



**HAL**  
open science

# Hybrid self organizing map and probabilistic quadratic loss multi-class support vector machine for mental tasks classification

Mounia Hendel, Abdelkader Benyettou, Fatiha Hendel

► **To cite this version:**

Mounia Hendel, Abdelkader Benyettou, Fatiha Hendel. Hybrid self organizing map and probabilistic quadratic loss multi-class support vector machine for mental tasks classification. *Informatics in Medicine Unlocked*, 2016, 4, pp.1 - 9. 10.1016/j.imu.2016.09.001 . hal-03033964

**HAL Id: hal-03033964**

**<https://hal.univ-lorraine.fr/hal-03033964>**

Submitted on 1 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License



# Hybrid self organizing map and probabilistic quadratic loss multi-class support vector machine for mental tasks classification



Mounia Hendel<sup>a,\*</sup>, Abdelkader Benyettou<sup>a</sup>, Fatiha Hendel<sup>b</sup>

<sup>a</sup> SIMPA Laboratory, Department of Computer Science, Faculty of Sciences, University of Science and Technology of Oran, Mohammed Boudiaf, USTO-MB, Oran, Algeria

<sup>b</sup> LARESI Laboratory, Department of Electronics, Faculty of Electric, University of Science and Technology of Oran, Mohammed Boudiaf, USTO-MB, Oran, Algeria

## ARTICLE INFO

### Article history:

Received 23 May 2016

Received in revised form

4 September 2016

Accepted 5 September 2016

Available online 6 September 2016

### Keywords:

SOM

M-SVM<sup>2</sup>

BCI

EEG

DWT

Mental Tasks

## ABSTRACT

Brain computer interface provides communication opportunity between the brain and the environment around a person with severe motor disabilities. However, the implementation of such interfaces requires a good signal processing scheme, whose performances depend principally on the technique used to select the best features, and the classification technique used to perform the discrimination between the different categories. This work proposes a new hybrid structure based on two stages with supervised and unsupervised learning. The first stage consists of a Self Organizing Map which allows to cluster the redundant irrelevant features and select the best descriptors. The second stage uses, a Probabilistic Quadratic Loss Multi-Class Support Vector Machine for the final classification. Five mental tasks namely: Baseline, Multiplication, Letter, Rotation, and Counting are considered, and an average accuracy between 81.73% and 91.90% is achieved. This result shows the effectiveness of the proposed method to enhance the performance of electroencephalogram classification problem.

© 2016 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Recently, there have been many interesting researches in the Brain Computer Interface (BCI) area. In fact, a BCI constitutes a direct communication device between the brain and the computer through brain signals. Its aim is to interpret the thoughts (mental tasks) of persons affected by a number of motor disabilities (without having to speak or to perform muscular activities) in a numerical form, that acts as a command to a computer [30,32].

We can find several invasive and non-invasive techniques to record the brain activities, with which we can determine the mental tasks: ElectroEncephaloGram (EEG), Functional Magnetic Resonance Imaging (fMRI), Magneto EncephaloGraphy (MEG), Near-Infrared Spectroscopy (NIRS), Positron Emission Tomography (PET), Event-Related Optical Signal (EROS) [30]. Among these techniques, EEG is the most used due to its ability to record the brain signals in a non-invasive manner and its low cost.

The determination of patterns from the EEG signals, include two approaches. The first approach is based on the analysis of Event-Related Potentials (ERPs). This last one represents the modification of the electrical potential generated by the brain in response to internal or external stimulus [3,22,33]. This signals are

generally of very low amplitude, so it is obligatory to perform multiple recordings and then averaging all of these measures, which allows to increase the Signal to Noise Ratio (SNR) and obtain reliable characteristics of the invoked potential. The invoked process can take a long time, and makes this method relatively slow to be useful in real-time control applications. The second approach, which is much more practical [3] and which is considered in this study, consists in analysing the five main frequency bands that characterize the EEG signals [30,32]: delta ( $\delta$ : 0.5–3.9 Hz), theta ( $\theta$ : 3.9–7.8 Hz), alpha ( $\alpha$ : 7.8–15.3 Hz), beta ( $\beta$ : 15.3–31.5 Hz), and gamma ( $\gamma$ : 31.5–40 Hz) bands.

However, the implementation of a typical BCI system requires a good signal processing scheme, whose performances are mainly associated to the technique used to select the best features from the EEG signals, and the classification technique used to interpret the obtained components. So, in this study, we focus especially on this two parts.

The selected features must be highly significant to achieve well discrimination between the different mental tasks. Many interesting approaches of EEG feature extraction have been proposed in the literature. They include Discrete Wavelet Transform (DWT) [29,9,1], Empirical Mode Decomposition (EMD) [8], Fast Fourier Transform [19], Short-Time Fourier Transform [1], Autoregressive features (AR) [14,17], and Discrete Fourier Transform [27]. Among the mentioned approaches, the DWT has been used successfully in feature extraction of EEG signals, because it provides a rigorous

\* Corresponding author.

E-mail address: [mounia\\_90@hotmail.com](mailto:mounia_90@hotmail.com) (M. Hendel).

mathematical approach that gives it the capacity of representing the non-stationary signals in time and frequency domain.

The EEG signals are obtained from several channels, and the feature vectors are constructed by concatenation of the descriptions that were obtained from the multiple used channels. This results in a large feature vectors while the number of samples of the training set is limited (which leads to the curse-of-dimensionality and the overtraining) [28]. Therefore, it is necessary, to perform feature selection and/or select a good classifier that is insensitive to these two problems.

The variable selection process is characterized by: the Feature Selection Algorithm (FSA), and the Evaluation Criterion of the Selected Features (ECSF). The FSA can be classified into two categories. Myopic algorithms that estimate the relevance of a descriptor without considering the other parameters, and Contextual algorithms that take into account the interactions (correlations) between all descriptors. The ECSF allows to evaluate the aptitude of the selected subset to discriminate between classes. Several methods are associated with this criterion and the most used are Filter and Wrapper methods. The first approach works independently from the classification algorithm (the selection is done before learning). However, the second approach uses a classification method to test the suitability of the selected variables. In this context, some approaches have been proposed, they include: Genetic Algorithm (AG) [18], Principal Component Analysis (PCA) [3], Common Spatial Patterns (CSP) [34], Forward Sequential Feature Selection (FSFS) [14], and Wilks Lambda measure (WLM) [8], Distinctive Sensitive Learning Vector Quantization (DSLVO) [24], etc.

