (EEG) signals [5]. However, the vast majority of these studies commonly seek to discover patterns in electrophysiological signals and images correlated with the diagnosis, prognosis, and evolution of a particular pathology or brain disorder and with the image analysis of normal/disease resting state fMRI [6, 7]. Comparatively, only very few works in this area use machine learning techniques for studying normal

brain cognitive high level functions; probably because in these cases, the interpretation of the effects of single brain regions or connections between these regions on the separation of *pattern classes* is more complicated, given that discriminative brain pattern is a description of the cumulative contributions of many features that contribute to cognitive underpinning of brain high-level functions.

In this work, we use machine learning techniques to discover patterns of synchrony in functional brain networks, constructed from the EEG registers of a group of healthy individuals while they were solving specially designed math problems. The problems were devised specifically to detect and measure analytic processing. An intuitive resolution could lead to a quick and simple but incorrect response that should be overridden analytically. This study aims to correlate types of responses (correct or incorrect) with specific patterns of neural synchronization. The primary finding is that classification of these patterns using data mining tools on datasets from complex cognitive processes related to math performance is achievable. For each pair of EEG channels,

corresponding to time windows associated with correct and incorrect answers given by the participants, correlations and phase synchrony were calculated. With these measures as entries, we construct connectivity (synchronization) networks as proxies of functional brain networks. A novel feature selection methodology that identifies the most relevant connections in these networks is proposed by using a nonlinear SVM-based classifier. This methodology allows us to determine not only a suitable network size but also the most relevant connections in the network, reducing the complexity and, therefore, facilitating the interpretation of mined patterns.

*1.1. Synchronization/Correlation Networks of Normal Brain Cognitive High-Level Functions Used in Resolution of Math Problems.* Lately, investigations have shifted from the study of local activation of large groups of neurons to the analysis of integration patterns among these groups. It is thought that the physiological bases of information processing and mental representation are provided by functional networks [8]. In fact, there is a great deal of current interest in the recent development of different techniques to extract large-scale functional and anatomical brain connectivity networks based on methods for creating correlation networks [9–11].

Researchers have developed a widely used method for creating correlation networks by using neural synchronization. Neural synchronization is a fundamental process in cortical computation which is believed to play an important role in information processing in the brain at both cellular and macroscopic levels [12, 13]. Brain oscillations that are ubiquitous phenomena in all brain areas become synchronized and consequently allow an implementation of the whole range of brain functions [12]. In particular, in our work, we use neural synchronization to measure the integrated activity of the functional brain network responsible for different math performances. Specifically, we use linear correlation and phase synchronization as measures for neural synchronization.

The correlation coefficient estimates linear coupling among signals of EEG channels, and its values are distributed over the unit interval. But the assumption that only linear interdependencies are relevant is actually not correct. Strictly speaking, linear correlation analysis based on Pearson's correlation coefficient and its derivatives can potentially miss important features of any dynamic system, particularly when we study brain functional network integration dynamics. Thus, in addition to linear correlation, we use phase synchronization between distant brain oscillating foci [14–18].

Phase synchrony in EEG channels assesses the stability of differences between phases of EEG signals at equivalent frequencies taken simultaneously by different electrodes. More simply stated, it is a measure of how the relative phase is distributed over the unit circle. If the two signals are phase synchronized, the relative phase will occupy a small portion of the circle, and the mean phase coherence is high. Phase synchronization has previously been considered to be a very good indicator of the functional coupling of neural activity in distant brain areas [19, 20]. To our knowledge, to date,

phase synchrony in EEG channels has not been used to study brain networks involved in mathematical activities. So, using it creates a relevant contribution to understanding the collaborative and integrative nature of neural functioning in mathematics.

The large-scale functional integration of different brain zones is a relevant aspect of understanding the neural mechanisms responsible for the use of diverse problem-solving strategies in mathematics. The cognitive underpinnings of several mathematical activities have previously been related to a widely distributed brain network that includes parietal, temporal, and frontal structures as their main nodes [21–23]. Our research uses electroencephalographic (EEG) analysis [18] for the study of the whole-brain connectivity network and shows how mathematical cognition depends upon the integration of activities from distributed brain regions.

Some researches on EEG analysis have shown that specific aspects of mathematical reasoning could be related to different features of electric activity in some frequency bands (see, e.g., [24, 25]). In [25], for example, it is shown that incorrect performance in simple mathematical tasks is preceded by higher delta activity (signal frequencies  $< 4$  Hz) in the lateral and medial areas of the right prefrontal cortex and by higher theta activity (4–8 Hz) bilaterally in the medial frontal zones. These slow wave patterns precede the subject's erroneous performance and show inhibited activity of the error-monitoring areas during erroneous mathematical calculations (i.e., these areas were simply not recruited). Therefore, a failure in the functional integration of these zones during problem resolution would be responsible for the subject's erroneous mathematical performance. On the other hand, correct answers were preceded by alpha activity (8–12 Hz) in the right posterior parietal area, a zone previously linked to mathematics. These early findings suggest that the size and integration of the functional network of different brain zones entailed in the resolution of problems are a relevant issue for understanding the neural mechanisms underlying math performance.

*1.2. Graph Theoretical (Network) Approaches and SVM Working Together.* Network theory is helpful in characterizing the interdependencies of various brain zones. However, graph theoretical (network) approaches in the study of brain functional networks suffer from some important methodological difficulties [26, 27]. For example, graph measures are strongly dependent on the network size (number of nodes), network density (percentage of links present), and degree (number of connections per vertex). This makes comparing results from different studies, which generally use distinct criteria to build functional networks, very difficult. Indeed, to construct unweighted networks, one has to apply a threshold on the connectivity values of the original weighted network. This results in scaling of the network properties as a function of the threshold [26]. The threshold can be chosen in a variety of ways, for example, based on an arbitrary choice, or using statistical criteria of connectivity strength, based on the average degree, or based on the density of the network. Fixing a standard number of vertices and the

average degree could solve these size effects but could also introduce spurious connections or ignore strong connections in the network [27].

Recently, the use of a minimum spanning tree (a sub-network of the original weighted network that connects all vertices in the network without forming loops and has the minimum total weight of all possible spanning trees), see [28], has been proposed to solve many of these methodological difficulties.

*1.3. Top-Down Approach and Main Motivations for This Study.* Students often arrive at universities without a well-formed background in abstract reasoning and with limited experience in the application of mathematical strategies. They lack proper understanding of some mathematical topics and often use inappropriate associations of different facts while trying to solve mathematical problems. These associations are fast internal reactions to external stimuli and appear to be related to the way in which the mind processes information.

Many authors in educational research have pointed out the persistence of student errors and misconceptions with respect to specific topics and tasks. For example, in [29], the authors observed that students react in a similar way to a wide variety of conceptually nonrelated problems that share some external common features. This fact led them to suggest that many responses described in the literature as alternative conceptions (misconceptions) could be better explained as evolving from a few common intuitive rules such as *More of A—More of B, Same A—Same B, Everything can be divided, and Over-generalized linearity*.

The present work applies a dual-process model of cognitive processing to these kinds of problems, testing the hypothesis that relative amounts of intuitive/analytic processing by the brain promote different strategies in the resolution of mathematical problems, leading to accurate or faulty solutions.

This work aims to solve these methodological difficulties by using some advanced tools from data mining. Specifically, the main methodological contribution is twofold: First, we extend the  $\xi$ -SOCP method [30], originally developed for linear binary classification, to nonlinear modeling thanks to the use of kernel functions. This model proposes a robust setting based on second-order cone programming, in which the traditional maximum margin approach for SVM is adapted by replacing the reduced convex hulls by ellipsoids [31], leading to a potentially superior classification performance [30, 31]. Additionally, we propose a novel feature selection methodology that identifies the most relevant connections in the network of interest while constructing the classifier using the  $\xi$ -SOCP method [30].

The rest of this article is structured as follows: Section 2 presents the methodology for capturing the data used in the modelling process. Section 3 provides a brief description of developments for feature selection and SVM, in which our  $\xi$ -SOCP method and the novel-embedded feature selection strategy is highlighted. Section 4 describes our results using neural synchronization datasets collected for this study. A summary of this paper can be found in Section 5, where

we provide the main conclusions of this study and address future developments.

## 2. Materials and Methods: Cognitive Neuroscience

*2.1. The Dual Process Theory.* As our theoretical framework, we use the dual process theory (DPT) [32, 33]. According to DPT, our cognition and behavior operate in parallel with two quite different modes, called system 1 (S1) and system 2 (S2), roughly corresponding to our commonly held notions of intuitive and analytical thinking. The S1 and S2 modes are activated by different parts of the brain and have different evolutionary origins (S2 being more recent evolutionary and, in fact, largely reflecting cultural evolution). Like perception, S1 processes are characterized by being fast, automatic, effortless, unconscious, and inflexible (hard to change or overcome). Unlike perception, S1 processes can be language-mediated and relate to events not in the here-and-now (i.e., events in faraway locations and in the past or future). In contrast, S2 processes are slow, conscious, effortful, and relatively flexible. The two systems differ mainly along the dimension of accessibility: how fast and how easily things come to mind. Although both systems can at times run in parallel, S2 often overrides the input of S1 when analytic tendencies are activated and cognitive resources are available. For example, it is known that in geometry-related math problems, students tend to handle attributes of the problems such as distance, size, and similarity that are automatically registered by S1 quickly and spontaneously. We used this fact to design tests with some salient stimuli in such a way that each alternative for answering the problem would clearly indicate whether the participant took an intuitive/wrong strategy or an analytic/correct one.

