ORIGINAL RESEARCH

Assessment of feature selection and classification methods for recognizing motor imagery tasks from electroencephalographic signals

Roberto Vega*[1], Touqir Sajed[1], Kory Wallace Mathewson[1], Kriti Khare[1], Patrick M. Pilarski[2], Russ Greiner[1], Gildardo Sanchez-Ante[3], Javier M. Antelis[3]

[1] *Deparment of Computing Science, University of Alberta, Edmonton, Canada*
[2] *Department of Medicine, University of Alberta, Edmonton, Canada*
[3] *Tecnológico de Monterrey, campus Guadalajara, Zapopan, México*

Received: June 30, 2016 **Accepted:** August 9, 2016 **Online Published:** September 1, 2016
DOI: 10.5430/air.v6n1p37 **URL:** http://dx.doi.org/10.5430/air.v6n1p37

ABSTRACT

Recognition of motor imagery tasks (MI) from electroencephalographic (EEG) signals is crucial for developing rehabilitation and motor assisted devices based on brain-computer interfaces (BCI). Here we consider the challenge of learning a classifier, based on relevant patterns of the EEG signals; this learning step typically involves both feature selection, as well as a base learning algorithm. However, in many cases it is not clear what combination of these methods will yield the best classifier. This paper contributes a detailed assessment of feature selection techniques, *viz.*, squared Pearson's correlation (R^2), principal component analysis (PCA), kernel principal component analysis (kPCA) and fast correlation-based filter (FCBF); and the learning algorithms: linear discriminant analysis (LDA), support vector machines (SVM), and Feed Forward Neural Network (NN). A systematic evaluation of the combinations of these methods was performed in three two-class classification scenarios: rest *vs.* movement, upper *vs.* lower limb movement and right *vs.* left hand movement. FCBF in combination with SVM achieved the best results with a classification accuracy of 81.45%, 77.23% and 68.71% in the three scenarios, respectively. Importantly, FCBF determines, based on the feature set, whether a classifier can be learned, and if so, automatically identifies the subset of relevant and non-correlated features. This suggests that FCBF is a powerful method for BCI systems based on MI. Knowledge gained here about procedural combinations has the potential to produce useful BCI tool, that can provide effective motor control for the users.

Key Words: Motor imagery, Classification, Feature selection, Machine learning

1. INTRODUCTION

A brain-computer interface (BCI) is a system that provides a person a way to interact with the external world without the use of the peripheral nervous system nor the motor pathways. They can be used, for example, to restore functionality of impaired limbs.[1,2] Most of the research in BCI is based on the non-invasive, electroencephalogram (EEG) technique, as this technology is reliable, affordable, portable, and provides high temporal resolution of the brain signals.[3] Many of these systems record and process the ongoing brain activity associated with a task performed by the user, such as moving their hands up, which helps to identify the control signals that can then be used to control external devices.[4,5] The most common task is the motor imagery (MI), wherein a sub-

*__Correspondence:__ Roberto Vega; Email: rvega@ualberta.ca; Address: Deparment of Computing Science, University of Alberta, Edmonton, AB, T6G 2E8, Canada.

ject imagines moving a specific part of his/her body, without actually executing the movement. For example, it is common to ask the subjects to imagine that they are moving their right (or left) hand or foot. The BCI system would then learn the brain patterns, such as changes in the power spectrum of the brain signals, associated with this action.[6] This task is preferred to other mental tasks (such as selective attention,[7] or self-regulation of brain rhythms[8, 9]) because it is more intuitive and less exhaustive for the user. One can then produce a BCI for this task: whenever the tool detects a brain pattern that matches the MI of moving a specific body part, the system will then actuate a corresponding external device: *e.g.*, when the pattern corresponds to "move left hand", the tool will move the physical device that corresponds to this action. Therefore, a key component for achieving this goal is learning which brain activity corresponds to which physical action.

The challenge of classifying between the actual movement of the right or left hand, as opposed to just imaging that movement, using EEG signals has been addressed before with an accuracy over 96%;[10] however, this accuracy drops substantially for the scenario of MI. Morash *et al.* used a naive Bayes classifier to classify MI of the right hand, left hand, tongue, and right foot. Their reported accuracy was, on average, below 65% in binary classification, and below 50% in multiclass classification.[11] Schlogl *et al.* reported an accuracy around 63%, on average, also in a 4-class problem using MI,[12] while Ge *et al.* reported an accuracy around 75% in a similar task.[13]

There are some works that report higher accuracy when classifying MI tasks, for instance Pfurtscheller *et al.* reported an accuracy of approximately 80% when distinguishing between moving the left hand vs the right hand.[6] However, they only reported the accuracy on 3 (out of 10) participants in the experiments – selected because they had the largest EEG differences. This situation is not uncommon. A different study excluded the data from 2 (out of 8) subjects from the experiments due to failure to adequately participate in the experiment, meaning that the signals that the researches expected to analyze were not present in their recordings.[11] This was also done after analyzing their data. Unfortunately, there is no standard way of determining if the recorded data contains patterns that can be encoded in a classifier or if the recordings should be discarded, which constitutes an important problem with EEG data: How to determine if a recorded data is suitable for analysis?

The general approach for applying machine learning in EEG involves four steps: 1) pre-processing: to improve the signal quality by reducing noise and artifacts; 2) feature extraction: to compute relevant task-related information from the pre-processed EEG signals; 3) feature selection: to find the features with high discriminative power; and 4) learning a classifier: to produce a classifier, by applying machine learning algorithms to the selected features of a pre-processed training dataset, that can then identify which movements a user is imaging. Basically, the result of this process is a classifier that can use the EEG signals, recorded during the execution of the mental task, to identify the required control signals that can be used to trigger a device. Although this approach is well known, there are many methodologies to perform feature selection and to learn a classifier; unfortunately, how to select the best combination of methods for this particular task is still an open problem.

Previous works have investigated the performance of several classifiers (linear discriminant analysis [LDA], support vector machines, or k-nearest neighbors for classification) with a single feature selector[10, 12] or several feature selectors (principal component analysis [PCA], locality linear projection, Fisher discriminant analysis, or Wilcoxon rank sum test) with a single classifier.[14, 15] Despite the high variety of techniques, very few studies have addressed the evaluation of combination of methods in an integrated way (*i.e.* several feature selection algorithms and several learning algorithms).

These integrated evaluations have been performed in the past for different tasks involving EEG signals, such as motor tasks or sleep stage classification.[16, 17] Bai *et al.*[16] systematically investigated how combining different methodologies for spatial filtering, temporal filtering, feature extraction and classification can influence the discrimination accuracy on motor-related tasks. They highlighted the difficulty in determining the most effective way of classifying EEG signals because there are no systematic approaches, and previous studies usually investigated several techniques independently, making it difficult to compare their efficiency. Sen *et al.*[17] made a similarly analyzed sleep stage classification, and found that the selection of the learning algorithm made an important difference in the classification, but the influence of the feature selection algorithm was imperceptible. For the case of MI, Koprinska *et al.* performed a similar study.[18] Unfortunately, they did not include non-linear feature selection techniques such as fast correlation-based filter nor kernel principal component analysis (kPCA), nor typical classification algorithms such as LDA or support vector machines.

This work aims to fill this gap in the MI literature by performing a systematic evaluation of different combinations of feature selection and classification algorithms in the recognition of MI tasks from EEG signals. For feature selection we tested two linear methods: squared Pearson's correlation (R^2)

and PCA; and two nonlinear methods: fast correlation-based filter (FCBF) and kPCA. For classifiers, we tested two linear methods: LDA and linear support vector machine (SVML); and two nonlinear methods: support vector machine with a radial basis function kernel (SVMR) and a feed-forward neural network (NN). All combinations of these methods were evaluated in three classification scenarios (each using EEG signals recorded in a MI task): rest *vs.* movement, upper *vs.* lower limb movement and right *vs.* left hand movement. Inclusion of these methods turned out to be very important, since in our experiments the combination of SVM and FCBF achieved the best results. FCBF also has several advantages over other well known feature selection techniques, as FCBF can automatically determine the number of relevant, non-correlated features to use, as well as determine if there is enough information in the data to perform accurate classification.