As for classification, many models have been implemented in the framework and the most common are: Artificial Neural Networks (ANN) [17,29,8,20,6], Support Vector Machine (SVM) [11,17,1], Bayes Quadratic (BQ) classifier [14], Linear Discriminant Analysis (LDA) [11], k-Nearest Neighbors (KNN) [11,19] and Extreme Learning Machine (ELM) [17]. Over the proposed classifiers, the ones that have given the best results and the most used are the SVM; they provide good generalization, great performance to the curse of dimensionality, and insensitivity to overtraining. However, they were originally designed for binary problems and their extension to multiclass problems is still an active research issue. We can find two approaches, the indirect approach that involves only bi-class SVMs and the direct approach that considers all classes simultaneously. The first one treats the problem partially and covers three main methods: 1-against-1, 1-against-all and Error Correcting Output Code. At the opposite, the second one treats the SVM learning problem as a global optimization one. We find four large direct models on which the scientific community is currently focusing: Weston and Watkins (WW) [31] model, Grammer and Singer (GS) [7] model, Lee et al. (LLW) [16] model, and the Quadratic Loss Multi-Class Support Vector Machine (M-SVM<sup>2</sup>) model [10] introduced by Guermeur and Monfrini.

The current study proposes a new hybrid structure based on two stages, with supervised and unsupervised learning. This approach is validated on EEG database of Keirn and Aunon [13], that consists on EEG recordings of 7 subjects performing five different mental tasks namely: Baseline, Multiplication, Letter, Rotation, and Counting. The first stage consists of a Self Organizing Map (SOM) that reduces the original variable set (150 components obtained from DWT decomposition) into a smaller feature set, enabling to cluster the redundant irrelevant features and select the best descriptors to enhance the performance of the final classifier. The second stage uses a Quadratic Loss Multi-Class Support Vector Machine (M-SVM<sup>2</sup>). It receives as input the obtained feature set from the first stage and outputs class posterior probability estimates.

The remainder of the paper is organized as follows. Section 2 summarizes some previous works using Keirn and Aunon

database. Section 3 gives a description of the used database, explains the methodology adopted to build the first feature set, and presents details of our proposed hybrid structure. Section 4 exposes the obtained results. Section 5 discusses the results, and Section 6 concludes the study.

## 2. Related works

One can find many models proposed in the literature for the interpretation of mental tasks using ElectroEncephaloGram (EEG) signals. Hereafter, we have selected those whose database is similar to ours (see Section 3.1).

Keirn and Aunon [14] proposed a method based on AR features and BQ classifier. They demonstrate the necessity to use ( $\delta$ ,  $\theta$ ,  $\alpha$ , and  $\beta$ ) bands and AR features to enhance the classification results. First, the EEG signals are subdivided into 1/4 s segments, and 36 AR features are estimated for each segment. Then, BQ classifier is proposed to classify combinations case of task pairs for five subjects, and achieved an average accuracy of 84.6% across all task pairs and over all the five subjects.

Palaniappan [21] divided the EEG signals into 20 segments with a length of 1/2 s. Then, computed for each segment, the Power Spectrum (PS) and the ASymmetry Ratios (ASR) for,  $\delta$ ,  $\theta$ ,  $\beta$ , and gamma ( $\gamma$ ) bands as features. The obtained components were presented as inputs of ENN classifier. He reported, from this study, that the performance was improved by the addition of the gamma band. Also, an average accuracy of: Subject 1: 79.8%, Subject 2: 69.1%, Subject 3: 70.2%, Subject 4: 80.3% were achieved over pairs tasks.

In [17], Liang et al., subdivided the EEG signals into 1/2 s segments with an overlap of 1/4 s between adjacent segments, and count for each segments the same AR features used by Keirn and Aunon [14]. Then, an ELM was proposed as a method of classification of five mental tasks. In addition, the authors smoothed the output vectors of classifier over 20 consecutive segments. The results indicate significant improvement in the testing accuracy for classifier, and an average accuracy between 57.37% and 86.70% were achieved.

Zhiwie and Minfen [35] proposed a method using Wavelet Packet Entropy (WPE) and SVM. They divided the EEG signals into 1 s segments that overlapped by 0.8 s, and performed the entropy algorithm on each of  $\delta$ ,  $\theta$ ,  $\alpha$ , and  $\beta$  bands. Then, the SVM classifier was trained to discriminate between 2-class, 3-class, 4-class, and 5-class. The proposed system achieved an average accuracy for five-class of: (76.3%) for subject 1 and (68.5%) for subject 2. It is shown from the study, that WPE can be a great method for feature extraction and the SVM can be a great method to discriminate between different tasks.

Diez et al. [8] applied their proposed features extraction method, on EEG segments of 1 s duration. This method is based on the EMD and the estimation of six features: Root Mean Square (RMS), Variance, Shannon Entropy, Lempel-Ziv Complexity Value, and Central and Maximum Frequencies. Then, reduction of dimensionality was performed, based on the WL parameter. Two classifiers were employed to performed the classification LDA and ANN. With this method, the best average classification of 91.17% was obtained cross all task pairs and using LDA classifier. These results conclude the effectiveness of EMD to process nonstationary and nonlinear signals as the EEG.

Tolic and Jovic [29] presented a method based on DWT and Backpropagation Neural Network (BPNN) to classify two, four and five mental tasks relative to subject 1 and subject 2. They decomposed each EEG signal in to 1/2 s segments, and they applied DWT to extract different frequency bands. The features are then calculated as wavelet energies of the obtained bands. The