The use of S2, consciously accessed, analytical processes trigger global and large-scale patterns of integrated neural activity. This fact appears as a variation on the global amount of synchrony between different brain areas. A greater proportion of S2 processes will appear as a greater amount of global synchrony. On the other hand, typical math errors due to semiautomatic use of heuristics will appear neurally as a reduced coupling of central work space neurons. Central work space neurons are thought to be particularly dense in the parietal, prefrontal, and cingulate cortices [21].

*2.2. Test Designing Based on Cognitive Neuroscience.* DPT enables understanding diverse phenomena because it predicts different judgments qualitatively depending on which reasoning system is used. DPT has been applied successfully to diverse domains and phenomena across a wide range of fields. While heuristic processing may render some manageable mathematics problems (by reducing the number of consciously driven operations), on some occasions, it can lead to errors and bias, reducing the effectiveness of a strategic plan of resolution. Available evidence and theory suggest that a converging suite of intuitive cognitive processes facilitates and supports some common rule-based flaw strategies in the resolution of math problems, which is a central aspect of deficient mathematical performance. In this

way, stereotyped errors come from the semiautomatic and insufficiently evaluated application of highly repeated S1 system heuristics for solving problems. Under most circumstances, S1 procedures lead to correct answers (e.g., linearity is a common property of many, but not to all, mathematical operations) but in certain cases, it can lead to mistakes. To avoid these errors, the subjects must inhibit their semiautomatic responses to allow proper, conscious evaluation of the problem [34]. Some neuroscience researches have linked response inhibition to prefrontal activity, especially in its medial zones [35]; error monitoring in general (see [36], for detailed review) and mathematical error monitoring in particular [22] have been linked to the frontal lobes, mainly to their medial structures.

However, individual differences in the tendency to override initially flawed intuitions in reasoning analytically could be associated with different mathematical performances. In fact, elaborative processing must entail a deeper level of consciously controlled stimulus analysis. This processing is assumed to involve more effortful, analytical thought and is less likely to lead to errors and biases, although sometimes it may prove to be dysfunctional due to effects such as *paralysis by analysis*—the tendency to become overwhelmed by too much information processing.

Some attributes of the problem denominated in DPT as *natural assessments* could lead to wrong strategies and answers, because students could ignore other, less accessible, attributes of the problem, or some instructions that should be considered in the resolution.

Another possible source of errors is called *attribute substitution*. According to [32], when people try to solve a complex problem, they often substitute attributes. That is to say, an individual assesses a specific real attribute of the problem heuristically by means of another attribute, which comes to mind more easily. The real attribute is less accessible, and another, related attribute which is more available replaces the first one. This substitution is so fast that S2 monitoring functions cannot be activated. The individual does not notice that he/she is really answering another question.

The math tasks in our experiments were designed to highlight different problem resolution strategies. An intuitive approach, for example, will produce a quick and easy, yet incorrect, answer that must be analytically overridden to be correct. In every case, participants choosing different resolution strategies will at the same time choose different alternatives to answer the math problem. Appendix B presents three of the math problems for illustrative purposes. The complete list of the 20 math problems can be found as supplementary material (available here).

**2.3. Preprocessing the Dataset from EEG Recording.** The raw data for the training and test subsets (see the next sections) were extracted from the EEGs of a group of engineering students that were recorded while each one of them was solving a set of 20 math problems. The relevant metadata for the various participants is presented in Appendix A. These EEGs (10–10 position convention) were registered in a semidark room with a low level of environmental noise while each

student was sitting in a comfortable chair. The data were recorded with the 64-channel Geodesic Sensor Net (EGI, USA) at the sampling frequency of 1000 Hz.

Since the sensors in the outer ring of the net were excluded from the analysis, because of low-quality signals, only 61 sensors were used for computations. The data were previously filtered (FIR, band-pass of 1–100 Hz), rereferenced against the common average reference, and segmented into nonoverlapping 1-s epochs using NS3 software.

As preliminary work for cleaning the dataset, we separated the oscillatory EEG-evoked electric activity from the induced one [37]. To do this, the EEG-evoked activity for each subject and his/her specific math problem was measured and averaged. This evoked activity was then subtracted from the total EEG activity through tests, subjects, and electrodes. The resulting EEG subtraction signal was analyzed with a fast Fourier transform on mobile overlapping and longtime windows between 5 and 10 seconds, because we did not know a priori what the interesting cognitive events to measure would be.

The measurement for each subject-math problem was segmented into time intervals ranging from  $-0.1$  s to 61 s. In  $t = 0$ , the math problem is presented, while at  $t = 60$  s, the question mark appears. The value  $-500$  ms is considered to be the baseline of before the occurrence of the problem.

**2.4. Constructing the Correlation and Synchronization Matrices from the Raw Dataset.** As will be shown in the following sections, a new method for feature extraction from EEG signals was developed by choosing elements of the correlation or synchronization matrices. The EEG time series recorded for each participant/math problem were used to construct the correlation and synchronization matrices of the functional brain networks with rows and columns representing sensors. These matrices contain information about (linear) interdependence and long-range synchronies between EEG channels. Both types of information would be used for classification purposes. Moreover, in the case of the synchronization matrix, we would also manage information about frequency bands.

The correlation coefficient  $r_{x,y}$  is perhaps one of the most well-known measures for (linear) interdependence between two signals  $x$  and  $y$ :

$$r_{x,y} = \frac{1}{N} \sum_{k=1}^N \frac{(x(k) - \bar{x})(y(k) - \bar{y})}{\sigma_x \sigma_y}, \quad (1)$$

where  $N$  is the length of the signals,  $\bar{x}$  and  $\bar{y}$  are the (sample) means of  $x$  and  $y$ , respectively, and  $\sigma_x^2$  and  $\sigma_y^2$  are the (sample) variances of  $x$  and  $y$ , respectively.

The correlation coefficient  $r_{x,y}$  quantifies the linear correlation between  $x$  and  $y$ . If  $x$  and  $y$  are not linearly correlated,  $r_{x,y}$  is close to zero; on the other hand, if both signals are identical, then  $r_{x,y} = 1$ .

Every correlation coefficient  $r_{i,j}$  is a bivariate measure that serves as a coupling coefficient that links the electrode nodes  $i$  and  $j$ . With these coefficients as entries, we construct a connectivity matrix (adjacency matrix) **Corr**, representing

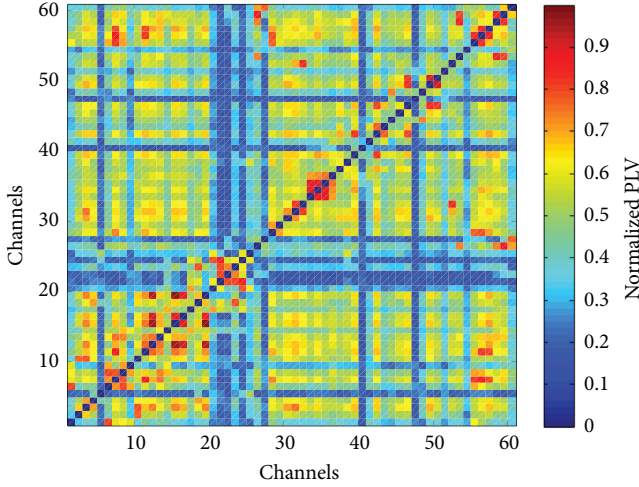


FIGURE 1: An example of the synchronization matrix for an individual and specific mathematics problem, in this case, corresponding to the beta (14–30 Hz) band. The colors indicate the extent to which two sensor locations are synchronized, which are quantified by the normalized PLV between every two EEG-recording sites  $i$  and  $j$  with  $i, j = 1, \dots, 61$ .

a functional brain network. Thus, we have a connectivity matrix  $61 \times 61$  composed of undirected and weighted edges consistent with the correlation coefficients. The matrix **Corr** is symmetric, so it has  $N(N-1)/2 = 1830$  independent elements. Zeros are placed in diagonal elements.

In order to discover to what extent two sensor locations were synchronized, we also used the phase locking value (PLV) [18, 19]. Sample PLV is one of the most widely used measures of brain synchronization. It quantifies the phase relationship between two signals with high temporal resolution without making any statistical assumptions on the data.

Given two time series of signals  $x(t)$  and  $y(t)$  and a frequency of interest  $f$ , the procedure computes a measure of phase locking between the components of  $x(t)$  and  $y(t)$  for each latency at frequency  $f$ . This requires the extraction of the instantaneous phase of every signal at the target frequency. The phases are calculated by convolving each signal with a complex wavelet function:

$$\Psi(t) = e^{-(t^2/2\sigma_t^2)} e^{2\pi i f t}, \quad (2)$$

that is,

$$W_x(t) = (\Psi \circ x)_t = \int \Psi(t') x(t' - t) dt' = A_x^W(t) e^{i\Phi_x^W(t)}, \quad (3)$$

where  $A_x^W(t)$  represents the signal amplitude. Following [19], we take  $\sigma_t = 7/f$  and we define  $W_y(t)$  in the same way as  $y(t)$ . Next, we can calculate the phase differences  $\Phi_{x,y}^W(t) = \Phi_x^W(t) - \Phi_y^W(t)$ . The phase locking value is then defined at time  $t$ , as the average value:

$$\text{PLV}_{x,y}(t) = \frac{1}{N} \left| \sum_{n=1}^N e^{i\Phi_{xy}^W(t,n)} \right|, \quad (4)$$

for all time-bins  $t$  and trial  $n \in \{1, \dots, N\}$ .

In our experiments, PLV measures were normalized relative to a baseline [38]. Specifically, this was done by using the 500 ms baseline before the onset of the math problem. The normalized signal was obtained by subtracting the average activity of the baseline from the raw signal and then dividing by the standard deviation of the baseline in a frequency-by-frequency manner.

By construction, PLV will be zero if the phases are not synchronized at all and will be one when the phase difference is in perfect, constant synchronization. The key feature of PLV is that it is only sensitive to phases, irrespective of the amplitude of each signal.