The rest of the paper is organized as follows: the EEG dataset, the feature selections and classification methods, and the evaluation process are described in Section 2; the results are described in Section 3; finally Sections 4 and 5 discuss the results and present the conclusions respectively.

2. Methods

All combinations of the proposed feature selection and classification methods were evaluated in three two-class classification scenarios using EEG signals from the publicly available BCI Competition 2008-Graz dataset A.[19,20]

2.1 EEG dataset

This dataset consists of EEG signals recorded from nine healthy subjects performing cue-based MI of the left hand, right hand, both feet and tongue. During the experiments, the subjects sat in front of a computer screen while auditory and visual cues instructed them on the execution of the task. Figure 1 presents the time sequence of a single trial. First, a visual and auditory cue (in the form of a fixation cross and a beep, respectively) indicated the beginning of the trial. Two seconds later, a visual cue in the form of an arrow pointing either to the left, right, up or down was presented on the screen for 1.25 seconds. The direction of the arrow indicated the subsequent MI of the left hand, right hand, both feet or tongue. The participants carried out the task until the fixation cross disappeared from the screen. Then, the participants had a short break of around 1.5 - 2.5 seconds that allowed them to relax before the next trial. For each subject, the experiment consisted of two sessions performed in two different days. Every sessions consisted of 72 trials per each body part.

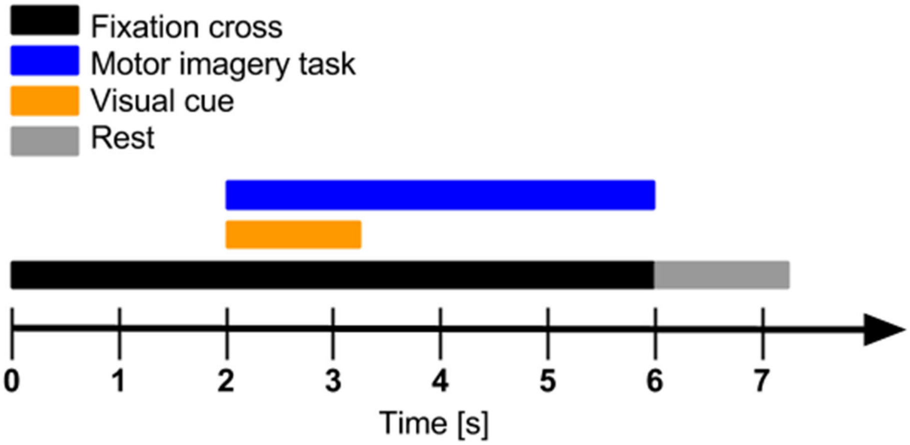

Figure 1. Temporal sequence of a single trial during the execution of the experiment of the BCI Competition 2008 – Graz dataset A. The next trial starts just after the rest period.

EEG signals were recorded from 22 Ag/AgCI electrodes located around the sensorimotor cortex according to the 10/20 system. The signals were recorded at a sampling frequency of 250 Hz with the reference at the left mastoid, and ground at the right mastoid. After the experimental sessions, EEG signals were bandpass filtered between 0.5 Hz and 100 Hz, and 50 Hz notch filtered to reduce power line interference. Then, trials were extracted using the start of the MI as reference. Finally, the experimenters identified and marked the artifact-free trials and the noisy trials (contaminated with electrooculographic and electromyographic activity) by visual inspection. In our present study, we only used trials that were marked as artifact-free, and each trial was trimmed to be only -3 to 3 seconds, thus the time interval [-3, 0)s corresponds to rest, while the time interval [0, 3)s corresponds to the execution of the MI task. In summary, the number of trials across all subjects, sessions and MI of the body parts was on average 65 ± 6 (minimum of 49 and maximum of 72).

2.2 EEG Analysis: Event-related desynchronization

EEG data were analyzed using a time-frequency analysis based on wavelets.[21] For each trial, the time-resolved power spectra was computed using complex Morlet wavelets in the frequency range between 2 Hz to 40 Hz at 1 Hz of resolution. For the time t and frequency f, the family of wavelets was $w(t,f) = Ae^{-t^2/2\sigma_t^2}$ with $A = \left(\sigma_t \pi^{1/2}\right)^{-1/2}$ and $\sigma_t = \left(2\pi\sigma_f\right)^{-1}$, which is characterized by the constant trade-off ratio f/σ_f that is typically fixed to 7 for the analysis of EEG signals.[22] Then, the average time-resolved power spectrum was computed across trials to compute the Event-Related De-Synchronization:

$$ERDS_i(t,f) = 100 \times \left[P_i(t,f) - P_{i,rest}(f)\right] / P_{i,rest}(f) \quad (1)$$

of each electrode i, which is the percentage of power increase (or decrease) relative to the rest interval [-3, 0)s, where $P_i(t,f)$ is the power spectra at time t and frequency f and $P_{i,rest}(f)$ is the average of power spectra at the rest interval [-3, 0)s at frequency f. Finally, we identify which ERDSs are significant using a bootstrap analysis at the significance level of α =0.05[23] using as baseline the time interval of the rest phase [-3, 0)s. Therefore, a significant event-related power decrease (aka ERD -*i.e.*, cortical activation state, which implied a decrease in synchrony of the underlying neuronal populations) is observed as a negative value, a significant event-related power increase (aka ERS -*i.e.*, cortical idling state) is observed as a positive value, and no significant power changes are observed as zero values.

2.3 Feature extraction: Power spectral density

It is well known that, in MI scenarios, the spectral power of alpha and beta rhythms (brain signals whose frequencies are between 8 Hz and 31 Hz.) changes during the imagined movement, especially in electrodes placed above the sensory-motor cortex contralateral to the moved body part.[24,25] In consequence, the standard approach for feature extraction in MI is to compute the power spectral density, which is also the most robust methods for feature extraction in MI tasks, and is the most preferred method for the spectral analysis of short segments and noisy signals, such as the EEG in MI tasks.[26]

There are several methods for the extraction of features related to the power spectrum of a signal. Herman, *et al.* compared the effect of using power spectral density, atomic decompositions, time-frequency energy distributions, as well as continuous and discrete wavelet approaches as feature extractors for the purpose of classifying EEG signals obtained in MI tasks. They concluded that PSD are the most robust and efficient methodology, among the tested technique, for extracting patterns that can be used for classification of MI tasks, using classification accuracy as metric.[26]

Intuitively, any stationary time series can be approximated as a superposition of sinusoids oscillating at different frequencies. The spectral density is an estimation of how the power of a signal is distributed among these frequencies. At the population level, the power spectral density (PSD) is defined as the Fourier transform of the autocovariance function. Both, the autocovariance function and the spectral density express the same information, but while the first one expresses it in terms of lags, the PSD expresses it in terms of cycles.[27] The periodogram is the sample-based counterpart of the power spectrum, and it is a tool used for the estimation of the PSD.[27]

In this work, we used the Welch's averaged modified periodogram method to estimate the PSD using Hanning-windowed epochs of length 500 ms with an overlap of 250 ms. These power spectral features were computed in the frequency range between 2 Hz and 40 Hz at a resolution of 1 Hz. This procedure yielded 39 power values per electrode, and resulted in a feature space dimensionality of m = 858 (*i.e.* 22 electrodes × 39 frequencies). Therefore, the feature vector is $x = (x_1, \cdots, x_{858})$ with an associated class label $y \in \{\text{rest, right, left, feet, tongue}\}$. These features were computed for each trial and electrode separately for the rest interval [-3, 0)s and for the MI interval [0, 3)s. Features computed in the rest interval were labelled as rest, while features computed in the MI interval were labelled as right, left, feet and tongue according to the movement that a participant was instructed to imagine.

2.4 Feature selection algorithms

This subsection describes the technical details of the feature selection methods evaluated in this study. Given a dataset $\{(x_i, y_i, i = 1, \cdots, N)\}$, where $x_i \in R^m$, $y_i = -1, +1$ (as the methods were evaluated in two-class scenarios), and N is the number of samples, the goal of the feature selection is to obtain a low-dimensional representation of the original dataset, but with high discriminative power.