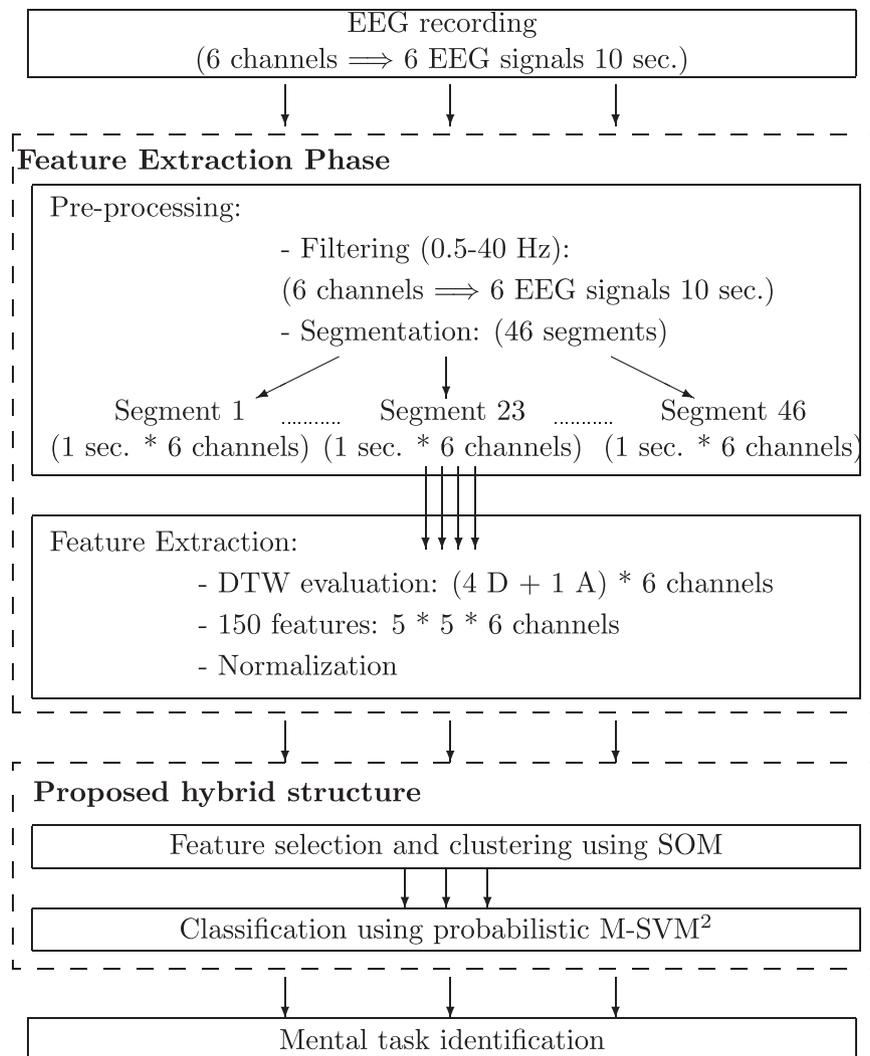


Fig. 1. Diagram scheme of the proposed method for one EEG recording.

resulting descriptor vectors are presented as input of a BPNN which achieved an average accuracy of 90.75% for subject 1 and five mental tasks.

The obtained results of different studies summarized in this section are given in Table 3.

### 3. Methods

In this section, we firstly give a description of the used database. Then, we expose details of the feature extraction method, and finally we present our proposed hybrid structure (Fig. 1).

#### 3.1. Description of the database

The EEG database used in this study was assembled by Keirn [13], and distributed online by Anderson [2]. It consists of 325 EEG recordings from seven subjects 21–48 years old performing five distinct mental tasks, namely:

- Baseline Task (BT), the subjects were invited to relax and think of nothing in particular.
- Multiplication Task (MT), all subjects were asked to solve a nontrivial multiplication problem (eg.  $28 \times 62$ ). The subjects were not vocalizing nor making movements when they were solving the problem.
- Letter Task (LT), subjects were instructed to mentally compose a letter to another person without vocalizing.

- Rotation Task (RT), the subjects were given 30 s to see a rotating 3D object. After removing the object, they were invited to imagine that object being rotating around an axis.
- Counting Task (CT), all subjects were asked to imagine a blackboard and to visualize numbers being written on it sequentially.

An electro-cap was used to obtain each EEG recording from six channels: C3, C4, P3, P4, O1 and O2, according to the 10–20 systems. The obtained signals were recorded for 10 s at 250 Hz during each task. We used all the subjects of the database except subject four because of the invalidation of the signals in several trials as reported by Liang et al. [17]. Each subject attended (1 session, 2 sessions, or 3 sessions) of five trials per session. Each trial contains five recordings, and each recording represents one of the five distinct mental tasks.

We note the existence of a seventh channel named ElectroOculoGraphy (EOG), that helps to locate the eye blinks, by performing a signal processing on it. In [14,17,35], the authors proposed detection schemes to locate and to perform ocular artifact correction. However, this channel was not used in this work, because the effects of eye blink is low at the central and occipital region where the six other channels were located [29].

#### 3.2. Feature extraction method

In this subsection, we describe the feature extraction method that includes the preprocessing and the parameter estimation steps. In the preprocessing step, we remove artifacts and divide the EEG signals into 46 1 s. EEG segments. Then, in the parameter estimation step, we use the DWT to extract from each segment the five main sub-bands that characterize the EEG signals  $\delta$ ,  $\theta$ ,  $\alpha$ ,  $\beta$  and  $\gamma$  bands, then we compute for each band and each EEG channel five descriptions to provide a feature vector of 150 components.

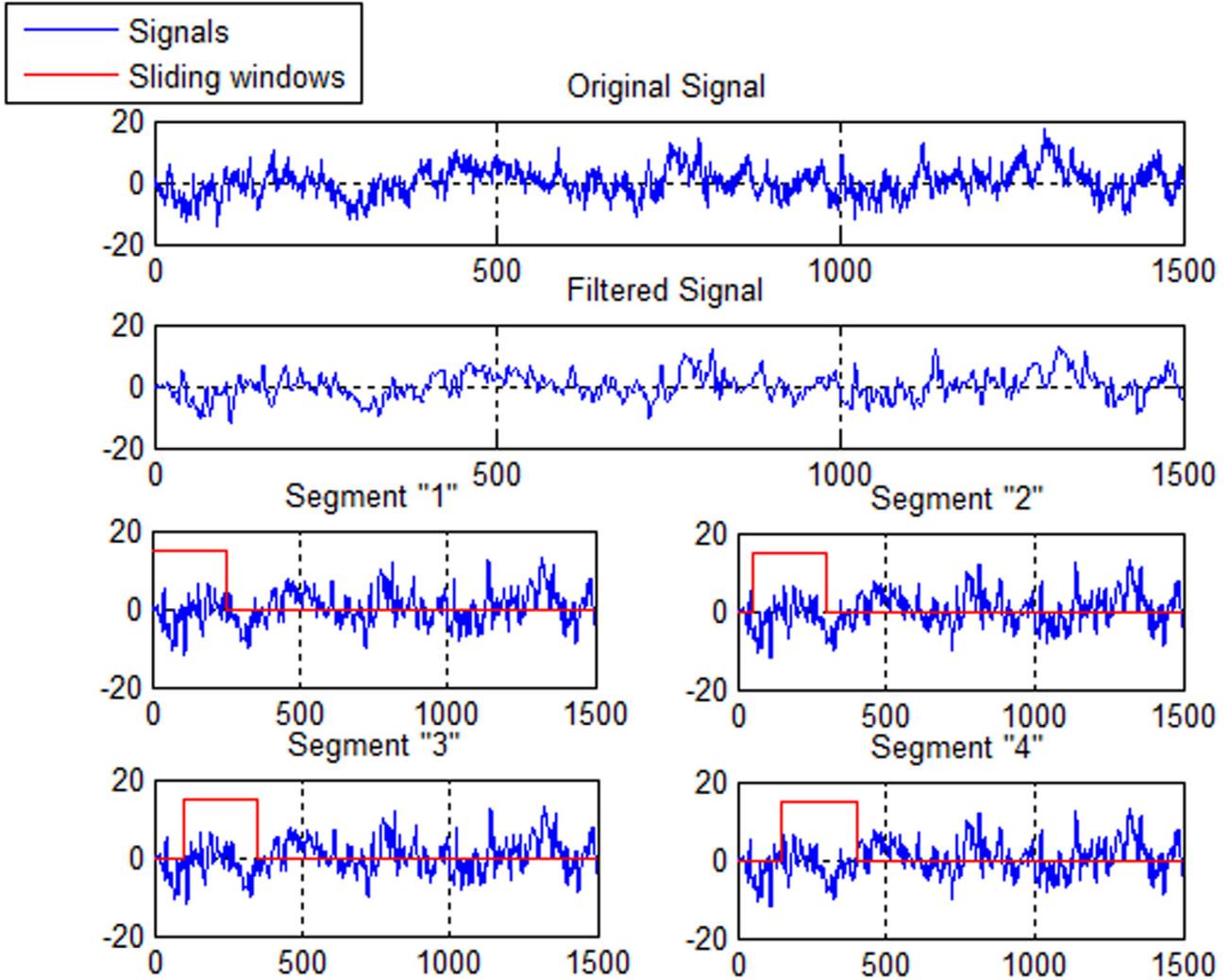


Fig. 2. Result of band-pass filter and segmentation.