From the  $N = 61$  EEG channels, we computed a symmetric  $61 \times 61$  synchronization matrix **S** for each participant and for each math problem within a specific frequency band. Each element  $S_{i,j}$  of the matrix **S** corresponds to the PLV computed for the electrode pair  $i$  and  $j$ ,  $\text{PLV}_{i,j}(t)$ . The matrix **S** is also symmetric, so it has  $N(N-1)/2 = 1830$  independent elements and, as before, zeros are placed in diagonal elements.

Each matrix element of **S** is the PLV computed for the corresponding pair of sensors. An illustrative example of the synchronization matrix is presented in Figure 1.

**2.5. From the Correlation (Corr) and Synchronization (S) Matrices to a Binary Classification Problem.** For classification purposes, we prepared two datasets using the correlation **Corr** and synchronization **S** matrices separately. Each data point corresponds to an answer of a participant to a specific mathematics problem. The participant could answer the math problem in a correct way ( $y = +1$ ) or incorrectly ( $y = -1$ ).

With the help of these matrices, we further constructed feature vectors  $\mathbf{x}$  of size  $N(N-1)/2$ , whose components are in one case the elements of the matrix **Corr**, and in other cases elements of the matrix **S**. Here,  $N$  is the number of EEG sensors used in signal recordings, and we also used the fact that both matrices **Corr** and **S** are symmetric.

This way, the datasets for the learning machine are given by the set:

$$\left( \mathbf{x}_1^{(a)}, y_1 \right), \dots, \left( \mathbf{x}_l^{(a)}, y_l \right), \quad (5)$$

with characteristic vectors  $\mathbf{x}^{(a)} \in \mathfrak{R}^{N(N-1)/2}$ ,  $a = \{\mathbf{Corr}, \mathbf{S}\}$  corresponding to **Corr** or **S** matrix entries, respectively, and the output label  $y_l \in \{-1, +1\}$ , where the subindex  $l$  denotes the number of participants multiplied by the number of math problems given to each participant.

Thus, with the previously preprocessed dataset collected from 14 participants, with a 61-channel ( $N = 61$ ) EEG Geodesic Sensor Net (EGI) and with 20 math problems for each of them, we have  $l = 14 \times 20 = 280$  sample vectors of dimension  $N(N-1)/2 = 1830$  to train SVM models and make a preliminary classification for each **Corr** and **S** cases.

For the case of synchronization matrix  $\mathbf{S}$ , we further developed the datasets considering the analysis for frequencies in three distinct domains, which we have called as follows: *Low*, corresponding to  $\delta$  (<4 Hz) and  $\theta$  (4–8 Hz) bands; *Medium*, corresponding to the  $\alpha$  (8–12 Hz) and  $\beta$  (14–30 Hz) bands; and *High*, corresponding to  $\gamma$  (30–80 Hz) bands. In the case of the *Medium* frequency domain, we also studied the  $\beta$  band in a separate way. A classification study is performed using each frequency domain for synchrony detection in order to find the one that lead to the best classification.

### 3. Materials and Methods: Support Vector Machines and Feature Selection

Among the existing machine learning methods, SVM has demonstrated superior performance in several domains and, in particular, in neuroscience [39]. Its appealing characteristics, such as the absence of local minima and an adequate generalization of new samples, thanks to the *structural risk minimization* principle [40], made SVM one of the preferred classification approaches among researchers and practitioners [41].

Feature selection is a very important topic in high-dimensional applications, and, in particular, in neurology [1]. Finding the adequate subset of relevant variables for a given data mining task reduces the risk of overfitting, improving the model's predictive performance, and provides important insight into the process that generates the data, enhancing the interpretability of the model [42]. Support Vector Machine, however, cannot derive the feature's importance within the respective model, and therefore variable selection methods need to be used in order to reduce the level of noise in high-dimensional datasets [43].

In this section, the traditional SVM model developed for binary classification by [44] is presented in both linear- and kernel-based versions. Subsequently, two recently developed extensions are discussed, namely, the twin SVM method [45, 46] and the SVM method based on second-order cone programming presented in [30] ( $\xi$ -SOCP). Finally, two feature selection approaches used to address the issue of high dimensionality are described: the Fisher score and RFE-SVM method.

Among all SVM variations, we chose Twin SVM and SOCP SVM due to their superior performance we observed in previous studies (see e.g., [30, 47]). Twin SVM has shown positive empirical performance compared with the standard SVM formulation, being also computationally more efficient since the construction of these classifiers can be done by splitting the optimization problem into two smaller subproblems [45, 46]. Regarding  $\xi$ -SOCP, it is based on robust optimization, considering the worst-case setting for the class conditional densities related to the two training patterns in binary classification. In machine learning, robustness is a valuable property since it reduces the risk of overfitting, guaranteeing that the test performance does not deteriorate too much compared to the training performance when slight changes in the data distribution occur [30, 47].

**3.1. Soft-Margin SVM.** Let  $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$  be a set of examples  $\mathbf{x}_i \in \mathfrak{R}^n$  with labels  $y_i \in \{-1, +1\}$ ,  $i = 1, \dots, m$ . The traditional soft-margin SVM formulation [44] finds a classifier of the form  $\mathbf{w}^T \mathbf{x} + \mathbf{b} = 0$  by solving the following model:

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i (\mathbf{w}^T \mathbf{x}_i + \mathbf{b}) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \\ & i = 1, \dots, m. \end{aligned} \quad (6)$$

For each training example, a slack variable  $\xi_i$  is introduced, while  $C$  is a positive parameter that controls the trade-off between margin maximization and model fit.

A nonlinear decision surface can be obtained by using a kernel function [48]. A maximum margin classifier is constructed in a higher dimensional space by computing the dual of Formulation 4 and applying the kernel trick, leading to the following problem:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i,s=1}^m \alpha_i \alpha_s y_i y_s K(\mathbf{x}_i, \mathbf{x}_s) - \sum_{i=1}^m \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m, \end{aligned} \quad (7)$$

where  $\alpha$  is a vector in  $\mathfrak{R}^m$  of the dual variables corresponding to the constraints in 4 and  $K : \mathfrak{R}^n \times \mathfrak{R}^n \rightarrow \mathfrak{R}$  is a kernel function. A typical choice of kernel is the *Gaussian kernel*, which usually leads to better results [43]. This kernel is as follows:

$$K(\mathbf{x}_i, \mathbf{x}_s) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_s\|^2}{2\sigma^2}\right), \quad (8)$$

where  $\sigma > 0$  is the kernel width parameter [46].

**3.2. Twin Support Vector Machine.** The twin SVM method [45] constructs two nonparallel hyperplanes instead of the single classifier used in the soft-margin SVM formulation. Formally, two hyperplanes of the form  $\mathbf{w}_1^T \mathbf{x} + \mathbf{b}_1 = 0$ ,  $\mathbf{w}_2^T \mathbf{x} + \mathbf{b}_2 = 0$  are obtained in such a way that each of the functions is closer to the samples of one of the two labels and, at the same time, is as far as possible from those points of the other class. The following two problems are solved in order to find the following hyperplanes:

$$\begin{aligned} \min_{\mathbf{w}_1, \mathbf{b}_1, \xi_2} \quad & \frac{1}{2} \|\mathbf{A}\mathbf{w}_1 + \mathbf{e}_1 \mathbf{b}_1\|^2 + \frac{c_3}{2} (\|\mathbf{w}_1\|^2 + \mathbf{b}_1^2) + c_1 \mathbf{e}_2^T \xi_2 \\ \text{s.t.} \quad & -(\mathbf{B}\mathbf{w}_1 + \mathbf{e}_2 \mathbf{b}_1) \geq \mathbf{e}_2 - \xi_2, \quad \xi_2 \geq 0, \\ \min_{\mathbf{w}_2, \mathbf{b}_2, \xi_1} \quad & \frac{1}{2} \|\mathbf{B}\mathbf{w}_2 + \mathbf{e}_2 \mathbf{b}_2\|^2 + \frac{c_4}{2} (\|\mathbf{w}_2\|^2 + \mathbf{b}_2^2) + c_2 \mathbf{e}_1^T \xi_1 \\ \text{s.t.} \quad & (\mathbf{A}\mathbf{w}_2 + \mathbf{e}_1 \mathbf{b}_2) \geq \mathbf{e}_1 - \xi_1, \quad \xi_1 \geq 0, \end{aligned} \quad (9)$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are the data matrix for the positive and negative class, respectively,  $c_i$  are trade-off positive parameters ( $i = \{1, 2, 3, 4\}$ ), and  $\mathbf{e}_1$  and  $\mathbf{e}_2$  are vectors of one's appropriate dimension. Previous formulation is known as twin-bounded SVM (TB-SVM) [46], which is similar compared

to the twin SVM (TW-SVM) method proposed by Jayadeva et al. [45] when setting  $c_1 = c_2 = \varepsilon$ .