2.4.1 *Square of pearson's correlation (R^2)*

The most common approach for feature selection in the recognition of MI tasks from EEG signals is to compute the R^2 value for every feature with respect to the class label, and then select the p features with higher value.[2] This method estimates the discriminative power of every feature independently by computing the square value of the Pearson's correlation coefficient between the values of the jth feature and the class vectors,

$$r_j = \frac{\sum_{i=1}^{N}(x_{ji} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_{ji} - \bar{x}_j)^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (2)$$

where x_{ji} is the ith sample of jth feature, y_i is the class label associated with the ith sample, the bar notation represents the average value across all samples, and $R_j^2 = r_j^2$.[28] In this method, the number of selected features p is heuristically determined. We selected the $p = 15$ features with the highest R_j^2 values. As the R^2 value is computed independently for each feature, the method does not account for correlations in the feature space;[28] however, it is very likely that these correlations occur when using the power spectra of the EEG, especially between neighboring channels and frequencies.

2.4.2 *Fast correlation based filter*

FCBF is a multivariate, nonlinear correlation measure used for automatic feature selection that is based on information theory. This measure assesses simultaneously correlations between the features and the class labels, as well as correlations between features. It is based on symmetrical uncertainty (SU):

$$SU(X,Y) = 2\left[\frac{H(X) - H(X \mid Y)}{H(X)H(Y)}\right] \quad (3)$$

$$H(X) = -\sum_i P(x_i) log_2[P(x_i)] \quad (4)$$

$$H(X \mid Y) = -\sum_j P(y_i) \sum_i P(x_i \mid y_j) log_2[P(x_i \mid y_i)] \quad (5)$$

where $SU(X,Y)$ is the SU between the random variables X and Y, $H(X)$ is the entropy of X and $H(X \mid Y)$ is the conditional entropy of X given Y.[29] However, SU assumes that the random variables X and Y are categorical, which is not the case for features based on the power spectral of the EEG. This limitation was solved by applying the multi-interval discretization method.[30] Automatic selection of features using FCBF involves two steps. In the first step, a subset of relevant features (*i.e.*, features that are correlated to the class) is selected by choosing all features for which $SU(X,Y) > \delta$, where δ is an heuristically chosen threshold that determines the minimum required relevance to be analyzed for redundancy (in this study we choose $\delta = 0$, since choosing higher values diminished the performance in the classification stage). In the second step, a redundant feature X is eliminated if there is a Markov blanket for X within a set of features F (*i.e.*, X does not provide any new information that is not already provided by the set of features F). Searching for an optimal subset of features in this way is combinatorial in nature, and it is prohibitive with a large number of features. An alternative, is to find this subset of features using approximate Markov blankets, which are defined as follows: feature X is an approximate Markov blanket for feature Y, with respect to Z, if $SU(X,Z) \geq SU(Y,Z)$ and $SU(X,Y) \geq SU(Y,Z)$.[31] In contrast to the R^2 and other methods, FCBF automatically selects the number of relevant and non-redundant features.

2.4.3 *Principal component analysis*

Principal component analysis (PCA) is a technique commonly used for dimensionality reduction in the task of recognizing MI from EEG signals.[32,33] This technique projects a dataset into different components that captures the variance of the data, where each component is a linear combination of the features. Moreover, the first component captures the most variance in the dataset, the second component captures the second most variance, and so on.[34,35] Thus, PCA solves the following optimization problem:

$$min_{W,Z} J(W,Z) = \left\| X - ZW^T \right\|_F^2 \quad (6)$$

subject to $WW^T = I_m$, where $X \in R^{N,m}$ is the dataset with zero mean along the columns, $W \in R^{m,m}$ is a matrix of orthogonal vectors, I_m is the m × m identity matrix $Z \in R^{N,m}$ is the projection of the data over the orthogonal vectors, and $\|A\|_F = \sqrt{tr(A^T A)}$ is the Frobenius norm of the matrix A. By selecting the first $p < m$ components in W (which explains most of the variance in the original dataset), we can reduce the dimensionality to p. Thus, the dimension reduced dataset is computed as $\hat{X} = \hat{Z}W^T$, where $\hat{W} \in R^{m,p}$ is the matrix with these p orthogonal components, and $\hat{Z} \in R^{N,p}$ is a low-dimensional representation of the original data.[36] In this study, we used the number of components required to retain 90% of the variance of the original dataset. Unlike the R^2 and FCBF methods, PCA does not discard any features from the original dataset, instead, all of them are transformed into a new lower dimensional space whose variables are linearly uncorrelated.

2.4.4 *Kernel-principal component analysis*

While PCA creates a dimension reduced feature space by taking the linear combination of features, kPCA computes the principal components in a high dimensionality feature space. We can view kPCA as applying a nonlinear transformation to the m-dimensional features x, $\phi(x) \in R^q$, to obtain a higher dimensional representation $q \gg m$.[37,38] However, this might be computationally expensive, especially if the number of dimensions in the original data is high. We therefore use the so-called kernel function to obtain the kernel matrix of the data $K(x_i, x_j)$, $i, j = 1, \cdots, N$. In kPCA the orthogonal principal components vectors of the kernel matrix can be used as the projections of the data onto the respective principal components.[38] Thus, the dimensionality reduced dataset is the first $p < m$ eigenvectors of the kernel matrix. In this study, we used the radial basis function kernel, $K[x_i, x_j = exp(\frac{-\|x_i - x_j\|^2}{2\sigma^2})]$ with $\sigma = 10^{-6}$ (we tested different values for this parameter in the training data, and found that this value provided the best results).

2.5 Learning algorithms

This subsection describes the technical details of the learning methods assessed in this study. As the methods were evaluated in several two-class classification scenarios, the technical details presented here focus on binary classification problems.

2.5.1 *Linear discriminant analysis*

The goal of LDA is to find a linear decision boundary that can separate data from two different classes. The discriminant function takes the form $f(x) = w^T x$. Here, $x = (1, x_1, x_2, \cdots, x_n)^T$ represents the $(n+1)$-dimensional feature vector of an instance to classify, and w is a learned vector of weights. Given a training dataset, the discriminant vector w^* is obtained by seeking the projection that maximizes the distance between the mean of the classes while minimizes their variance, thus providing a classification that is optimal when the two classes are Gaussian with equal covariances.[39] This requires solving the optimization problem:

$$w^* = \text{arg max}_w \text{J(w)} = \frac{w^T S_B w}{w^T S_W w} \tag{7}$$

where S_B is the between-class covariance, and S_W is the within class covariance.[34] This classifier is very simple and demands low computational requirements; however, its performance might be deficient for non-linearly separable datasets.

2.5.2 *SVML and SVMR*

The goal of this classifier is to compute a separating hyperplane that discriminates between classes in a way that maximizes the separating margins between it and the nearest data points in each class (called support vectors).[40] The hyperplane is found by solving the following optimization problem:

$$min_{w,b,\zeta} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \zeta_i \tag{8}$$

subject to $y_i[w^T \phi(x_i) + b] \geq 1 - \zeta_i, \zeta_i \geq 0, \forall i = 1, ..., m$, where $\phi(\cdot)$ maps x into a higher dimensional space, w is the weight vector, b is the bias term, ζ_i are slack variables introduced because a separating hyper-plane might not exist when there is overlap between classes, and C is a regularization parameter that determines the trade off between margin width and training error.[41] When $\phi(x) = x$, the separating hyperplane is linear, leading to the linear SVML. It is also possible to create non-linear decision boundaries through a kernel function given by $K(x_i, y_i) = \phi(x_i)\phi(x_j)$.[42] The most common kernel is the radial basis function (RBF), which is the same as the one used with kPCA. This leads to a Support Vector Machine with RBF kernel (SVMR), which has been extensively used in the recognition of MI tasks from EEG signals.[43] In our experiments, we considered such SVMR, and selected the best values for parameters C and σ using the training data.