### 3.2.1. Preprocessing

We first apply using a Hamming window, a Finite Impulse Response (FIR) band pass filter (of type 1 and order 49) with a cut-off frequency between [0.5, 40] Hz to include all the five sub-bands and remove artifacts of internal and external sources (muscle movements, power line interference, electrode movements, etc.). Then, each EEG signal is divided into 46 descriptions representing the content of a sliding window of 1 s with an overlap of 0.8 s between adjacent segments [35]. So, the first segment corresponds to a window located between [1, 250] samples, the next segment between [51, 300] samples, and so on. Fig. 2 illustrates the results obtained after filtering and segmentation.

### 3.2.2. Parameters estimation

The DWT is a powerful method able to represent the non-stationary signals such the EEG signals, in time and frequency domain. Moreover the coefficients at each stage of DWT are almost stationary and low auto-correlated, which provide better characteristics of signals.  $s+1$ . In DWT analysis, a signal is subdivided into an approximation (A) and a detail (D), obtained respectively by the use of low-pass and high-pass filters. The approximation at stage  $s$  is split into a next-level approximation and detail  $S+1$ , and the process is continued. In this study, we use 5 stages of decomposition and Daubechies 4 (db4) wavelets to subdivide each 1 s. EEG segment into five different frequency bands (see Fig. 3): A5 ( $\delta$ : 0.5–3.9 Hz), D5 ( $\theta$ : 3.9–7.8 Hz), D4 ( $\alpha$ : 7.8–15.3 Hz), D3 ( $\beta$ : 15.3–31.5 Hz) and D2 ( $\gamma$ : 31.5–40 Hz) according to the following expressions for approximation:

$$\left[ 0, \frac{F_s}{2^{(n+1)}} \right], \quad (1)$$

and detail:

$$\left[ \frac{F_s}{2^{(n+1)}}, \frac{F_s}{2^n} \right]. \quad (2)$$

We extract from the sub-bands (D2, D3, D4, D5 and A5), five statistical features as descriptors of the signal.

- Signal energy: The energy of a signal is calculated as following:

$$E_x = \frac{1}{N} \sum_{n=1}^N (x(n)^2). \quad (3)$$

Where,  $x(n)$  is the signal, and  $N$  is the number of samples.

- Signal Variance: The variance of a signal is determined by the following mathematical equation:

$$\sigma_x^2 = \frac{1}{N} \sum_{n=1}^N (x(n) - \bar{x})^2. \quad (4)$$

Where,  $x(n)$  is the signal,  $\bar{x}$  is the signal sample mean and  $N$  is the number of samples.

- Variance of the autocorrelation function of a signal: The autocorrelation function represents a metric to measure the similarity between a signal  $x(i)$  and its shifted version. Its variance is calculated according to the following equation:

$$\sigma_{R_{xx}(l)} = \sigma \left( \sum_{n=i}^{N-|k|-l} [x(n) - x(n-l)]^2 \right). \quad (5)$$

where,  $l$  is the time shift index.  $i=l$ ,  $k=0$  for  $l \geq 0$ , and  $i=0$ ,  $k=1$  for  $l < 0$ .

- Min ( $x(n)$ ): represents the smallest amplitude.
- Max ( $x(n)$ ): represents by the highest amplitude.

As mentioned in Section 3.1, the electrical activity of the brain is reflected by six electrodes C3, C4, P3, P4, O1 and O2, for each task. Hence, our feature vectors will

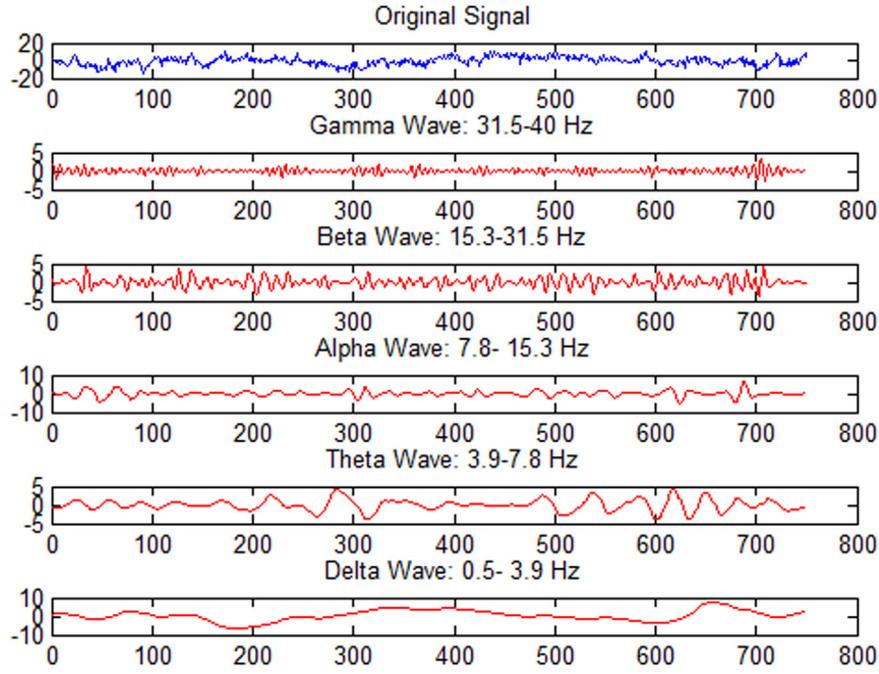


Fig. 3. Detail coefficients (D2, D3, D4 and D5) and the fifth approximation A5 for a given segment.