Similarly to the soft-margin formulation, a nonlinear decision surface can be obtained by applying the kernel trick. The kernel-based twin SVM method solves the following two problems:

$$\begin{aligned} \min_{\mathbf{u}_1, \mathbf{b}_1, \xi_2} \quad & \frac{1}{2} \|K(\mathbf{A}^T, \mathbb{X})\mathbf{u}_1 + \mathbf{e}_1 \mathbf{b}_1\|^2 + \frac{c_3}{2} (\|\mathbf{u}_1\|^2 + \mathbf{b}_1^2) + c_1 \mathbf{e}_2^T \xi_2 \\ \text{s.t.} \quad & -(K(\mathbf{B}^T, \mathbb{X})\mathbf{u}_1 + \mathbf{e}_2 \mathbf{b}_1) \geq \mathbf{e}_2 - \xi_2, \quad \xi_2 \geq 0, \end{aligned} \quad (10)$$

$$\begin{aligned} \min_{\mathbf{u}_2, \mathbf{b}_2, \xi_1} \quad & \frac{1}{2} \|K(\mathbf{B}^T, \mathbb{X})\mathbf{u}_2 + \mathbf{e}_2 \mathbf{b}_2\|^2 + \frac{c_4}{2} (\|\mathbf{u}_2\|^2 + \mathbf{b}_2^2) + c_2 \mathbf{e}_1^T \xi_1 \\ \text{s.t.} \quad & (K(\mathbf{A}^T, \mathbb{X})\mathbf{u}_2 + \mathbf{e}_1 \mathbf{b}_2) \geq \mathbf{e}_1 - \xi_1, \quad \xi_1 \geq 0, \end{aligned} \quad (11)$$

where  $\mathbb{X} = [\mathbf{A}^T \ \mathbf{B}^T]$  is the matrix of both training patterns (sorted by class) and  $K : \mathfrak{R}^n \times \mathfrak{R}^n \rightarrow \mathfrak{R}$  is the kernel function.

Finally, a new data point is assigned to label +1 ( $k = 1$ ) or -1 ( $k = 2$ ) according to its proximity to the two hyperplanes. That is,  $\mathbf{x} \in \mathfrak{R}^n$  belongs to the label  $k^*$  iff

$$k^* = \operatorname{argmin}_{k=1,2} \left\{ \frac{|\mathbf{w}_k^T \mathbf{x} + b_k|}{\|\mathbf{w}_k\|} \right\}. \quad (12)$$

The twin SVM method has been recently applied in neuroscience [49] and, in particular, in pattern analysis with EEG signal data [50, 51].

**3.3.  $\xi$ -Second-Order Cone Programming SVM.** In this study, we also used the robust SVM version based on second-order cones presented by [30]. For instance, if we suppose that  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are random vector variables that generate samples of positive class (brain synchrony pattern of a participant that made a good resolution of the math problem) and negative class (brain synchrony pattern of participant that made an incorrect resolution of the math problem), respectively, we should construct a maximum margin linear classifier such that the false-negative and false-positive error rates do not exceed  $\eta_1 \in (0, 1]$  and  $\eta_2 \in (0, 1]$ , respectively, in the following Quadratic Chance-Constrained Programming (QCCP) problem:

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C\xi \\ \text{s.t.} \quad & \Pr\{\mathbf{w}^T \mathbf{X}_1 + \mathbf{b} \geq 1 - \xi\} \geq \eta_1, \\ & \Pr\{\mathbf{w}^T \mathbf{X}_2 + \mathbf{b} \leq -1 + \xi\} \geq \eta_2, \\ & \xi \geq 0. \end{aligned} \quad (13)$$

False negative could appear, for example, due to not completely reliable math assessment. In this case, there is a possibility of correct answers to a math problem, despite that

it is registered as a pattern of low synchronization in the brain activity, that is, when a student correctly solves the problem without performing a deep analytical thinking.

The proposed robust setting suggests classifying each label correctly, up to the rate  $\eta_k$ , even for the worst data distribution. Thanks to Chebyshev's inequality [52, Lemma 1], this approach leads to the following deterministic problem:

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C\xi \\ \text{s.t.} \quad & \mathbf{w}^T \boldsymbol{\mu}_1 + \mathbf{b} \geq 1 - \xi + \kappa_1 \sqrt{\mathbf{w}^T \Sigma_1 \mathbf{w}}, \\ & -(\mathbf{w}^T \boldsymbol{\mu}_2 + \mathbf{b}) \geq 1 - \xi + \kappa_2 \sqrt{\mathbf{w}^T \Sigma_2 \mathbf{w}}, \quad \xi \geq 0, \end{aligned} \quad (14)$$

where  $\boldsymbol{\mu}_k$  and  $\Sigma_k$  are the means and covariance matrices for each class  $k = 1, 2$  and  $\kappa_k = \sqrt{\eta_k / (1 - \eta_k)}$ , and  $C > 0$  is a tradeoff parameter. The constraints appearing in the previous problem are called second-order cone constraints [53]. Thus, we refer to this problem as the  $\xi$ -SOCP formulation.

Similar to [54], a nonlinear version can be also derived for  $\xi$ -SOCP via the kernel trick. The kernel-based  $\xi$ -SOCP method is as follows:

$$\begin{aligned} \min_{\mathbf{s}, \mathbf{b}, \xi} \quad & \frac{1}{2} \mathbf{s}^T \mathbf{K} \mathbf{s} + C\xi \\ \text{s.t.} \quad & \mathbf{s}^T \mathbf{g}_1 + \mathbf{b} \geq 1 - \xi + \kappa_1 \sqrt{\mathbf{s}^T \Xi_1 \mathbf{s}}, \\ & -(\mathbf{s}^T \mathbf{g}_2 + \mathbf{b}) \geq 1 - \xi + \kappa_2 \sqrt{\mathbf{s}^T \Xi_2 \mathbf{s}}, \quad \xi \geq 0, \end{aligned} \quad (15)$$

where  $\mathbf{K} = [\mathbf{K}_{11}, \mathbf{K}_{12}; \mathbf{K}_{21}, \mathbf{K}_{22}] \in \mathfrak{R}^{m \times m}$ , with  $\mathbf{K}_{11} = \mathbf{A} \mathbf{A}^T$ ,  $\mathbf{K}_{12} = \mathbf{K}_{21}^T = \mathbf{B} \mathbf{A}^T$ ,  $\mathbf{K}_{22} = \mathbf{B} \mathbf{B}^T$  and

$$\begin{aligned} \mathbf{g}_k &= \frac{1}{m_k} \begin{bmatrix} \mathbf{K}_{1k} \mathbf{e}_k \\ \mathbf{K}_{2k} \mathbf{e}_k \end{bmatrix}, \\ \Xi_k &= \frac{1}{m_k} \begin{bmatrix} \mathbf{K}_{1k} \\ \mathbf{K}_{2k} \end{bmatrix} \left( \mathbf{I}_{m_k} - \left( \frac{1}{m_k} \right) \mathbf{e}_k \mathbf{e}_k^T \right) [\mathbf{K}_{1k}^T \ \mathbf{K}_{2k}^T]. \end{aligned} \quad (16)$$

**3.4. Feature Selection for SVM.** In this work, two feature selection strategies that have been used frequently for binary classification with SVM are considered: the Fisher score and RFE-SVM [42]. The first technique assesses variable relevance before applying SVM, constructing a ranking of attributes that can be used as input for the SVM model. This ranking is based on the distance between  $\mu_j^+$  and  $\mu_j^-$ , the means for the  $j$ th attribute in the positive and negative labels are as follows:

$$F(j) = \left| \frac{\mu_j^+ - \mu_j^-}{\left( \sigma_j^+ \right)^2 + \left( \sigma_j^- \right)^2} \right|, \quad (17)$$

where  $\sigma_j^+$  ( $\sigma_j^-$ ) is the standard deviation for the positive (negative) label. RFE-SVM, in contrast, performs a feature selection process embedded in the model, eliminating those variables that have the lowest contribution iteratively [42]. The variable contribution (the SVM margin when removing a given attribute) can be written in terms of the dual variables as follows:

$$W^2(\boldsymbol{\alpha}) = \sum_{i,s=1}^N \alpha_i \alpha_s y_i y_s K(\mathbf{x}_i, \mathbf{x}_s). \quad (18)$$

The RFE-SVM algorithm was successfully applied to linear Twin SVM in [55]. However, the RFE algorithm has not been extended before to  $\xi$ -SOCP, to the best of our knowledge. In this work, we implement the RFE algorithm for the  $\xi$ -SOCP method. This strategy together with the kernel-based  $\xi$ -SOCP formulation is a novel methodological contribution of this work. We should note that the characteristics of feature vectors  $\mathbf{x}$  are elements of the synchronization matrices, which represent the weights of network edges. So, this procedure determines the most significant features or, in other words, the most noticeable network edges that make the difference in separation of the strategies used in the resolution of mathematical problems, in a correct or incorrect way.

#### 4. Results and Discussion

We applied the classification and feature selection approaches described in the previous section on five different datasets corresponding to correlation data **Corr** and to synchronization data **S** which comprise three frequency domains (*Low*, *Medium*, and *High*) and the  $\beta$  frequency band. Each of the datasets has 280 samples (100 right answers and 180 wrong answers) described by 1830 variables.

For model evaluation, we chose a nested cross-validation (CV) strategy: training and test subsets were obtained using a 10-fold CV (outer loop), and the training subset was further split in training and validation subsets in order to find the right hyperparameter setting. The final feature ranking and classification were then performed with the full training subset from the outer loop for the best combination of hyperparameters, and the classification performance was computed by averaging the test results. This way, the test subsets from the outer loop remain unseen during the hyperparameter selection procedure. The following values for the hyperparameters were studied:  $C$ ,  $c_i$  and  $\sigma \in \{2^{-7}, \dots, 2^7\}$  and  $\eta_k \in \{0.2, 0.4, 0.6, 0.8\}$ .

All experiments were performed on an HP Envy dv6 with 16 GB RAM, 750 GB SSD, an Intel Core Processor i7-2620 M (2.70 GHz), and using Microsoft Windows 8.1 OS (64 bits). The toolbox LibSVM [56] was used for standard SVM approaches, while the SeDuMi Matlab Toolbox [57] and the codes provided by Shao et al., the author of Twin-Bounded SVM [46] (publicly available in <http://www.optimal-group.org/>), were used for  $\xi$ -SOCP and TB-SVM, respectively.

Table 1 presents the best performance of the model selection procedure for all classification methods (standard SVM, TB-SVM, and  $\xi$ -SOCP in their linear and nonlinear versions) and for all five datasets without performing feature selection. The average performance among all techniques for each data is also reported. The two best performances among all methods are highlighted in bold type.