2.5.3 *Feed forward neural network (NN)*

Feed Forward NN are general function approximators that can be used as classifiers. The commonly known multi-layer perceptron consists of an array of inputs, an intermediate (aka hidden) layer of neurons with nonlinear activation functions $f(\cdot)$, and an output layer of neurons with linear or nonlinear activation functions.[44] This architecture allows us to create non-linear classification boundaries. The output of the jth neuron is the result of applying an activation function to the linear combination of the weights of the connection between neurons and the neuron's inputs:

$$f(\sum_{i=1}^{n} x_i w_{ij} + \theta_j) \tag{9}$$

where x_i denotes the ith input value, $w_{i,j}$ denotes the synaptic strength between the ith neuron and the jth neuron, θ_j is a bias and n is the number of neurons of the previous layer that are connected with this jth neuron. In our experiments, the NN was implemented using a sigmoid activation function, a single hidden layer with (on different runs) 10, 20, 30 or 40 neurons (the number of neurons to use is determined during the training phase), and a single output neuron. The number of input neurons is the number of features in the dataset.

2.6 Evaluation process

The $4 \times 4 = 16$ combinations of feature selection and classification methods were evaluated in three two-class classification scenarios: 1) rest *vs.* movement $\in$ {left, right, feet, tongue}; 2) upper $\in$ {left, right} *vs.* lower $\in$ {feet} limb movement; 3) right *vs.* left hand movement. To assess performance, we used the data of each subject and session as an individual dataset. We used eighteen datasets in the study, and divided it in two stages: Selection and evaluation. In the selection phase, fourteen sets were randomly selected and used to learn the best combination of feature selector and classification methods. In the evaluation phase, the four remaining sets were used to evaluate the quality of the selected feature selector and classifier that presented the highest performance in the selection phase. For each set in the selection phase, the following paradigm was applied:

(1) Randomly separate 80% of the trials as training set, D_T, and use the remaining 20% as validation set, D_V.
(2) Use 5-fold cross validation on D_T to select the parameters of the classifier (if needed). The feature selection algorithm was embedded in this cross validation process, *i.e.* the feature selection algorithms were executed in each fold.
(3) Train the classifier using D_T and the selected parameters. Unlike the previous step, we do not use cross validation here.
(4) Apply the classification models to D_V and compute the performance metric, defined as the percentage of correctly classified labels, aka classification accuracy (ACC).

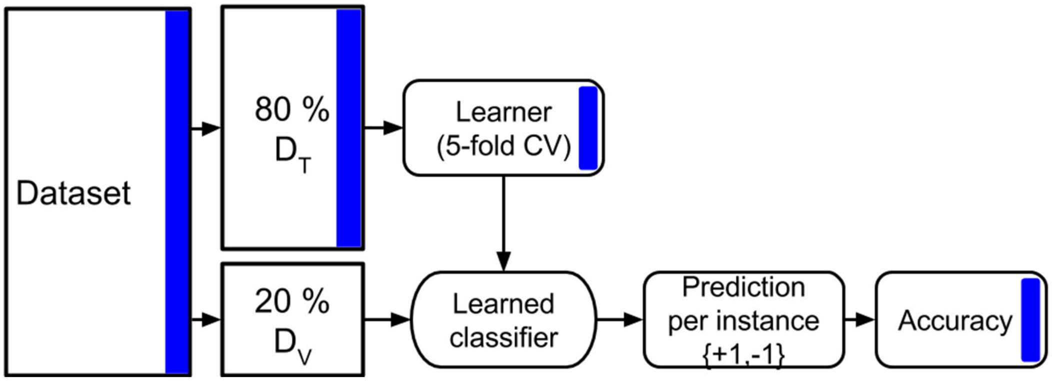

Figure 2. Graphical representation of the process carried out to assess the performance of the methods. The training set, D_T (80% of the data), is used to train the classifier, while the validation set, D_V (20% of the data), is used to assess performance. The parameters of the classifier, as well as the relevant features, are estimated through an inner 5-fold cross validation. Note that not all the stages have access to the true labels (represented by a blue bar). In particular, the learner uses both, the data and the labels, but the learned classifier only have access to the data on D_V. The quality of the classifier is estimated by comparing the predictions on D_V with their corresponding true labels.

Figure 2 depicts the evaluation process followed in the work.

This paradigm was applied 30 times, which differed based on different D_T and D_V splits.[45] Then, the distributions of ACC were constructed by gathering the values ACC of all datasets in the selection phase, and computing the Mean ± STD. across all of them. Finally, we repeated the same paradigm to the four remaining sets in the evaluation phase. The only difference is that in the third step, instead of using all the combinations of learning algorithms and feature selectors, we used only the one selected in the training stage.

The significant chance level of the classification accuracy (ACC_{sig}) was computed with the cumulative binomial distribution,[46] using the lowest number of trials across all datasets (215, 49 and 53 trials for rest *vs.* movement, upper *vs.* lower and right *vs.* left classification scenario, respectively), the number of classes $\mathrm{N}_{classes}$ = 2 and a confidence level of α = 0.05. Therefore, ACC_{sig} for the three classification scenarios were 56%, 61% and 60%. Significant differences between the median of the distribution of ACC and the significant chance level ACC_{sig} were examined using the Wilcoxon signed rank test at the confidence level of α = 0.05.

3. Results

3.1 ERDS maps

Figure 3 shows the scalp topography map of significant ERDS activity (relative to the baseline from -2 to 0 s) obtained using all the available data of one session in a representative subject (*i.e.*, all trials in one out of the 18 datasets). These results show significant desynchronization (P < .05) in all the electrodes during the MI interval [0, 3) s in the motor-related α (8 Hz-12 Hz) and β (12 Hz-30 Hz) frequency bands; however, no significant desynchronization or synchronization (P > .05) is observed during the rest interval [-3, 0)s. This ERDS analysis was also carried out for each type of MI movement independently. The results also showed significant desynchronization (P < .05) during the MI interval but in different scalp locations according to the type of movement (results not presented here). For instance, the left hand MI revealed more prominent desynchronization in electrodes above the contralateral right motor cortex; the right hand MI revealed more prominent desynchronization in electrodes located on the contralateral left motor cortex; while both, feet and tongue MI revealed more prominent desynchronization in electrodes above the midline. In summary, this analysis shows significant power-spectral changes associated with the MI task.

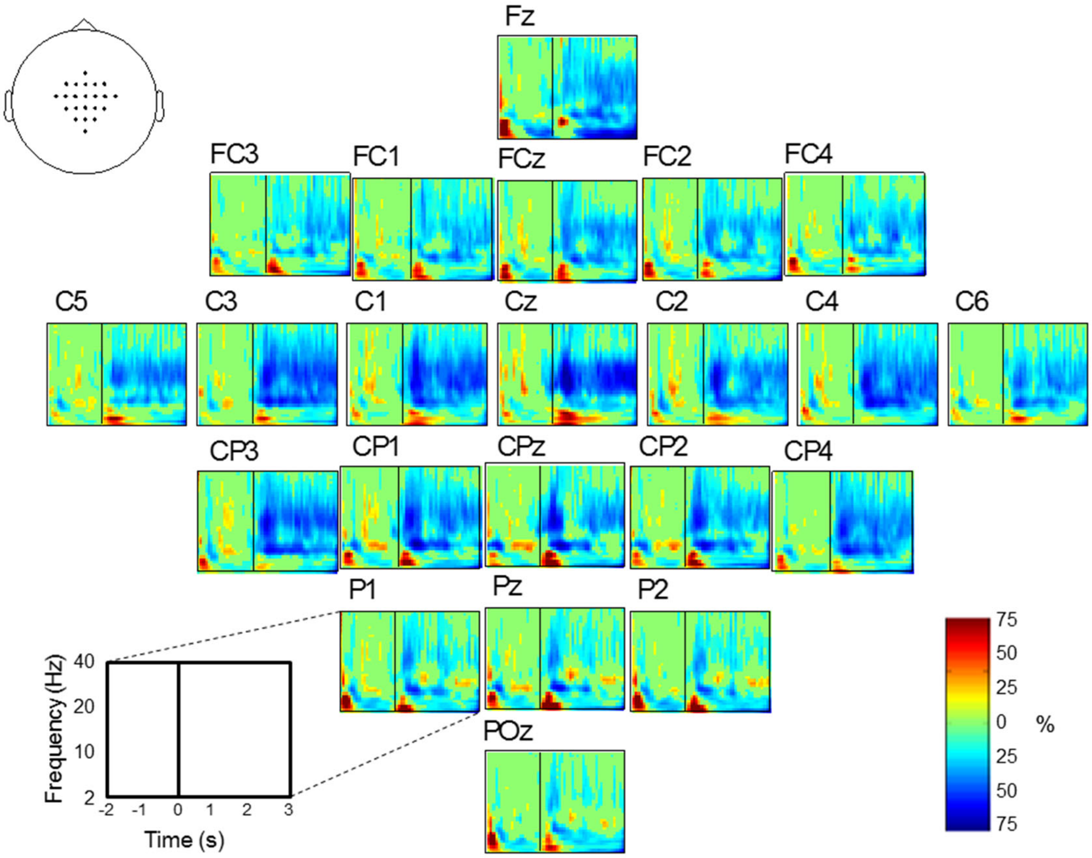

Figure 3. Scalp topographical map of ERDS in one session of a representative subject. Frequency in ordinate from 2 Hz to 40 Hz at a resolution of 1Hz. Time in abscissa from -2 to 3 s. MI onset occurs at t = 0 s (solid black line in all graphs). The colorbar represent the increase (red) or decrease (blue) of the power spectral density with respect to the baseline. It is possible to appreciate that after the imagination of the movement starts (t = 0) there is a decrease in the power spectral density in the alpha and beta bands in the electrodes corresponding to the motor area.