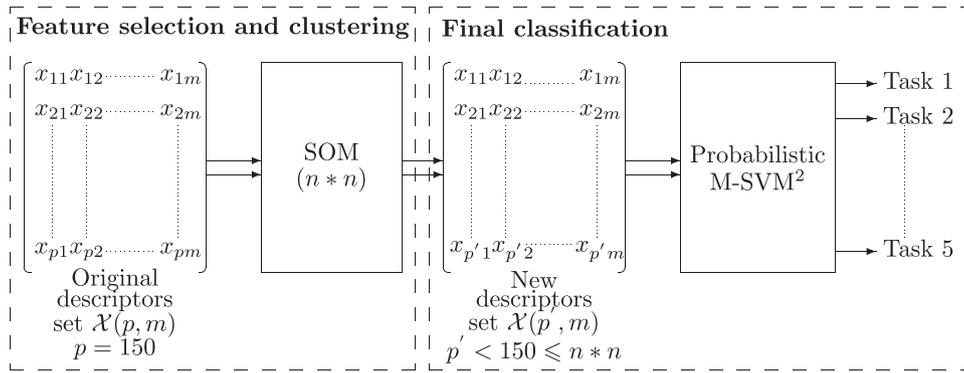


Fig. 4. Proposed hybrid structure.

be composed of 150 coefficients (6 channels \* 5 frequency bands \* 5 features). On the other hand, and in order to standardize all descriptors at the same level  $[-1, +1]$ , the normalization process is applied as follows:

$$\tilde{x}_{ij} = \text{tansig} \left( \frac{(x_{ij} - \bar{x}_j)}{\sigma_{x_j}} \right). \quad (6)$$

where  $\bar{x}_j$  and  $\sigma_{x_j}$  are respectively the mean and the standard deviation of the  $j^{\text{th}}$  parameter of the feature vectors, across each subject and each session. *tansig* is a hyperbolic tangent sigmoid transfer function.

### 3.3. Proposed hybrid structure

We report in this subsection, the methodology adopted to construct our hybrid system, based on SOM map and probabilistic M-SVM<sup>2</sup> to recognize a mental task (Fig. 4).

#### 3.3.1. Feature selection and clustering using SOM

In general, an observation is characterized by a set of descriptors that are not all informative. Indeed, some of them can be insignificant, correlated or not relevant to the classification task. Therefore, the selection of relevant parameters presents a major interest which enables to reduce the volume of information to be treated and consequently to reduce the complexity of the classification algorithms and also improve their generalization performance.

In this study, we propose a new alternative of feature selection using SOM algorithm. This choice is mainly based on the fact that the SOM algorithm has the particularity to cluster high-dimensional sets in an unsupervised manner. This approach can be considered as contextual since it takes into account the

interactions (correlations) between all parameters, and Wrapper because we use an MLP classifier to evaluate and select a subset of parameters.

### Unsupervised learning of an SOM

The simulated system is a two-dimensional SOM with  $n * n$  artificial neurons:

- Each input vector  $X1$  represents the line of the description space  $\mathcal{X}(p, m)$  (with  $m$  the number of samples of the learning set, and  $p$  the number of parameters = 150) and it is linked to each neuron of the map.
- The external connections of the neuron  $i \in [1, n * n]$  with the input vector  $X1$  is materialized by a synaptic weight vector  $V_i = [v_{i1}, v_{i2}, \dots, v_{im}]$ .
- Also, each neuron of the map is related to all the other neurons. The weights of the internal connections  $W_{ir}$  of the neuron  $i$  with its neighbors  $r(t)$  are given by the function called the “Mexican hat”.

During the learning phase, the network adapts itself to different input vectors by adapting the external map weight connection as following:

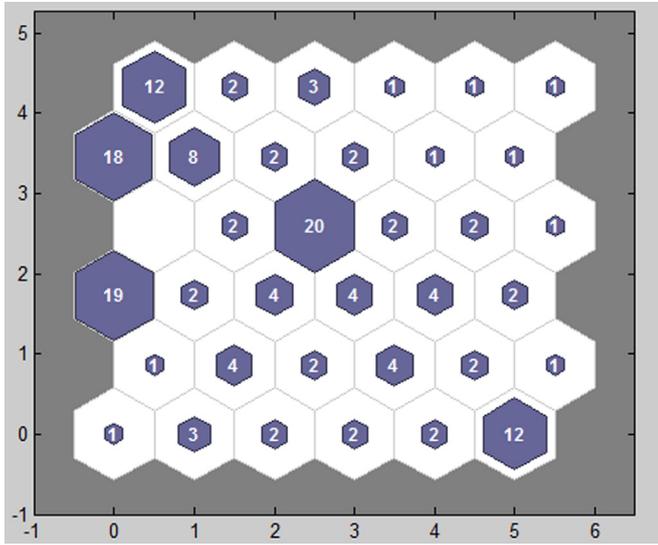


Fig. 5. Result of the feature selection and clustering using SOM.

- Firstly, the euclidean distance between the input vector and each weight vector is calculated according to the following expressions:

$$d_i(t) = \|X1(t) - V_i(t)\| \quad (7)$$

- Then, the neuron  $c$  that gives the minimum distance from the input vector is selected as the winner and characterized by the same distance:

$$d_c = \min(d_i(t)), \quad i \in [1, n^*n] \quad (8)$$

- Finally, the weights of the winner neuron and its neighbors  $r(t)$  are updated as follows:

$$V_c(t+1) = V_c(t)\alpha(t)\|X1(t) - V_i(t)\| \quad (9)$$

$$V_i(t+1) = V_i(t)\alpha(t)\|X1(t) - V_i(t)\|_{i \in r(t)} \quad (10)$$

with  $\alpha(t)$  the learning rate used to modulate the new weight;  $\alpha(t)$  and  $r(t)$  decrease with  $t$ .

At the end of the training process, we obtain a map that is self-organized in  $(n^*n)$  neurons. Each neuron represents one or more components. The most relevant parameters (that are independent of the other) are individually represented by a single neuron. However, the correlated parameters are grouped in the same neuron. So, we can extract the parameters that constitute each neuron of the obtained map. Also, there are neurons that have no descriptors and represent the death neurons, these neurons are

ignored and removed. The resulting map will then represent the new learning set (Fig. 5).

The parameters of each new example are first grouped in similar combinations to those of the obtained map in the learning step. Then, the new reduced vector is calculated by averaging the parameter values of each neuron. Therefore, we obtain a vector  $\mathcal{X}(p', m)$  with  $p' < 150 \leq n^*n$ .

### 3.3.2. Classification using probabilistic M-SVM<sup>2</sup>

The M-SVM<sup>2</sup> is a Kernel machines Method [26,25], that operates in the functional class induced by a Mercer kernel [5]. In what follows,  $\mathcal{X}$  is the description space,  $\mathcal{Y}$  is the set of the categories  $\in \llbracket 1, Q \rrbracket$  with  $3 \leq Q < +\infty$ ,  $\kappa$  is a kernel of Mercer on  $\mathcal{X}^2$ ,  $(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})$  is the Reproducing kernel Hilbert Space (RKHS) with kernel  $\kappa$ , and  $C$  is the penalty parameter. For  $m \in \mathbb{N}^*$ , let  $\mathbf{z}_m = ((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times \llbracket 1, Q \rrbracket)^m$ ,  $\mathbf{y}_m = (y_i)_{1 \leq i \leq m}$ ,  $\xi \in \mathbb{R}^{Qm}(\mathbf{y}_m)$ . A  $Q$ -category M-SVM<sup>2</sup> with a kernel  $\kappa$ , a RKHS  $(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})$ , and a “training set”  $\mathbf{z}_m$  is a large margin discriminant model trained by solving a convex quadratic programming problem of the form: (for more detail see [10]).