From Table 1, we observe that *Medium* and **Corr** are the ones with better average performance among the five

TABLE 1: Performance summary for different classification approaches. All datasets.

	<i>High</i>	<i>Medium</i>	<i>Low</i>	<b>Corr</b>	Beta
SVM <sub>l</sub>	64.6	68.6	67.9	67.9	65.0
SVM <sub>nl</sub>	67.5	67.9	67.9	66.1	65.0
TB-SVM	66.1	66.8	67.5	66.4	64.6
TB-SVM <sub>nl</sub>	65.0	66.5	65.7	67.1	61.8
$\xi$ -SOCP	58.6	63.2	63.9	65.7	65.3
$\xi$ -SOCP <sub>nl</sub>	65.7	67.1	64.3	66.1	63.6
Average	64.6	<b>66.7</b>	66.2	<b>66.5</b>	64.2

TABLE 2: Performance summary for different feature selection approaches. *Medium* and **Corr** datasets.

	<i>Medium</i>	<i>n</i>	<b>Corr</b>	<i>n</i>
Fisher + SVM <sub>l</sub>	66.8	50	67.9	50
Fisher + SVM <sub>nl</sub>	67.1	1000	67.5	50
Fisher + TB-SVM	68.2	500	67.1	50
Fisher + TB-SVM <sub>nl</sub>	67.5	1000	67.9	500
Fisher+ $\xi$ -SOCP	67.1	250	68.2	250
Fisher+ $\xi$ -SOCP <sub>nl</sub>	66.4	500	<b>68.6</b>	100
RFE + SVM <sub>l</sub>	<b>68.2</b>	20	68.2	10
RFE + SVM <sub>nl</sub>	68.2	250	67.5	20
RFE + TB-SVM	67.5	500	67.1	20
RFE+ $\xi$ -SOCP	65.4	1000	68.2	10

different datasets. For the remaining analysis, we focus on these two datasets.

Next, we studied both feature selection approaches (Fisher score and recursive feature elimination) with the following number of selected attributes: 10, 20, 50, 100, 250, 500, 1000, and 1830 (i.e., with no features removed). The nested cross-validation strategy was performed for each subset of features, and the best performance is reported in Table 2, indicating the optimal number of selected variables for each case. The best performance among all methods is highlighted in bold type for each dataset.

In Table 2, we first observe that feature selection can improve predictive performance compared to a case with all available information, confirming what the specialized literature on this topic suggests [42]. In our case, an improvement of around 2% is achieved by eliminating those attributes that introduce noise in the modelling process. The best strategies are  $\xi$ -SOCP in its kernel-based version in combination with Fisher score and standard linear SVM in combination with RFE for the **Corr** and *Medium* datasets, respectively. Notice that the best approach is the one with fewer selected attributes in case of ties in accuracy.

Next, we construct the accuracy curves for the different subsets of selected variables, based on the four best methods presented in Table 2. For both frequency domains, we plot the classification performance of the four selected strategies



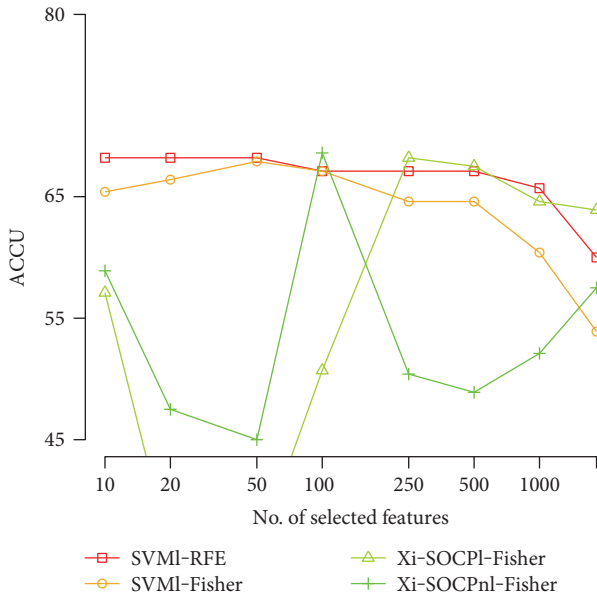


FIGURE 2: Accuracy versus the number of selected variables for **Corr** dataset.

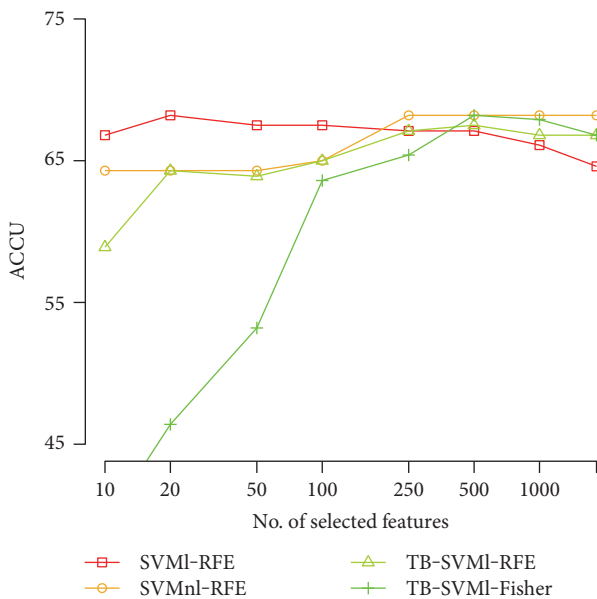


FIGURE 3: Accuracy versus the number of selected variables for **Medium** dataset.

to assess stability and overall predictive power. These results are presented in Figures 2 and 3.

In Figures 2 and 3, we first observe that no method outperformed others, and a reasonably broad range of classifiers is needed to define a good classifier adequately. The overall best performance is achieved with  $\xi$ -SOCPl as the classification method and using Fisher score for feature ranking, although the most stable strategy corresponds to standard SVM with linear kernel, and the RFE algorithm for variable elimination. Notice also that the best performance is usually

achieved between 10 and 100 attributes, discarding more than 90% of the available information.

Finally, we select the most relevant variables for each dataset. This is performed by combining the two best ranking strategies, namely, Fisher score and standard linear SVM in combination with RFE, and selecting the common variables (notice that each of these variables means a link between two EEG channels) that appear in both rankings (those with the highest importance).

Figures 4(a) and 4(b) show resulting networks constructed with statistically significant connections for the best classification methods and the set of connections that best discriminates the two classes of correct and incorrect answers during solving of math problems for correlation **Corr** and synchronization **S** matrices, respectively.

## 5. Conclusions

The discriminative brain pattern is a description of the cumulative contributions of many features. Therefore, the interpretation of the effects of single brain regions or connections between regions on the separation of the pattern classes is a complicated matter. However, some marked contributions aimed to solve this problem are present in this study.

It is well known that neural synchrony, which is involved in the large-scale transient integration of functional areas widely distributed over the brain, is required for normal cognitive operations [18]. Figure 4(a) shows precisely how the correlation can and should be used to measure this integration of different functional areas in normal processes related with math problem resolution.

Despite the apparent differences in patterns of connections (that were obtained by different methodologies), there are nodes that are common to both figures: AF4, F1, Fz, FT7, CF3, C1, C5, C6, P3, P7, and O1. These nodes precisely correspond to a widely distributed brain network previously identified as related to several mathematical activities [21–23].

The most discriminative connections (selected features) for the **S** case happen in the *Medium* range of frequencies from 8 to 30 Hz. Interestingly, these relevant connections are present in both the  $\alpha$  and  $\beta$  bands. The  $\alpha$  frequency band is widely associated with attention processes [58–62], and there is a large consensus that under conditions of inattention, some stimuli fail to be seen because the subject's attention is occupied with a different task and/or with another salient stimulus [63–65]. In such cases, perceptual, lexical, and/or semantic processing of the math problems could occur in the absence of conscious perception. As a result, stronger nonconscious effects are observed that may lead to erroneous mathematical performance.

In [66], it was suggested that focused attention elicits  $\alpha$  large phase locking during the processing of a target stimulus. This phase response can be interpreted as reflecting temporal attention. Thus, the  $\alpha$  phase should also play a crucial role in the attentional blink phenomenon, which represents reduced ability to report a second target after identifying the first target in visual stimuli. The explanations of the phenomenon proposed so far have focused primarily on some cognitive

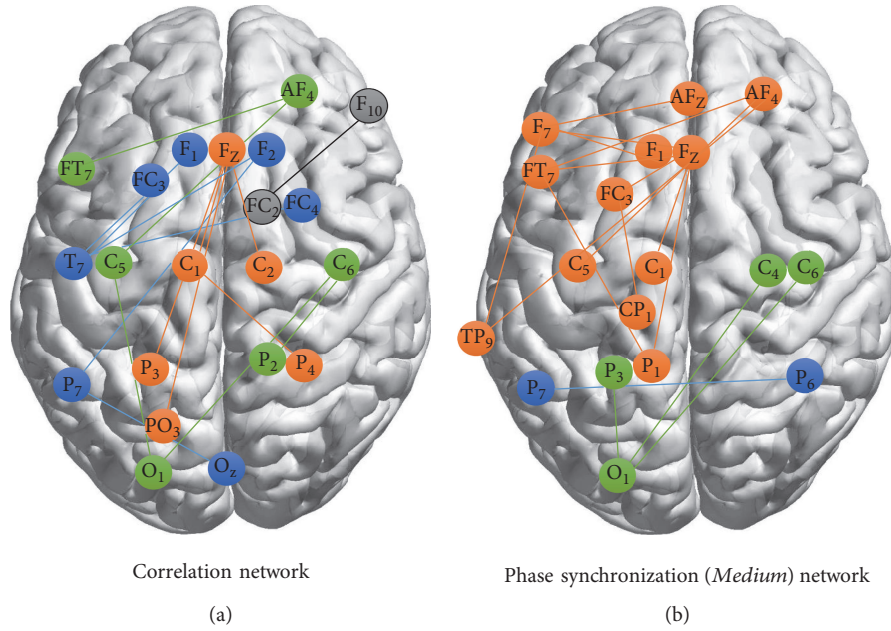


FIGURE 4: Graphic representation in which the vertices indicate EEG sensor locations (according to 10–10 position convention), and the edges represent the most significant features in the SVM classification and feature selection procedure. The colors represent different disconnected network components. BrainNet Viewer software [75] was used for the graph visualization.