3.2 Feature selection results

Table 1 summarizes the number of selected features across all datasets. The selected number of features differs in all the methods. Note that PCA and kPCA perform a data transformation, thus technically these methods requires all available features, however the number of components used for classification is lower than the total number of features. FCBF is the method with highest variability in the number of selected features and, in some cases, it did not find any useful feature.

Table 1. Summary (median, minimum and maximum) of selected features across all datasets for the four feature selection methods in the three classification scenarios. The minimum and maximum values are shown in parenthesis.

	Rest *vs.* movement	**Right *vs.* left**	**Upper *vs.* lower**
R^2	15 (15 - 15)	15 (15 - 15)	15 (15 - 15)
FCBF	20 (4 - 307)	3 (0 - 21)	10 (0 - 199)
PCA	28 (8 - 53)	21 (6 - 39)	22 (7 - 45)
kPCA	2 (16 - 31)	22 (11 - 31)	23 (11 - 31)

3.3 Classification accuracy results: Selection phase

Figure 4 displays, separately for the three classification scenarios, the distribution of ACC for all combinations of feature selection and classification methods. Note that in the three classification scenarios, the distributions of ACC are absent for feature selector FCBF in combination with classifiers SVML and NN. This was because, during the training process, the learning algorithms were not able to create a separation surface.

Table 2. The Mean ± STD. values of the ACC metric achieved for all combinations of feature selection and classification methods in classification scenario rest *vs.* movement.

	LDA	SVML	SVMR	NN
R^2	66.70% ± 6.18%	67.52% ± 6.59%	75.33% ± 4.10%	71.05% ± 7.04%
FCBF	72.56% ± 6.39%		79.18% ± 5.50%	
PCA	71.50% ± 6.85%	71.82% ± 7.31%	**83.92% ± 5.88%**	70.37% ± 7.08%
kPCA	77.48% ± 5.83%	75.80% ± 6.75%	**83.41% ± 4.19%**	74.67% ± 6.21%

For the rest *vs.* movement classification scenario (see Figure 4a), all combinations of feature selection and classifier presented a distribution of ACC with median significantly different (P < .05, Wilcoxon signed rank test) and greater than ACC$_{sig}$. In this scenario, FCBF-SVMR, PCA-SVMR and kPCA-SVMR presented the highest averaged performance with Mean ± STD. of 79.18 ± 5.50, 83.41 ± 4.19 and 83.92 ± 5.88%, respectively (see Table 2).

Table 3. The Mean ± STD. values of the ACC metric achieved for all combinations of feature selection and classification methods in classification scenario upper *vs.* lower limb movement.

	LDA	SVML	SVMR	NN
R^2	68.28% ± 6.83%	69.60% ± 5.61%	69.29% ± 5.67%	65.92% ± 5.12%
FCBF	**74.38% ± 7.48%**		**74.28% ± 7.70%**	
PCA	73.19% ± 7.20%	72.58% ± 5.98%	72.58% ± 6.23%	70.15% ± 6.27%
kPCA	74.09% ± 7.37%	73.36% ± 6.54%	73.41% ± 6.41%	70.54% ± 6.97%

For the upper *vs.* lower limb movement classification scenario (see Figure 4b), all combinations of feature selection and classifier also presented a distribution of ACC with median significantly greater than ACC$_{sig}$ (P < .05, Wilcoxon signed rank test). In this case, the highest averaged performance was achieved by FCBF-LDA and FCBF-SVMR with mean ± std values of 74.38 ± 7.48 and 74.28 ± 7.70% (see Table 3). It is important to mention that FCBF could not find any relevant feature in 2 out of 14 datasets, so we could not perform classification using this method on these 2 cases. The reasons for this behavior will be explained in section 4. If we remove these 2 datasets from the analysis, the performance of the other combinations of feature selection and classifier slightly increases (compared to the results achieved when they were included).

Table 4. The Mean ± STD. values of the ACC metric achieved for all combinations of feature selection and classification methods in classification scenario right *vs.* left limb movement.

	LDA	SVML	SVMR	NN
R^2	57.68% ± 10.81%	57.55% ± 10.92%	52.53% ± 16.25%	58.55% ± 13.23%
FCBF	**67.46% ± 12.10%**		65.79% ± 12.99%	
PCA	60.56% ± 12.90%	60.58% ± 12.76%	60.23% ± 13.00%	58.42% ± 10.73%
kPCA	66.40% ± 13.89%	65.73% ± 13.80%	65.95% ± 13.36%	62.63% ± 12.36%

Finally, for the right *vs.* left hand movement classification scenario (see Figure 4c), combinations of feature selector FCBF with classifiers LDA and SVMR, and combinations of feature selector kPCA with classifiers LDA, SVML and SVMR, presented a distribution of ACC with median significantly greater than ACC$_{sig}$ (P < .05, Wilcoxon signed rank test). In this scenario, the highest averaged performance was achieved by features selectors FCFB or kPCA in combination with any of the classifiers LDA, SVML and SVMR with a Mean ± STD. around of 66.00 ± 15.00% (see Table 4). As in the previous scenario, FCBF could not find any relevant feature in 6 out of 14 datasets. When the ACC results of these datasets were removed from the other combinations of feature selection and classifier, their performance also slightly increases, however the median of the distributions were not significantly different nor greater than ACC$_{sig}$.