#### Learning problem of an M-SVM<sup>2</sup>, primal formulation

$$\min_{h, \xi} \left\{ \frac{1}{2} \sum_{k=1}^Q \|w_k\|^2 + C \xi^T M \xi \right\} s.t. \quad \begin{cases} \forall i \in \llbracket 1, m \rrbracket, \forall k \in \llbracket 1, Q \rrbracket \setminus \{y_i\}, h_k(x_i) \leq \frac{1}{Q-1} + \xi_{ik} \\ \sum_{k=1}^Q h_k = 0 \end{cases}$$

where  $M \in \mathcal{M}_{Qm, Qm}$  is a matrix of rank  $(Q-1)m$ , with a general term defined as follows:

$$m_{ik, jl}^{(2)} = (1 - \delta_{y_i, k})(1 - \delta_{y_j, l}) \left( \delta_{k, l} + \frac{\sqrt{Q-1}}{Q-1} \right) \delta_{i, j},$$

In our experiments we used the M-SVM package (M-SVMpack) [15], which allowed us to implement all the different M-SVM models (WW model, CS model, LLW model, and M-SVM<sup>2</sup> model).

**Post-processing the outputs of the M-SVM<sup>2</sup>** To obtain the probabilistic M-SVM<sup>2</sup>, we extend Platts bi-class solution [20] to the multi-category case, according to the softmax function, [23], given by Eq. (11):

$$\forall k \in \llbracket 1, Q \rrbracket, \quad \tilde{h}_k = \frac{\exp(h_k)}{\exp\left(\sum_{l=1}^Q h_l\right)} \quad (11)$$

## 4. Results

As mentioned in Section 3.1, we have used all the subjects of the database except subject four. Each subject was evaluated separately, session by session,

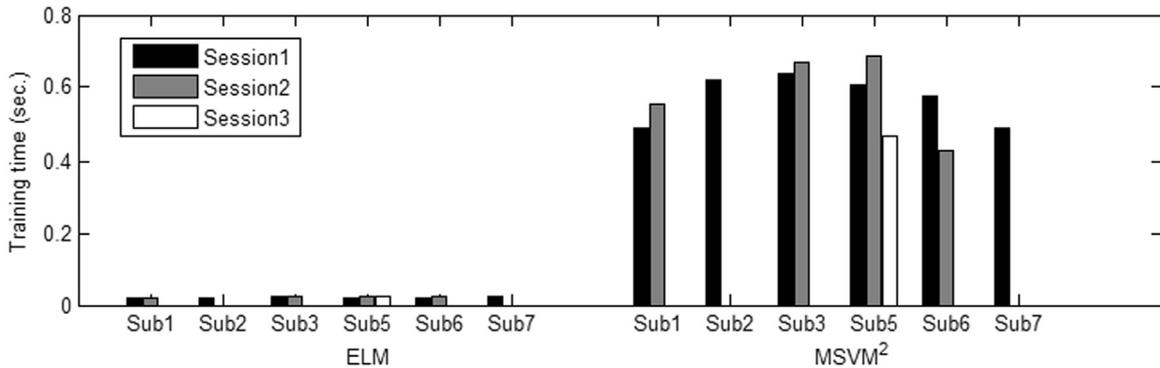


Fig. 6. Training time (s) of ELM and MSVM<sup>2</sup>.

**Table 1**  
classification results of the proposed hybrid structure (%).

Subject and session	BT	MT	LT	RT	CT	Accuracy+ confidence interval of 95%	Reject	Reject+accuracy
<b>Subject 1</b>								
<b>Session1</b>	84.78	90.43	98.69	87.82	84.78	89.3 [84.64–92.66]	6.60	95.9
<b>Session2</b>	96.08	99.56	93.04	84.78	86.08	91.90 [86.94–94.29]	6.17	98.07
<b>Subject 2</b>								
<b>Session1</b>	70.43	83.47	89.56	95.65	91.30	86.08 [81.01–89.96]	7.73	93.81
<b>Subject 3</b>								
<b>Session1</b>	70.00	86.52	76.52	65.65	72.60	74.25 [68.23–79.47]	18.95	93.2
<b>Session2</b>	87.82	79.13	86.52	77.82	77.39	81.73 [76.22–86.18]	14.52	96.25
<b>Subject 5</b>								
<b>Session1</b>	92.17	64.34	88.26	82.60	66.08	78.69 [72.94–83.48]	17.73	96.42
<b>Session2</b>	83.47	65.65	88.69	73.47	66.08	75.47 [69.52–80.58]	18.08	93.55
<b>Session3</b>	86.08	86.08	94.34	83.47	66.95	83.38 [78.02–87.63]	10.17	93.55
<b>Subject 6</b>								
<b>Session1</b>	95.21	91.73	95.65	85.21	91.30	91.82 [87.55–94.71]	7.73	99.55
<b>Session2</b>	90.43	97.39	86.95	83.91	95.21	90.77 [86.33–93.87]	7.04	97.81
<b>Subject 7</b>								
<b>Session1</b>	96.08	87.82	97.82	86.08	87.39	91.03 [86.63–94.08]	7.82	98.85
<b>Average</b>								98.08%
<b>sd ±</b>	–	–	–	–	–	–	–	2.3

because the mental state and reasoning of persons performing the same task can be different. Then, to illustrate the results, five-fold cross-validation procedure was performed using five distinct trials. We trained on three trials (15 tasks\*46 segments=690 segments), used a fourth trial on validation (5 tasks\*46 segments=230 segments), and tested on the fifth trial (5 tasks\*46 segments=230 segments).

#### 4.1. Selection of SOM and M-SVM<sup>2</sup> parameters

In general, to find the best hyperparameters for any method, it is recommended to perform the model selection. This is done on the validation data. Then, the best set of parameters obtained is used to build the model for future testing.

For the SOM Model, the main parameter that should be optimized is the one related to the size of the map. So, to find the best map (using the training set), five ( $n * n$ ) topologies have been tested (with  $n \in [4, 8]$ ) and each topology has been repeated twenty times, for a total of 100 epochs for each validation set. Then, for each epoch, we modify the validation set according to the obtained map (resulting learning set), then an MLP network was used for evaluating the performance of each topology (Wrapper method). The MLP was trained on the resulting learning set and tested on modified validation set. Finally, the best selected map will allow to generate the new testing set. Thus, the learning and the validation sets (of the best selected map) added to the resulting testing set will be used on the second floor.

As for the M-SVM<sup>2</sup> model, the parameters that should be optimized are those related to the penalty parameter and kernel parameter. We note that three types of kernels are available in the MSVMpack (Linear, Polynomial, and Gaussian RBF). Here, we save the “Gaussian RBF” kernel which is the most used one. So, for M-SVM<sup>2</sup> with “Gaussian RBF” kernel, two parameters are optimized:  $C$  and kernel parameter  $\gamma$  over five-fold, cross-validation. For each fold set we test  $15 \times 15 = 225$  different combinations of  $(C, \gamma)$ ,  $\gamma = [2^4, 2^3, 2^2, \dots, 2^{-10}]$  and  $C = [2^{12}, 2^{11}, 2^{10}, \dots, 2^{-2}]$  [12]. For each combination, the validation performance is calculated by training on the training set and testing on the validation set. In the prediction phase, we train on the same training set using the pair of  $(C, \gamma)$  that gives the best validation rate, and predict on the test set.