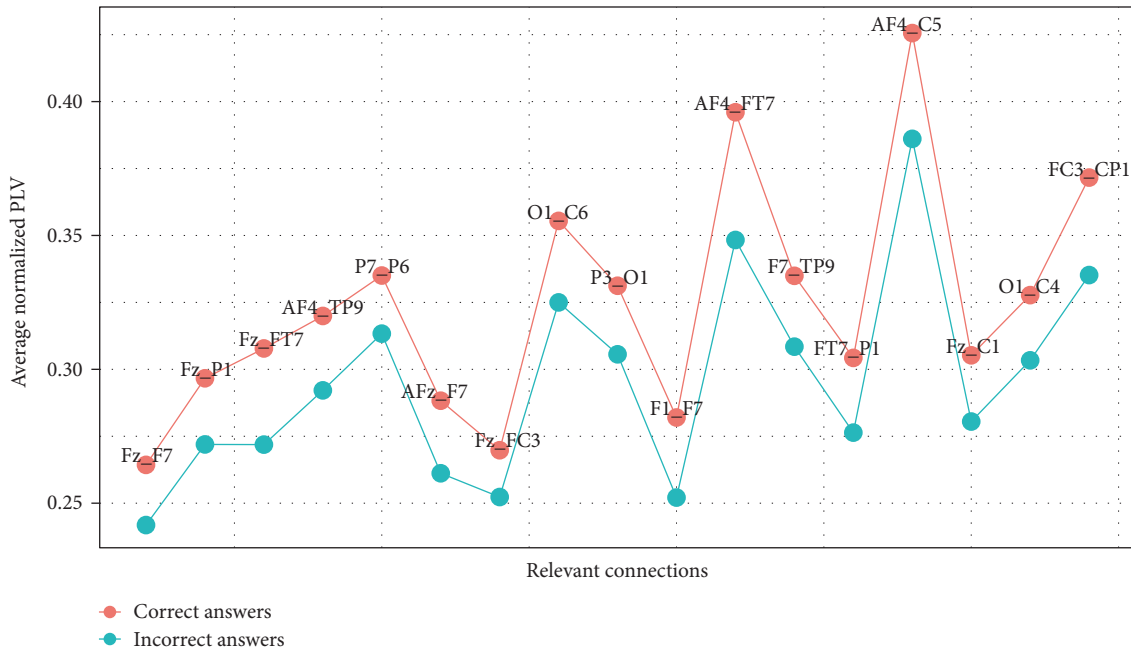


FIGURE 5: Average synchrony for relevant connections in the case of *Medium* dataset. Note that for every relevant connection, the average PLV is higher for correct answers than for incorrect ones.

aspects, such as attentional filters, capacity limitation, and retrieval failure processes [67, 68].

Figure 5 shows the average PLV calculated for relevant connections that were determined by the procedure of feature selection explained in Section 3.4. This average was taken for all measurements made for both classes of mathematical problems, that is, the ones that were answered properly and the ones that were answered incorrectly.

Figure 5 depicts two important results. First, the measures of the synchronization in the relevant connections indicate two types of different behaviors in regard to the cases of correct responses and incorrect responses. In the first case, synchronization is consistently higher than in the case of incorrect responses, indicating a different kind of resolution strategy whose neurological correlate shows a coordinated integration of different brain areas.

Secondly, synchronization measured by PLV values is not excessively high. The interpretation of this fact is interesting. It indicates that by building a proxy of a functional brain network, using synchronization networks able for distinguishing underlying normal cognitive processes in solving complex problems, the crucial point is not the detection of a high level of synchronization but rather the determination of which areas or groups are integrated with a stable synchronization. Traditional methods based on the choice of a threshold value in correlation/synchronization networks (e.g., using a threshold value greater than 0.5) would exclude much of the analysis described in this paper, which inevitably cause the loss of relevant information related to the neurological correlates that support the cognitive processes studied.

The primary finding of this study is that classification using automated SVM of datasets from complex cognitive processes related to math performance is feasible. Moreover, feature selection is a valuable procedure for reducing the complexity and, therefore, facilitating the interpretation of the mined patterns.

The observed differences in the math performance of participants and the success of SVM classification with a reasonable statistical significance suggest the potential use of this methodology as a novel approach to study patterns of connectivity in functional brain networks related to normal cognitive processes beneath the execution of complex tasks.

A still unsolved problem in graph theoretical approach for studying brain functional network integration is that network threshold can be chosen in a variety of ways based on an arbitrary choice or using statistical criteria of connectivity strength, based on the average degree or based on the network density. Particularly, when the threshold is based on a fixed number of connections in the network, this choice may result in either inclusion of spurious or noisy connections in networks (for too high density values or too high average degree) or the exclusion of relevant connections in networks (for too low density values or too low average degree) [27]. The procedure proposed here, based on feature selection tools, allows to determine not only a suitable network size by decreasing the size of feature vector but also the most relevant connections in network, that is, the ones that contribute most to the classification. As was shown in Table 2, feature selection improves predictive performance, which means that the optimal choices of the numbers of significant variables (connections) happen without including noisy connections or exclusion of significant connections.

There are some limitations to our study. First, the present analysis was performed on EEG data in sensor space, which contains some inherent spurious correlations because volume conduction causes the signal at each sensor to be a mixture of blurred activity from different inner cortical sources. More accurate inferences about anatomical locations need a source reconstruction of the activity in the cortex [69]. Second, in pattern classification, there are always uncertainties, for example, training datasets may contain incomplete information, there is input noise, there is noise in measurements, or underlying process is stochastic. As a result of such a probabilistic setting, uncertainties arise in

TABLE 3: Metadata for all participants.

Subject	Gender	Age	Correct answers	% C.A	Incorrect answers	% I.A	RT (ms)
1	Male	21	7	35	13	65	Null
2	Male	22	4	20	16	80	Null
3	Female	20	5	25	15	75	Null
—	—	—	—	—	—	—	—
4	Male	20	9	45	11	55	33091.80
5	Male	20	9	45	11	55	50645.45
6	Male	21	6	30	14	70	38395.85
7	Male	24	10	50	10	50	46157.05
8	Female	21	8	40	12	60	36193.85
9	Male	19	4	20	16	80	43862.60
10	Male	22	4	20	16	80	40500.10
11	Male	20	6	30	14	70	36475.95
12	Female	20	4	20	16	80	32184.35
13	Male	23	6	30	14	70	37211.10
14	Male	23	7	35	13	65	39002.50
15	Male	22	6	30	14	70	36721.80
16	Male	20	15	75	5	25	42437.95
17	Female	23	6	30	14	70	32250.65
Ave.	—	21.24	7	36	13	64	38938.00

learning from data. In order to get more reliable and reproducible results, we constructed a classifier whose misclassification rate does not exceed a defined maximum tolerable limit. For this, we presented in Section 3.3 a methodology for classification with uncertainties using SVM. However, general approaches of chance-constrained problems require the use of more suitable numerical methods for solving problems with linear constraints of probability to get computationally tractable approximations.

## Appendix

### A. Basic and Behavioral Data

The subjects in the experiment were first-year engineering students enrolled in calculus and algebra courses. All participants were right-handed, healthy individuals, without any known neurological disorder, and with an average age of  $21.24 \pm 1.43$ . The basic metadata is presented in Table 3.

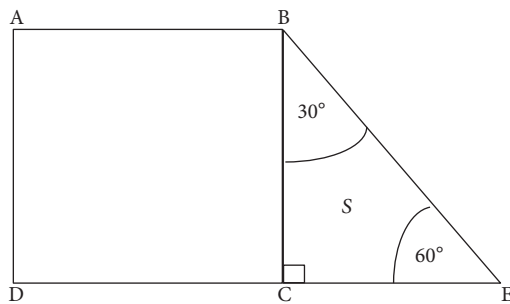
### B. Mathematical Problems

*Example 1* (attribute substitution). According to Kahneman et al. [32], attribute substitution often occurs when people try to solve a complex problem. This happens when an individual heuristically assesses a specific real attribute of the problem by means of another attribute, which comes easier to the mind. The real attribute is less accessible, and another related attribute, which is more available, replaces the first one. This substitution occurs so fast that the S2 monitoring functions cannot be activated, so it happens in a practically unconscious way. In geometry problem resolution, for

example, students quickly handle attributes such as distance, dimensions, and similarity that are automatically/unconsciously recorded. These attributes are called “natural assessments.” Managing these estimators can lead to errors, essentially because students ignore less accessible attributes that are critical to successfully solve the problem. We used this fact in Problem 1. In the resolution of this problem, some students ignored the values of angles and seem to have in mind the idea that the square has twice the area of the triangle. The average score obtained by students in this problem was 1.26 of a maximum of 4, showing that this was a fairly common error.