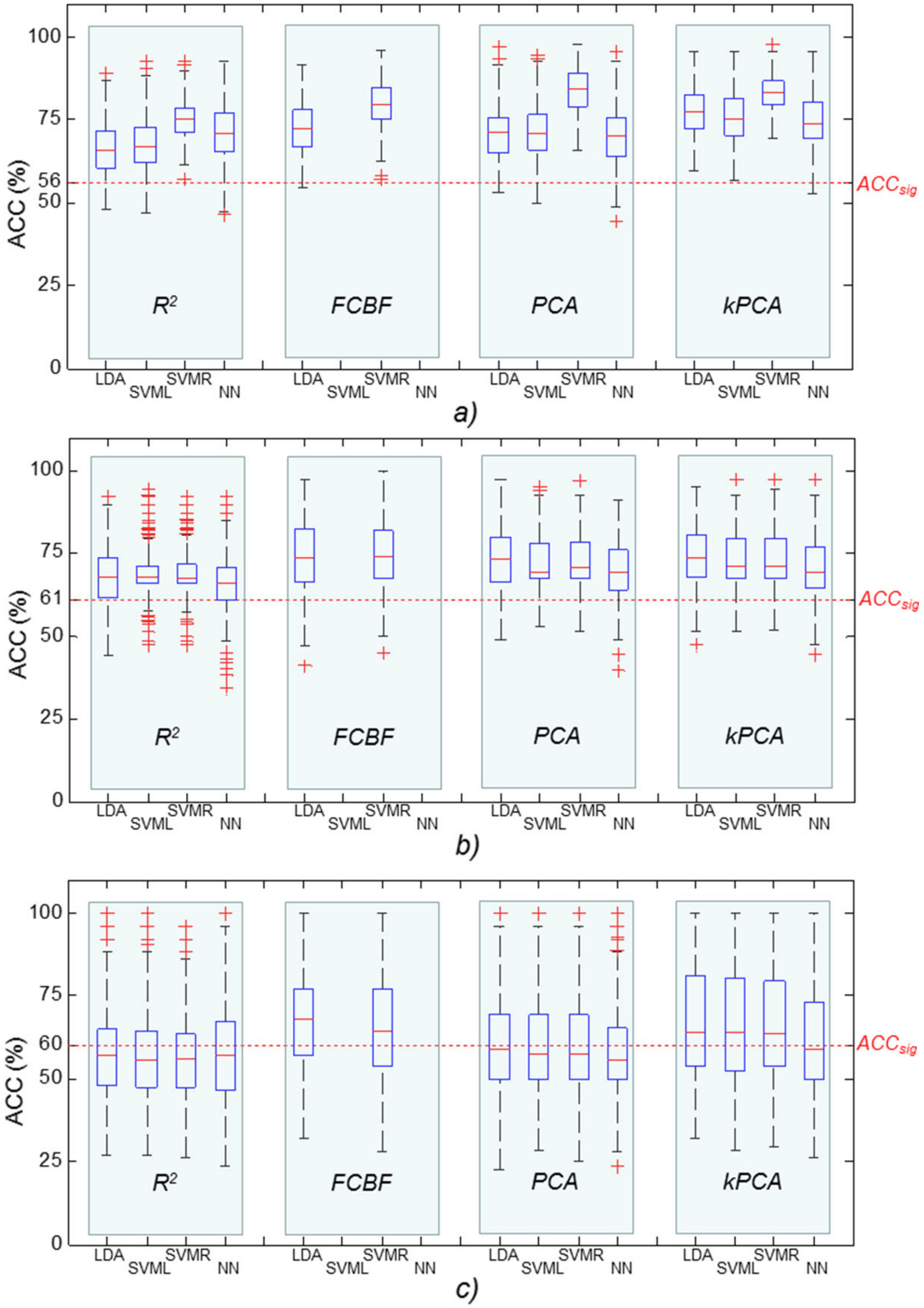

Figure 4. Distributions of the ACC metric for all combinations of feature selection and classification methods achieved in the selection phase for three classification scenarios: a) rest *vs.* movement; b) upper *vs.* lower limb movement; c) right *vs.* left hand movement. The horizontal dotted red line in each classification scenario represents the significant chance level ACC$_{sig}$ ($P < .05$). In order to be considered a significant result, the median of the accuracy distribution should be above the dotted line. It is easy to appreciate that while in scenarios 1 and 2 all the combination of feature selection and classifier did better than chance, in the third scenario only the non linear methods had a performance better than chance.

These results show that the best performance in the first classification scenario was provided by the feature selection PCA or kPCA in combination with classifier SVMR. For the second and third scenarios the best feature selector was FCBF combined with either LDA, or SVMR, with no statistically significant difference in their results. Because of the ability of FCBF to detect problematic datasets we selected FCBF as feature selector, while for classification we selected SVMR because it obtained the highest results across the in the first scenario (for the second and third scenario LDA was slightly better, but the difference was not statistically significant). This combination will be used for evaluation in the subsequent sub-section.

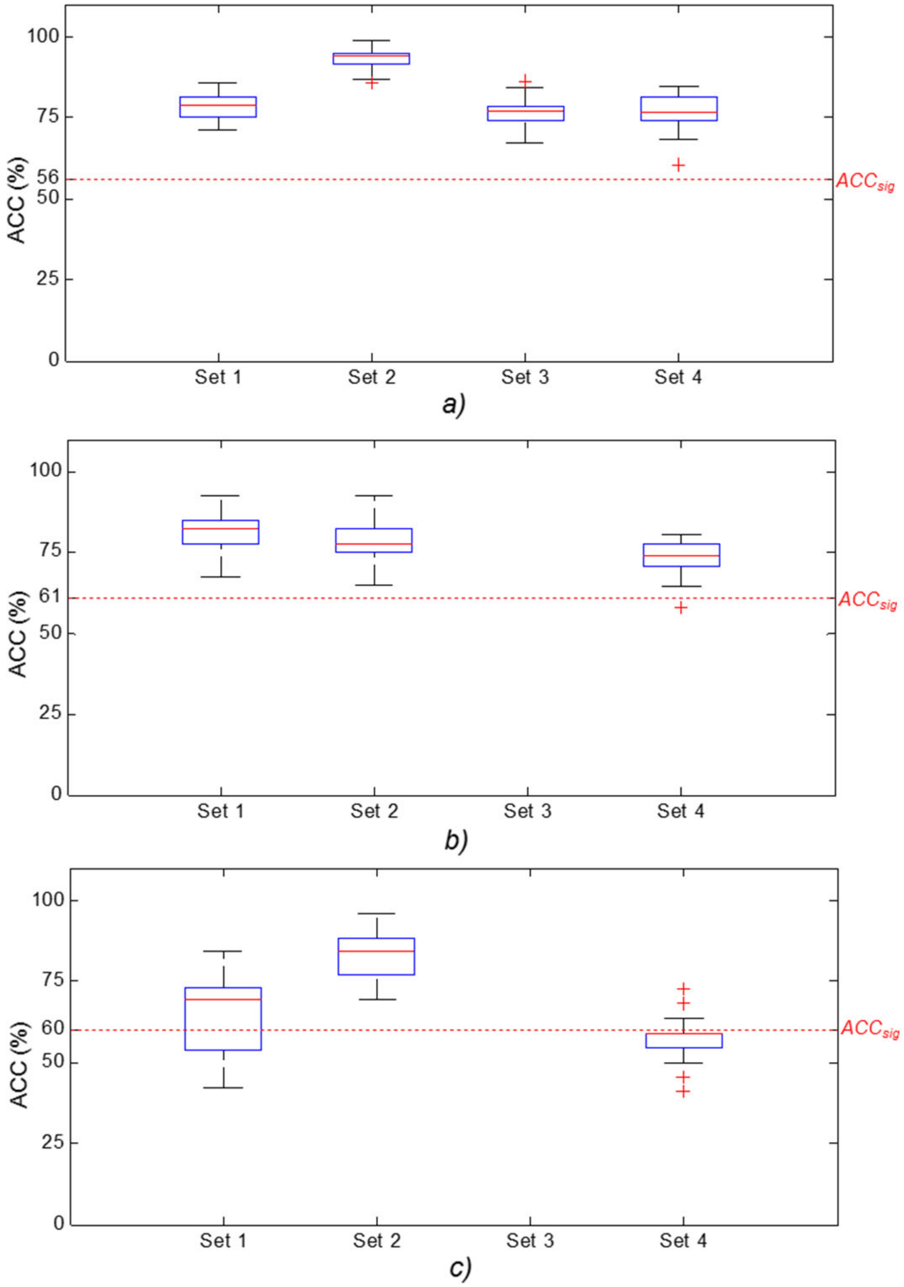

Figure 5. Distributions of the ACC metric obtained by combination FCBF-SVMR in the four datasets used for evaluation for three classification scenarios: a) rest *vs.* movement; b) upper *vs.* lower limb movement; c) right *vs.* left hand movement. The horizontal dotted red line in each classification scenario represents the significant chance level ACC$_{sig}$ (P < .05). In order to be considered a significant result, the median of the accuracy distribution should be above the dotted line. In almost all the cases we obtained results above the chance level. In two cases FCBF considered that there were not enough information in the EEG signals to make a proper classification (Set 3 in the second and third scenario).

3.4 Classification accuracy results: Evaluation phase

Figure 5 shows the distribution of ACC achieved by feature selection FCBF in combination with classifier SVMR in the four datasets used for evaluation. These results are presented separately for the three classification scenarios. Note that one distribution of ACC is absent in classification scenarios upper *vs.* lower limb movement and right *vs.* left hand movement, as FCBF could not find any relevant feature and thus classification was not performed. These results shows that all but one dataset (the fourth in the right *vs.* left movement classification scenario) presented a distribution of ACC with median statistically significantly greater than the significant chance level of accuracy ACC_{sig} ($P < .05$, Wilcoxon signed rank test). The averaged values of Mean ± STD. of ACC across all four datasets were 81.45 ± 8.26%, 77.23 ± 6.81% and 68.71 ± 13.89% for classification scenarios rest *vs.* movement, upper *vs.* lower limb movement and right *vs.* left hand movement, respectively. These results confirm the significant and high classification accuracy provided by the feature selection FCBF in combination with classifier SVMR, and that classification accuracy is highest for the rest *vs.* movement classification scenario.