#### 4.2. Evaluation of the proposed classification scheme

Table 1 exposes the results obtained with the proposed hybrid structure as follows:

- Accuracy for each task that represents an average accuracy per mental task over the five-fold cross-validation that were conducted.
- The test accuracy for each session that represents an average accuracy of all mental tasks per session.
- The reject accuracy for each session that represents an average accuracy of accumulated rejects over the five-fold cross-validation.

**Table 2**  
learning and testing time (in s) for MSVM<sup>2</sup>.

Session	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7
<b>Session1</b>							
<b>Learning</b>	0.488	0.620	0.640	–	0.608	0.574	0.486
<b>Testing</b>	0.012	0.020	0.016	–	0.023	0.017	0.020
<b>Session2</b>							
<b>Learning</b>	0.554	–	0.668	–	0.688	0.428	–
<b>Testing</b>	0.01	–	0.004	–	0.008	0.004	–
<b>Session3</b>							
<b>Learning</b>	–	–	–	–	0.464	–	–
<b>Testing</b>	–	–	–	–	0.014	–	–

Also, to associate a degree of reliability to the proposed system, we used the following statistical measure proposed by [4]. It allows to give the interval of correct classification rate for future test sets, with a confidence rate of  $\alpha\%$ .

$$I(\alpha, N) = \frac{T + \frac{Z_\alpha^2}{2N} \pm Z_\alpha \sqrt{\frac{T(1-T)}{N} + \frac{Z_\alpha^2}{4N^2}}}{1 + \frac{Z_\alpha^2}{N}} \quad (12)$$

where  $N$  is the size of testing set,  $T$  is the correct classification rate, and  $Z_\alpha = 1.96$  for  $\alpha = 95\%$ .

On the other hand, we post-process the outputs of the M-SVM<sup>2</sup> model in order to generate the class posterior probability estimates. Then, if the probability associated to a class is higher than 0.75, the task is considered correctly classified. Otherwise the task is assigned to the reject class. It is more informative to take a decision if we know that there is 90% of chance that the task belongs to a given class.

According to the obtained results, it can be seen that the proposed classification system is able to produce a promising accuracy between 74.25% and 91.90%. In fact, one can see from the obtained confidence intervals, that even in the worst case, the proposed system is able with a confidence of 95%, to achieve a good classification rate for future test sets.

We also notice that, the accuracy of the reject class (per session) in addition to the accuracy of the same session is around (96.08% $\pm$ 2.3) for all subjects (where sd is the standard deviation and 96.08% is an average). This observation proves, that the number of misclassified tasks achieved by the proposed system is low, even though it is unable to well classify a given task (it is better to say “I don’t know” than giving a false prediction).

The learning time and the testing time of MSVM<sup>2</sup> classifier are reported in Table 2 as an average training/testing time per session over the five-fold cross-validation that were conducted.

**Table 3**  
Results of related works (%).

Method	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Average
<b>AR+BQ [14]</b> 2 classes	-	-	-	-	-	-	-	84.16
<b>(PS+ASR)+ENN [21]</b> 2 classes	79.80	69.10	70.20	80.30	-	-	-	74.85
<b>AR+ELM [17]</b> 5 classes	S2: 86.70	S1: 75.99	S2: 57.37	-	S3: 62.79	S1: 72.62	S1: 74.81	71.71
<b>WPE+SVM [35]</b> 5 classes	S2 76.30	S1 68.50	-	-	-	-	-	72.4
<b>EMD+LDA [8]</b> 2 classes	95.00	92.45	80.93	92.73	85.13	94.61	97.36	91.17
<b>DWT+BPNN [29]</b> 5 classes	S2: 90.75	S1: 73.30	-	-	-	-	-	82.02
<b>Our method</b> 5 classes	S2: 91.90	S1: 86.08	S2: 81.73	-	S3: 83.38	S1: 91.82	S1: 91.03	87.65

It can be seen from the table, that the accumulated testing time is between 0.004 s and 0.023 s, and accumulated training time is between 0.464 s and 0.764 s. This concluded that the MSVM<sup>2</sup> requires a reasonable and proper computational time for real applications.

## 5. Discussion

To prove the contribution of the proposed approach to enhance the performances of BCI, it is interesting to compare our results with those of methods summarized in Section 2. (See Table 3).

- In [14] five subjects were considered and ten combinations case of task pairs were tested. The results were given as the total classification accuracy, averaged across all ten task pairs and over all the five subjects.
- Palaniappan [21] and Diez et al. [8], have respectively considered four and five subjects. The results were obtained for each subject as the average values over all the ten task pairs and over all the attended sessions. Also, the overall mean accuracy of the considered subjects was reported.
- For [17,35,29] and the proposed method, five tasks were tested simultaneously. The results were presented as the best average accuracy of the attended sessions per subject, and the global rate of reported sessions over the considered subjects (two subjects were utilized in [29,35], and six subjects were used in [17] and our method).

However, so that our comparison will be completely fair, it is necessary to take into consideration the number of classes and the number of subjects considered in the different studies. As consequence, we focused our comparison only with the summarized methods that consider five classes.

- Liang et al. [17] achieve classification accuracy between 57.37% and 86.70%, and global rate of 71.71%. The proposed classification system is able to produce a promising accuracy between 81.73% and 91.90%, and global rate of 87.65%.
- Zhiwie and Minfen [35], achieve a classification accuracy of 76.30% for subject 1 and 68.50% for subject 2. Also, Tolic and Jovic [29] archive an accuracy of 90.75%, 73.30% for subjects 1, and subjects 2, respectively. The proposed method exceeds these results with an accuracy of 91.90% for subject 1 and 86.08% for subject 2.

As mentioned above, the results were presented in Table 3 as the best average accuracy of the attended sessions per subject.

Since in general, when we want to adopt a real system at a given subject, it is necessary to perform several learning sessions, and maintain thereafter the one that gives the best performance for efficient future use.

Also, we wanted to discuss the computational time between MSVM<sup>2</sup> and compared methods. However, the presented related work, have not exposed the computational time of their classifiers, except [17], that exposed the training time. So, we expose in Fig. 6 the training time of ELM and MSVM<sup>2</sup>. We can observe that the computation time of ELM is smaller in comparison with that of the MSVM<sup>2</sup>. However, in general, we does not take into consideration, the time required for adjusting the parameters of a real system, and the time required for its learning in order to adopt it to a given subject. Which is more important, is the necessary time to recognize a given mental task. Indeed, as mentioned in Section 4 the MSVM<sup>2</sup> has a low generalization time.