**Problem 1.** In the figure, the area of the triangle BCE is  $S$ . The area of the square ABCD is as follows:

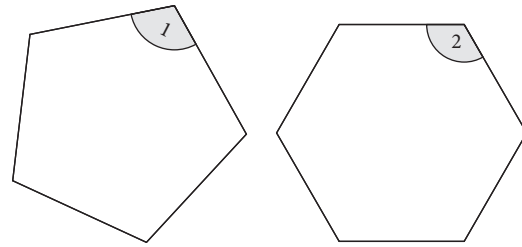
- (a)  $2S$ .
- (b)  $2\sqrt{3}S$ . ✓



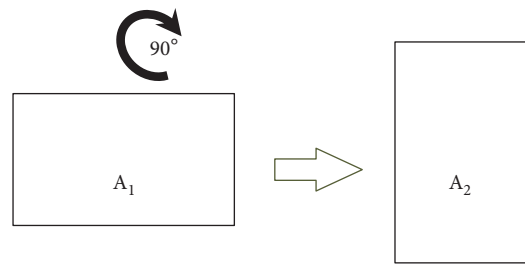
**Example 2** (intuitive rules). In [29, 70, 71], the authors observed that students react in a similar way to a wide variety of conceptually nonrelated problems but share some common features. This fact allowed them to suggest that many misconceptions described by educational research literature could be explained as consequences of some few common intuitive rules, such as “More of A—More of B,” “Same A—Same B,” “Everything can be divided,” and “Overgeneralized linearity.” The fundamental contribution of the intuitive rules is the observation that human response is often determined by irrelevant externalities of the tasks and not by the really important concepts and ideas. Intuitive rules showed that subjects during the resolution of a problem, under certain circumstances, can completely ignore some important data to save the immediacy. Using this fact, Problem 3 and Problem 4 were constructed. Problem 3 is intuitive and evident; the score obtained in this problem was the maximum possible: 4.0. However, equal areas do not necessarily imply equal volumes, and this mistake led to an average score 1.89 from a maximum of 4.0 for Problem 4. Problem 2 was also constructed by following intuitive rules, in particular, “Same A—Same B” [29].

**Problem 2**

- (a) Angle 2 is greater than angle 1. ✓
- (b) Angle 1 is equal to angle 2.



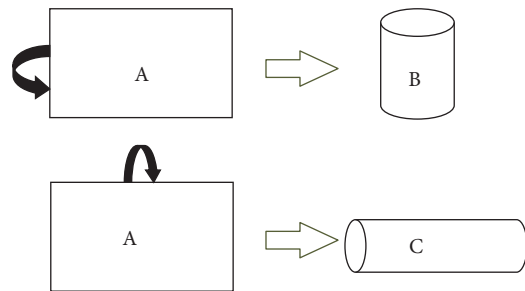
**Problem 3.** The rectangle  $A_1$  is rotated  $90^\circ$  to obtain the rectangle  $A_2$ .



Indicate the correct alternative:

- (a) The area of rectangle  $A_1$  is different from the area of triangle  $A_2$ .
- (b) The areas of both rectangles are equal. ✓

**Problem 4.** The rectangle A is wound in the ways indicated in the figures to construct the cylinders B and C.



Indicate the correct alternative:

- (a) The volume of cylinder C is different from the volume of cylinder B.
- (b) The volumes of both cylinders are equal.

**Example 3** (mathematical sets and conceptual metaphors). Lakoff and Nunez [72] stated that the concept of mathematical set is internalized using the image-schema CONTAINER [73]. The container is the source domain, whereas the mathematical set is the target domain of a conceptual metaphor. The various properties of the set are conceptualized through the metaphor “sets are containers.” Thus, for example, “an element belongs to the set” is conceptualized as “the element is inside the container,” and “A is a subset

of B” is conceptualized as “The container A is inside the container B.” In contrast, Fischbein and Baltsan [74] postulated that the mathematical set is conceptualized as an image-schema COLLECTION. Although Fischbein does not use the term metaphor, using *tacit model* instead, there are no major differences if we assume “sets are collections” as a new metaphor about the mathematical sets. We focus on common mistakes to detect the use of this type of conceptualizations. For this, we build Problem 6 to Problem 10 (see the supplementary material). In Problem 5, for example, the conceptualization of mathematical sets as collections was difficult in practice. It is counterintuitive to think of a collection of elements without any elements. For this reason, we should expect incorrect answers in cases of conceptualizations through collections. The average score on this problem was 0.42 out of a maximum of 4.0.

*Problem 5.* Given the proposition,

$$\begin{aligned}
 P = & \text{“with the solutions of the system of equations} \\
 & 4x + 3y = 2, \\
 & 4x + 3y = -2, \\
 & \text{”} \tag{B.1}
 \end{aligned}$$

it is possible to build a set,” indicate the correct alternative:

- (a)  $p$  is false.
- (b)  $p$  is true. ✓

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This research was partially supported by CONICYT, FONDECYT Projects 1160894 (Paul Bosch and Julio López) and 1160738 (Sebastián Maldonado), and the Universidad del Desarrollo Institutional Projects 20141013112757683278 (Paul Bosch) and 20111016102627393115 (Mauricio Herrera). This work was funded by the Complex Engineering Systems Institute (CONICYT-PIA FB0816).

## Supplementary Materials

All mathematical problems. (*Supplementary Materials*)

## References

- [1] P. Battista, C. Salvatore, and I. Castiglioni, “Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study,” *Behavioural Neurology*, vol. 2017, Article ID 1850909, 19 pages, 2017.
- [2] L. Squarcina, U. Castellani, M. Bellani et al., “Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques,” *NeuroImage*, vol. 145, pp. 238–245, 2017.
- [3] Y. Chen, J. Storrs, L. Tan, L. Mazlack, J.-H. Lee, and L. Lua, “Detecting brain structural changes as biomarker from magnetic resonance images using a local feature based SVM approach,” *Journal of Neuroscience Methods*, vol. 221, pp. 22–31, 2014.
- [4] B. S. C. Wade, S. H. Joshi, B. Gutman, and P. Thompson, “Machine learning on high dimensional shape data from subcortical brain surfaces: a comparison of feature selection and classification methods,” *Pattern Recognition*, vol. 63, pp. 731–739, 2017.
- [5] L. Vezard, P. Legrand, M. Chavent, F. Faita-Ainseba, and L. Trujillo, “EEG classification for the detection of mental states,” *Applied Soft Computing*, vol. 32, pp. 113–131, 2015.
- [6] S. Hojjati, A. Ebrahimzadeh, A. Khazaei, and A. Babajani-Feremi, “Predicting conversion from MCI to AD using resting-state fmri, graph theoretical approach and SVM,” *Journal of Neuroscience Methods*, vol. 282, pp. 69–80, 2017.
- [7] B. Rashid, M. Arbabshirani, E. Damaraju et al., “Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity,” *NeuroImage*, vol. 134, pp. 645–657, 2016.
- [8] S. L. Bressler, “Large-scale cortical networks and cognition,” *Brain Research Review*, vol. 20, no. 3, pp. 288–304, 1995.
- [9] M. Rubinov and O. Sporns, “Complex network measures of brain connectivity: uses and interpretations,” *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.
- [10] O. Sporns, “Structure and function of complex brain networks,” *Dialogues in Clinical Neuroscience*, vol. 15, pp. 247–262, 2013.
- [11] O. Sporns, “Contributions and challenges for network models in cognitive neuroscience,” *Nature Neuroscience*, vol. 17, no. 5, pp. 652–660, 2014.
- [12] G. Buzsáki, *Rhythms of the Brain*, Oxford University Press, New York, NY, USA, 1st edition, 2006.
- [13] P. Uhlhaas and W. Singer, “Neural synchrony in brain disorders: relevance for cognitive dysfunctions and pathophysiology,” *Neuron*, vol. 52, no. 1, pp. 155–168, 2006.
- [14] S. Doesburg, A. Roggeveen, K. Kitajo, and L. Ward, “Large-scale gamma-band phase synchronization and selective attention,” *Cerebral Cortex*, vol. 18, no. 2, pp. 386–396, 2008.
- [15] J.-P. Lachaux, E. Rodriguez, L. Michel, L. Antoine, J. Martinerie, and J. Francisco, “Studying single-trials of phase synchronous activity in the brain,” *International Journal of Bifurcation and Chaos*, vol. 10, no. 10, pp. 2429–2439, 2000.
- [16] F. Mormann, K. Lehnertz, P. David, and C. Eelger, “Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients,” *Physica D*, vol. 144, pp. 358–369, 2000.
- [17] A. Ossadtchi, R. Greenblatt, V. Towle, M. Kohrman, and K. Kamada, “Inferring spatiotemporal network patterns from intracranial EEG data,” *Clinical Neurophysiology*, vol. 121, pp. 823–835, 2010.
- [18] F. Varela, E. J.-P. Lachaux, E. Rodriguez, and J. Martinerie, “The brain web: phase synchronization and large-scale integration,” *Nature Reviews Neuroscience*, vol. 2, no. 4, pp. 229–239, 2001.
- [19] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. Varela, “Measuring phase synchrony in brain signals,” *Human Brain Mapping*, vol. 8, no. 4, pp. 194–208, 1999.
- [20] P. Uhlhaas, G. Pipa, B. Lima et al., “Neural synchrony in cortical networks: history, concept and current status,” *Frontiers in Integrative Neuroscience*, vol. 17, no. 3, 2009.