4. Discussion

This work studied the performance of four feature selection and four classification methods in the recognition of MI tasks from electroencephalographic brain signals. The ERDS analysis of the EEG data revealed significant cortical activity in electrodes located above the motor cortex and in the alpha and beta frequency bands. This analysis showed the presence of recognizable power spectral changes associated with the MI task. Therefore, the question is how to select both: a subset of power spectral-based features and also a classification method that together best recognize MI tasks.

This work addressed this issue by evaluating combinations of linear or nonlinear feature selection, and various classification methods in three two-class classification scenarios: rest *vs.* movement, upper *vs.* lower limb movement and right *vs.* left hand movement. The selection phase showed that all combinations of feature selection and classification methods presented significant classification accuracy in the recognition of rest *vs.* movement and upper *vs.* lower limb movement; however, the classification accuracy in the recognition of right *vs.* left hand movement was significant only for FCBF with LDA and SVMR, and kPCA with any of LDA, SVML, SVMR.

Regarding the feature selection methods, FCBF and kPCA provided the highest performance across the three classification scenarios. FCBF automatically selects a subset of features that are correlated to the class and are not redundant. This means it can select power spectral features extracted from the different EEG channels without user intervention. In contrast, kPCA is required to use all feature as each of its features is a linear combination of all features contained in the original dataset. Therefore, FCBF is preferred over kPCA as FCBF effectively uses only a small subset of the features contained in the original dataset; this allows us to use only a subset of the electrodes, *i.e.*, only the ones whose power spectral features have some discriminative power. This is important for real applications of EEG-based BCI as using fewer electrodes reduces the complexity of the initial setup of the EEG recording system.

Note that FCBF did not find any relevant features in some evaluation sets. This problem may be due to a discretization process applied to the data. FCBF assigns a score to each feature via entropy-based measures, and thus it can only be used with discrete data. As the power spectral features are continuous values, we applied a multi-interval discretization, which separates each feature into bins, using a heuristic that tries to minimize the information entropy of the classes in the dataset.[30] In the cases when FCBF could not find any relevant feature, the number of bins that minimized the information entropy was one, *i.e.*, all the samples would be in the same group regardless of the class. If this happens in all the features, then all the samples would have exactly the same input vector, but different labels, meaning no feature would be relevant for classification. Further investigation is required to understand if this (discretization) is the problem. Importantly, when we eliminated the evaluation sets for which FCBF found no relevant features, the performance in the other combinations of feature selector and classifier also increased. This indicates that it was the removed evaluation sets that had low classification accuracy, FCBF automatically identified by finding no features with discriminative for those evaluation sets. This property provides FCBF an advantage over other feature selection algorithms, since it can be useful for identifying problematic datasets. This means the system designer can now know when this happens, allowing them to consider an alternative approach. We recommend, when possible, not using the datasets where FCBF can find no relevant features, as eliminating these datasets increased the average performance of the tested classifiers. However, if removing entire datasets is not feasible, we would then recommended using kPCA, which is a non-linear combinations of all the features in the dataset, and whose results were the closest to FCBF.

Regarding the classification methods, both LDA and SVMR achieved the highest performance in the three classification scenarios. Indeed, no significant differences were found between the classification accuracy of these classifiers. Given that brain signals are non-stationary, a low-variance classifiers such as LDA might outperform more complex classifiers with lower bias and higher variance, such as SVM.[43] However, we chose SVMR because it presented slightly higher average classification accuracy, and because it is the most common technique used for the recognition of MI tasks from EEG signals. In summary, the best performance in the three classification scenarios was provided by the feature selection FCBF in combination with classifier SVMR.

We evaluated the FCBF in combination with classifier SVMR over the three classification scenarios. This evaluation phase showed significant classification accuracy in the three classification scenarios. Indeed, the average classification accuracy was 81.45 ± 8.26%, 77.23 ± 6.81% and 68.71 ± 13.89% for rest *vs.* movement, upper *vs.* lower limb movement and right *vs.* left hand movement, respectively. Note that classifying between rest and movement yielded the higher classification results, while classifying between right *vs.* left hand movement yielded the lowest classification results. We attribute this difference in performance across the three classification scenarios to two factors: 1) the difference in the power spectral density of the brain signals are stronger between imagining a movement *vs.* imaging no movement, than between different movements, and 2) we had more samples for the rest *vs.* movement than for upper *vs.* lower limb movement and right *vs.* left hand movement, and as expected, lower number of samples leads to poorer performance.

Note that it is difficult to compare our results with previous results,[20,47] as our evaluation process – with the three two-class classification scenarios and the metrics (see subsection 2.6) – differ from their methodology, which focused on training the classifier using data from session one and then testing the resulting system on data from session two, in a four-class classification problem.

5. Conclusion

This paper has characterized the performance of different feature selection and classification techniques in the context of MI using EEG signals. The proper combination of both techniques have the potential of improving the development of brain computer interfaces that aim, among other applications, to restore the functionality of impaired limbs, control external devices, or interact with multimedia applications. Much of the existing literature compares a single feature selection algorithm with several learning algorithms, or several feature selection techniques with a single classifier. The few works that make a broader comparison do not include, to the best of our knowledge, the most common learning algorithms, such as SVM or LDA, nor non-linear algorithms for feature selection, such as kPCA and FCBF. This paper aimed to fill this gap. We first analyzed the event-related desynchronization to verify the presence of recognizable power spectral changes associated with the MI task. After that, we used the power spectral density over 39 frequencies and 22 electrodes as candidate features in three binary classification problems. Finally, we compared the performance of all pairs of four feature selection methods with four learning algorithms. The combination of Support Vector Machine with Radial Basis Kernel and the Fast Correlation Based Filter produced the best results. As far as we know, this is the first time that FCBF is used in this context. This combination, besides achieving high accuracy, had the ability to identify datasets where classification was problematic. When these datasets were removed, the accuracy of the classifier increased by 6%. At the same time, the FCBF automatically selected the number of most relevant, non-redundant features that are required to perform classification. It is important to note that FCBF did not work as well when combined with a SVM that uses a linear kernel, nor with NN (using 10, 20, 30, or 40 neurons in the hidden layer), emphasizing the need of comparing not only several feature selectors, but also several learners.These results also suggest that FCBF might be strong candidate for being used in real-time classification of EEG signals. The behavior of this feature selector under this scenario is part of the future work.

Acknowledgements

The authors want to thank the Consejo Nacional de Ciencia y Tecnología (CONACYT), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Innovates Center for Machine Learning (AICML) for the financial support provided for conducting this research.

Conflicts of Interest Disclosure

The authors declare that they have no conflict of interest.

References

[1] Allison B, Dunne SM, Leeb R, et al. Towards practical brain-computer interfaces: Bridging the gap from research to real-world applications. Biological and Medical Physics. Biomedical Engineering. Springer Verlag. Heidelberg; 2012.

[2] Wolpaw J, Birbaumer N, McFarland D, et al. Brain-computer interfaces for communication and control. Clinical Neurophysiology. 2002; 113 (6): 767-91. http://dx.doi.org/10.1016/S1388-2

457(02)00057-3

[3] Da Silva FL, Niedermeyer E. Electroencephalography: Basic principles, clinical applications, and related fields. 3rd edition. Lippincott Williams & Wilkins; 1993.