## 6. Conclusion

In this study, we have proposed and implemented a new hybrid structure based on two stages to discriminate between five distinct tasks taken from the Keirn and Aunon database. In the first stage, we have proposed a new alternative of feature selection using SOM algorithm with unsupervised learning. The map received as input the 150 components obtained from wavelet decomposition. In the second stage, we have used a Probabilistic M-SVM<sup>2</sup> with supervised learning for the final classification. The M-SVM<sup>2</sup> received as input the vector parameters that was reduced by the SOM map. The proposed approach surpassed the results from the previous systems proposed in this framework, with a best average accuracy across sessions per subject between 81.73% and 91.90%. These observations conclude that our method can be an excellent alternative to enhance the performances of BCI systems.

## References

- [1] Ahmadi A, Jafari R, Hart J. Light-weight single trial eeg signal processing algorithms: computational profiling for low power design. In: IEMBS'11; 2011. p. 4426–30.
- [2] Anderson C. URL (<http://www.cs.colostate.edu/eeg/main/data/>); 1989.
- [3] Anderson C, Devulapalli S, Stolz E. Eeg signal classification with different signal representation. In: Girosi, F., Makhoul, J., Manolakos, E., Wilson, E. (Eds.), IEEE Service Centre. Piscataway, NJ; 1995, Vol. V. p. 475–83.
- [4] Bennani Y, Bossaert F. Predictive neural networks for traffic disturbance detection in the tele-phonenetwork. IEEE Comput Eng Syst Appl 1996;1085–90.
- [5] Berline A, Thomas-Agnan C. Reproducing Kernel Hilbert spaces in probability and statistics. Boston: Kluwer Academic Publishers; 2004.
- [6] Chai R, Ling S, Hunter G, Tran Y, Nguyen H. Brain-computer interface classifier for wheelchair commands using neural network with fuzzy particle swarm optimization. IEEE J Biomed Health Inform 2014;8(5):1614–24.
- [7] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. J Mach Learn Res 2001;2:265–92.
- [8] Diez P, Mut V, Laciari E, Torres A, Avila E. Application of the empirical mode decomposition to the extraction of features from eeg signals for mental task classification. In: EMBC'09; 2009. p. 2579–82.
- [9] Estrada E, Nazeran H, Sierra G, Ebrahimi F, Setarehdan S. Wavelet-based eeg denoising for automatic sleep stage classification. In: CONIELECOMP'11; 2011. p. 295–8.
- [10] Guermeur Y, Monfrini E. A quadratic loss multi-class svm for which a radius-margin bound applies. Informatica 2011;22(1):73–96.
- [11] Hariharan H, Vijejan V, Sindhu R, Divakar P, Saidatul A, Yaacob Z. Classification of mental tasks using stockwell transform. Comput Elect Eng 2014;40:1741–9 Elsevier.
- [12] Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 2002;13:415–25.
- [13] Keirn Z. Alternative modes of communication between man and machines. masters thesis. Electrical Engineering Department. USA: Purdue University; 1998.
- [14] Keirn Z, Aunon J. A new mode of communication between man and his surroundings. IEEE Trans Biomed Eng 1990;37(12):1209–14.

- [15] Lauer F, Guermeur Y. MSVMpack: a multi-class support vector machine package. *J Mach Learn Res* 2011;12:2269–72 URL (<http://www.loria.fr/lauer/MSVMpack>).
- [16] Lee Y, Lin Y, Wahba G. Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data. *J Am Stat Assoc* 2004;99(465):67–81.
- [17] Liang N, Saratchandran P, Huang G, Sundararajan N. Classification of mental tasks from eeg signals using extreme learning machine. *Int J Neural Syst* 2006;16(1):29–38.
- [18] Sabetia M, Boostania R, Priceb SK, G. Selection of relevant features for eeg signal classification of schizophrenic patients. *ELSEVIER. Biomed Signal Process Control*. 2(2); 2007. p. 122–34.
- [19] Majkowski A, Kolodziej M, Rak R. Implementation of selected eeg signals processing algorithms in asynchronous bci. In: *MeMeA'12*; 2012. p. 1–3.
- [20] Nasehi S, Pourghassem H. Mental task classification based on hmm and bpnn. In: *CSNT'13*; 2013. p. 210–4.
- [21] Palaniappan R. Utilizing gamma band to improve mental task based brain-computer interface designs. *IEEE Trans Neural Syst Rehabil Eng* 2006;14(3):299–303.
- [22] Picton T, Lins O, Scherg M. The recording and analysis of event-related potentials. In: *Boller F, Grafman J. (Eds.), Handbook of Neuropsychology*. Vol. 10. Elsevier, Ch. 1; 1995. p. 3–73.
- [23] Platt J. Probabilities for SV machines. In: *Smola A, Bartlett P, Schölkopf B, Schuurmans D. (Eds.), Advances in Large Margin Classifiers*. Cambridge, MA: The MIT Press. Ch. 5; 2000. p. 61–73.
- [24] Pregenzer M, Pfurtscheller G. Frequency component selection for an eeg-based brain to computer interface. *IEEE Trans Rehabil Eng* 1999;7(4):413–9.
- [25] Schölkopf B, Smola A. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA: The MIT Press; 2002.
- [26] Shawe-Taylor J, Cristianini N. *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press; 2004.
- [27] Shedeed H. A new method for person identification in a biometric security system based on brain eeg signal processing. In: *WICT'11*; 2011, p. 1205–10.
- [28] Hastie T, Tibshirani R, JF. *The elements of statistical learning: data mining, inference, and prediction*. Second ed. New York: Springer; 2009.
- [29] Tolic M, Jovic F. Classification of wavelet transformed eeg signals with neural network for imagined mental and motor tasks. *Int J Neural Syst* 2013;45(1):130–8.
- [30] Vaid S, Singh P, Kaur C. Eeg signal analysis for bci interface: a review. In: *Proceedings of IEEE conference on advanced computing and communication technologies*; 2015. p. 143–7.
- [31] Weston J, Watkins C. *Multi-class support vector machines*. Tech. Rep. CSD-TR-98-04. Royal Holloway, University of London, Department of Computer Science; 1998.
- [32] Wolpaw J, Birbaumer N, Heetderks W, McFarland D, pH, P, Schalk G, et al. Brain-computer interface technology: a review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*. 8; 2000. p. 164–73.
- [33] Woodman G. A brief introduction to the use of event-related potentials (erps) in studies of perception and attention. *Atten Percept Psychophys* 2010;72(8). <http://dx.doi.org/10.3758/APP.72.8.2031>.
- [34] Wang Y, X G, Gao S. Common spatial pattern method for channel selection in motor imagery based brain-computer interface. *IEEE Engineering in Medicine and Biology*. 5; 2005. p. 5392–5.
- [35] Zhiwei L, Minfen S. Classification of mental task eeg signals using wavelet packet entropy and svm. In: *ICEMI'07*; 2007. p. 3–907–3–910.