- [21] J. Danker and J. Anderson, "The roles of prefrontal and posterior parietal cortex in algebra problem solving: a case of using cognitive modeling to inform neuroimaging data," *NeuroImage*, vol. 35, pp. 1365–1377, 2007.
- [22] V. Menon, K. Mackenzie, S. Rivera, and A. Reiss, "Pre-frontal cortex involvement in processing incorrect arithmetic equations: evidence from event-related fMRI," *Human Brain Mapping*, vol. 16, no. 2, pp. 119–130, 2002.
- [23] L. Zago, L. Petit, M. Turbelin, F. Andersson, M. Vigneau, and N. Tzourio-Mazoyer, "How verbal and spatial manipulation networks contribute to calculation: an fMRI study," *Neuropsychologia*, vol. 49, no. 9, pp. 2403–2414, 2008.
- [24] T. Fernandez, T. Harmony, M. Rodriguez et al., "EEG activation patterns during the performance of tasks involving different components of mental calculation," *Electroencephalography and Clinical Neurophysiology*, vol. 94, no. 3, pp. 175–182, 1995.
- [25] T. Fernandez, T. Harmony, J. Silva-Pereyra et al., "Specific EEG frequencies at specific brain areas and performance," *Neuroreport*, vol. 11, no. 12, pp. 2663–2668, 2000.
- [26] A. Fornito, A. Zalesky, and E. T. Bullmore, "Network scaling effects in graph analytic studies of human resting-state fMRI data," *Frontiers in Systems Neuroscience*, vol. 4, no. 22, 2010.
- [27] B. C. van Wijk, C. J. Stam, and A. Daffertshofer, "Comparing brain networks of different size and connectivity density using graph theory," *PLoS One*, vol. 5, no. 10, article e13701, 2010.
- [28] P. Tewarie, E. van Dellen, A. Hillebrand, and C. J. Stam, "The minimum spanning tree: an unbiased method for brain network analysis," *NeuroImage*, vol. 104, pp. 177–188, 2014.
- [29] D. Tirosh and R. Stavy, "Intuitive rules: a way to explain and predict students' reasoning," *Educational Studies in Mathematics*, vol. 38, pp. 51–66, 1999.
- [30] S. Maldonado and J. Lopez, "Alternative second-order cone programming formulations for support vector classification," *Information Sciences*, vol. 268, pp. 328–341, 2014.
- [31] S. Nath and C. Bhattacharyya, "Maximum margin classifiers with specified false positive and false negative error rates," *Proceedings of the SIAM International Conference on Data Mining*, 2007.
- [32] D. Kahneman and A. Tversky, "Extensional vs. intuitive reasoning: the conjunction fallacy in probability judgement," *Psychological Review*, vol. 90, no. 4, pp. 293–315, 1983.
- [33] K. E. Stanovich and R. F. West, "Individual differences in reasoning: implications for the rationality debate?," *Behavioural and Brain Sciences*, vol. 23, no. 5, pp. 645–665, 2000.
- [34] R. West, M. Toplak, and K. Stanovich, "Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions," *Journal of Educational Psychology*, vol. 100, no. 4, pp. 930–941, 2008.
- [35] D. Sharp, V. Bonnelle, X. De Boissezon et al., "Distinct frontal systems for response inhibition, attentional capture, and error processing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 13, pp. 6106–6111, 2010.
- [36] S. Taylor, E. Stern, and W. Gehring, "Neural systems for error monitoring," *The Neuroscientist*, vol. 13, no. 2, pp. 160–172, 2007.
- [37] C. Tallon-Baudry and O. Bertrand, "Oscillatory gamma activity in humans and its role in object representation," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 151–162, 1999.
- [38] S. Makeig, S. Debener, J. Onton, and A. Delorme, "Mining event-related brain dynamics," *Trends in Cognitive Sciences*, vol. 8, no. 5, pp. 204–210, 2004.
- [39] A. Cerasa, I. Castiglioni, C. Salvatore et al., "Biomarkers of eating disorders using support vector machine analysis of structural neuroimaging data: preliminary results," *Behavioural Neurology*, vol. 2015, Article ID 924814, 10 pages, 2015.
- [40] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.
- [41] S. Wang, W. Jiang, and K.-L. Tsui, "Adjusted support vector machines based on a new loss function," *Annals of Operations Research*, vol. 174, no. 1, pp. 83–101, 2010.
- [42] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction, Foundations and Applications*, Springer, Berlin, 2006.
- [43] S. Maldonado, R. Weber, and J. Basak, "Kernel-penalized SVM for feature selection," *Information Sciences*, vol. 181, no. 1, pp. 115–128, 2011.
- [44] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
- [45] R. Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.
- [46] Y. H. Shao, C. H. Zhang, X. B. Wang, and N. Y. Deng, "Improvements on twin support vector machines," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 962–968, 2011.
- [47] S. Maldonado and J. Lopez, "Synchronized feature selection for support vector machines with twin hyperplanes," *Knowledge-Based Systems*, vol. 132C, pp. 119–128, 2017.
- [48] B. Scholkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.
- [49] S. Wang, M. Chen, Y. Li, Y. Shao, Y. Zhang, and S. Du, "Morphological analysis of dendrites and spines by hybridization of ridge detection with twin support vector machine," *PeerJ*, vol. 4, article e2207, 2016.
- [50] Q. She, Y. Ma, M. Meng, and Z. Luo, "Multiclass posterior probability twin svm for motor imagery EEG classification," *Computational Intelligence and Neuroscience*, vol. 2015, Article ID 251945, 9 pages, 2015.
- [51] S. Soman and Jayadeva, "High performance EEG signal classification using classifiability and the twin SVM," *Applied Soft Computing*, vol. 30, pp. 305–318, 2015.
- [52] G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. Jordan, "A robust minimax approach to classification," *Journal of Machine Learning Research*, vol. 3, pp. 555–582, 2003.
- [53] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Mathematical Programming*, vol. 95, no. 1, pp. 3–51, 2003.
- [54] P. Bosch, J. Lopez, H. Ramirez, and H. Robotham, "Support vector machine under uncertainty: an application for hydroacoustic classification of fish-schools in chile," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4029–4034, 2013.
- [55] Z.-M. Yang, J.-Y. He, and Y. Shao, "Feature selection based on linear twin support vector machines," *Procedia Computer Science*, vol. 17, pp. 1039–1046, 2013.
- [56] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

- [57] J. F. Sturm, "Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11, no. 12, pp. 625–653, 1999, Special issue on Interior Point Methods (CD supplement with software).
- [58] C. H. Brunia, "Waiting in readiness: gating in attention and motor preparation," *Psychophysiology*, vol. 30, no. 4, pp. 327–339, 1993.
- [59] J. J. Foxe, G. V. Simpson, and S. P. Ahlfors, "Parieto-occipital ~10 Hz activity reflects anticipatory state of visual attention mechanisms," *Neuroreport*, vol. 9, no. 17, pp. 3929–3933, 1998.
- [60] W. Klimesch, M. Doppelmayr, H. Russegger, T. Pachinger, and J. Schwaiger, "Induced alpha band power changes in the human EEG and attention," *Neuroscience Letters*, vol. 244, no. 2, pp. 73–76, 1998.
- [61] F. Lopes da Silva, "Neural mechanisms underlying brain waves: from neural membranes to networks," *Electroencephalography and Clinical Neurophysiology*, vol. 79, pp. 81–93, 1991.
- [62] M. Steriade, P. Gloor, R. R. Llinas, F. H. L. Dasilva, and M. M. Mesulam, "Basic mechanism of cerebral rhythm activities," *Electroencephalography and Clinical Neurophysiology*, vol. 76, no. 6, pp. 481–508, 1990.
- [63] S. Dehaene and J.-P. Changeux, "Neural mechanisms for access to consciousness," in *The Cognitive Neurosciences III*, M. Gazzaniga, Ed., pp. 1145–1157, MIT Press, 2003.
- [64] S. Dehaene and L. Naccache, "Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework," *Cognition*, vol. 79, no. 1-2, pp. 1–37, 2001.
- [65] S. Kouider and S. Dehaene, "Levels of processing during nonconscious perception: a critical review of visual masking," *Philosophical Transactions of the Royal Society of London B Biological Sciences*, vol. 362, no. 1481, pp. 857–875, 2007.
- [66] S. Hanslmayr, A. Aslan, T. Staudigl, W. Klimesch, C. S. Herrmann, and K. H. Bauml, "Prestimulus oscillations predict visual perception performance between and within subjects," *NeuroImage*, vol. 37, no. 4, pp. 1465–1473, 2007.
- [67] E. W. Martin, J. T. Enns, and K. L. Shapiro, "Turning the attentional blink on and off: opposing effects of spatial and temporal noise," *Psychonomic Bulletin and Review*, vol. 18, no. 2, pp. 295–301, 2011.
- [68] K. L. Shapiro, J. E. Raymond, and K. M. Arnell, "The attentional blink," *Trends in Cognitive Sciences*, vol. 1, no. 8, pp. 291–296, 1997.
- [69] M. Chavez, M. Valencia, V. Latora, and J. Martinerie, "Complex networks: new trends for the analysis of brain connectivity," *International Journal of Bifurcation and Chaos*, vol. 20, no. 6, pp. 1677–1686, 2010.
- [70] P. Tsamir, "When intuition beats logic: prospective teachers' awareness of their same sides-same angles solutions," *Educational Studies in Mathematics*, vol. 65, no. 3, pp. 255–279, 2007.
- [71] P. Tsamir and D. Tirosh, "Summing up and looking ahead: a personal perspective on infinite sets," in *Proceedings 30th Conference of the International Group for the Psychology of Mathematics Education*, vol. 1, pp. 49–63, 2006.
- [72] G. Lakoff and R. Nunez, *Where Mathematics Comes from: How the Embodied Mind Brings Mathematics into Being*, Basic Books, New York, NY, USA, 2000.
- [73] T. Oakley, *The Oxford Handbook of Cognitive Linguistics, Chapter Image schemas*, Oxford University Press, New York, NY, USA, 2007.
- [74] E. Fischbein and M. Baltsan, "The mathematical concept of set and collection' model," *Educational Studies in Mathematics*, vol. 37, pp. 1–22, 1999.
- [75] M. Xia, J. Wang, and Y. He, "Brainnet viewer: a network visualization tool for human brain connectomics," *PLoS One*, vol. 8, no. 7, article e68910, 2013.