[4] Becedas J. Brain-machine interfaces: Basis and advances. IEEE Transactions on Systems, Man, and Cybernetics, part C: Applications and Reviews. 2012; 42 (6): 825-35. http://dx.doi.org/10.1109/TSMCC.2012.2203301

[5] Lebedev MA, Nicolelis MA. Brain-machine interfaces: Past, present and future. Trends in Neurosciences. 2006; 29 (9): 536-46. PMid:16859758. http://dx.doi.org/10.1016/j.tins.2006.07.004

[6] Pfurtscheller G, Neuper C, Flotzinger D, et al. EEG-based discrimination between imagination of right and left hand movement. Electroencephalography and Clinical Neurophysiology. 1997; 103 (6): 642-51. http://dx.doi.org/10.1016/S0013-4694(97)00080-1

[7] Farewell LA, Donchin E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and Clinical Neurophysiology. 1998; 70 (6): 510-23. http://dx.doi.org/10.1016/0013-4694(88)90149-6

[8] Middendorf M, Mcmillan G, Calhoun G, et al. Brain-computer interfaces based on the steady-state visual-evoked response. IEEE Transactions on Rehabilitation Engineering. 2000; 8 (2): 211-4. PMid:10896190. http://dx.doi.org/10.1109/86.847819

[9] Wolpaw JR, McFarland DJ, Neat GW, et al. An EEG-based brain-computer interface for cursor control. Electroencephalography and Clinical Neurophysiology. 1991; 78 (3): 252-9. http://dx.doi.org/10.1016/0013-4694(91)90040-B

[10] Blankertz B, Curio G, Müller KR. Classifying single trial EEG: Towards brain computer interfacing. In: Diettrich S, Becker S, Ghahramani Z (Eds.). Advances in Neural Information Processing Systems 14. MIT Press. 2002: 157-64.

[11] Morash V, Bai O, Furlani S, et al. Classifying EEG signals preceding right hand, left hand, tongue and right foot movements and motor imageries. Clinical Neurophysiology. 2008; 119: 2570-8. PMid:18845473. http://dx.doi.org/10.1016/j.clinph.2008.08.013

[12] Schlögl A, Lee F, Bischof H, et al. Characterization of four-class motor imagery EEG data for the BCI-competition 2005. Journal of Neural Engineering. 2005; 2 (4): L14-22. PMid:16317224. http://dx.doi.org/10.1088/1741-2560/2/4/L02

[13] Ge S, Wang R, Yu D. Classification of four-class motor imagery employing single-channel electroencephalography. PLoS ONE. 2014; 9(6). http://dx.doi.org/10.1371/journal.pone.0098019

[14] Garcia-Laencina PJ, Rodriguez-Bermudez G, Roca-Dorda J. Exploring dimensionality reduction of EEG features in motor imagery task classification. Expert Systems with Applications. 2014; 41 (11): 5285-95. http://dx.doi.org/10.1016/j.eswa.2014.02.043

[15] Rodriguez-Bermudez G, Garcia-Laencina PJ, Roca-Dorda J. Efficient automatic selection and combination of EEG features in least squares classifiers for motor imagery brain-computer interfaces. International Journal of Neural Systems. 2013; 23 (4). PMid:23746288. http://dx.doi.org/10.1142/S0129065713500159

[16] Bai O, Lin P, Vorbach S, et al. Exploration of computational methods for classification of movement intention during human voluntary movement from single trial EEG. Clinical Neurophysiology. 2007; 118: 2637-55. PMid:17967559. http://dx.doi.org/10.1016/j.clinph.2007.08.025

[17] Sen B, Peker M, Cavusoglu A, et al. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. Journal of Medical Systems. 2014; 38 (3). PMid:24609509. http://dx.doi.org/10.1007/s10916-014-0018-0

[18] Koprinska I. Feature selection for brain-computer interfaces. In: New Frontiers in Applied Data Mining. Springer. 2010: 106-17. http://dx.doi.org/10.1007/978-3-642-14640-4_8

[19] Brunner C, Leeb R, Müller-Putz G, et al. BCI competition 2008-Graz dataset A.

[20] Wang D, Miao D, Blohm G. Multi-class motor imagery EEG decoding for brain-computer interfaces. Frontiers in Neuroscience. 2012; 6: 151. PMid:23087607. http://dx.doi.org/10.3389/fnins.2012.00151

[21] Tallon-Baudry C, Bertrand O, Delpuech C, et al. Oscillatory gamma-band (30-70 Hz) activity induced by a visual search task in humans. Journal of Neuroscience. 1997; 17 (2): 722-34. PMid:8987794.

[22] Yuan H, Perdoni C, He B. Relationship between speed and EEG activity during imagined and executed hand movements. Journal of Neural Engineering. 2012; 7 (2).

[23] Graimann B, Pfurtscheller G. Quantification and visualization of event-related changes in oscillatory brain activity in the time frequency domain. Progress in Brain Research. 2006; (159): 79-97. http://dx.doi.org/10.1016/S0079-6123(06)59006-5

[24] McFarland DJ, Miner L, Vaughan T, et al. Mu and beta rhythm topographies during motor imagery and actual movements. Brain Topography. 2000; 12 (3): 177-86. PMid:10791681. http://dx.doi.org/10.1023/A:1023437823106

[25] Pfurtscheller G, Brunner C, Schlögl A, et al. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. NeuroImage. 2006; 31 (1): 153-9. PMid:16443377. http://dx.doi.org/10.1016/j.neuroimage.2005.12.003

[26] Herman P, Prasad G, McGinnity T, et al. Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2008; 16 (4): 317-26. PMid:18701380. http://dx.doi.org/10.1109/TNSRE.2008.926694

[27] Shumway R, Stoffer D. Time series analysis and its applications. Third edition. Springer; 2011. http://dx.doi.org/10.1007/978-1-4419-7865-3

[28] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003; 3 (7/8): 1157-82.

[29] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research. 2004; 5: 1205-24.

[30] Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. IJCAI. 1993: 1022-9.

[31] Chidlovskii B, Lecerf L. Scalable feature selection for multiclass problems. Machine Learning & Knowledge Discovery in Databases. 2008: 227.

[32] Naeem M, Brunner C, Pfurtscheller G. Dimensionality reduction and channel selection of motor imagery electroencephalographic data. Computational Intelligence & Neuroscience. 2009: 1-8. PMid:19536346. http://dx.doi.org/10.1155/2009/537504

[33] Yu X, Chum P, Sim KB. Analysis of the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system. Optik - International Journal for Light and Electron Optics. 2014; 125: 1498-502. http://dx.doi.org/10.1016/j.ijleo.2013.09.013

[34] Bishop CM. Pattern recognition and machine learning (Information science and statistics). Springer-Verlag. New York, Inc Secaucus, NJ, USA, 2006.

[35] Duda RO, Hart PW, Stork DG. Pattern classification. New York, NY. Wiley; 2001.

[36] Murphy KP. Machine Learning: A probabilistic perspective. The MIT Press; 2012.

[37] Schölkpf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation. 1998; 10 (5): 1299-319. http://dx.doi.org/10.1162/089976698300017467

[38] Schölkpf B, Smola A, Müller KR. Kernel principal component analysis. In: Advances in kernel methods, support vector learning. MIT Press. 1999: 327-52.

[39] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. 2nd edition, Springer; 2009. http://dx.doi.org/10.1007/978-0-387-84858-7

[40] Guyon I, Gunn S, Nikravesh M, et al. Feature extraction: Foundations and applications (Studies in fuzziness and soft computing). Springer-Verlag. New York, Inc. Secaucus, NJ, USA, 2006.

[41] Schölkpf B, Smola A. Learning with kernels: support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning. MIT Press; 2002.

[42] Burges CJ. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998; 2: 121-67. http://dx.doi.org/10.1023/A:1009715923555

[43] Lotte F, Congedo M, Lcuyer A, et al. A review of classification algorithms for EEG-based brain computer interfaces. Journal of Neural Engineering. 2007; 4(2): R1-13. PMid:17409472. http://dx.doi.org/10.1088/1741-2560/4/2/R01

[44] Haykin SS. Neural Networks: A comprehensive foundation. Upper Saddle River. NJ, Prentice Hall; 1999.

[45] Tabachnick BG, Fidell LS. Using multivariate statistics. Pearson/Allyn & Bacon. Boston; 2007.

[46] Combrisson E, Jerbi K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. Journal of Neuroscience Methods. 2015; 250: 126-36. PMid:25596422. http://dx.doi.org/10.1016/j.jneumeth.2015.01.010

[47] Suk HI, Lee SW. A novel bayesian framework for discriminative feature extraction in brain-computer interfaces. IEEE transactions on Pattern Analysis and Machine Intelligence. 2013; 35 (2): 286-99. PMid:22431526. http://dx.doi.org/10.1109/TPAMI.2012.69