

Performance enhancement of mental task classification using EEG signal: a study of multivariate feature selection methods

Akshansh Gupta · R. K. Agrawal · Baljeet Kaur

Published online: 7 September 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract In the recent years, the research community has shown interest in the development of brain–computer interface applications which assist physically challenged people to communicate with their brain electroencephalogram (EEG) signal. Representation of these EEG signals for mental task classification in terms of relevant features is important to achieve higher performance in terms of accuracy and computation time. For feature extraction from the EEG, empirical mode decomposition and wavelet transform are more appropriate as they are suitable for the analysis of non-linear and non-stationary time series signals. However, the size of the feature vector obtained from them is huge and may hinder the performance of mental task classification. To obtain a minimal set of relevant and non-redundant features for classification, six popular multivariate filter methods have been investigated which are based on different criteria: distance measure, causal effect and mutual information. Experimental results demonstrate that the classification accuracy improves while the computation time reduces considerably with the use of each of the six multivariate feature selection methods. Among all the combinations of feature extraction and selection methods that are investigated, the combination of wavelet transform and linear regression performs the best.

Ranking analysis and statistical tests are also performed to validate the empirical results.

Keywords Empirical mode decomposition · Wavelet transform · Bhattacharyya distance · Kullback–Leibler distance · Ratio of scatter matrices · Linear regression · Minimum redundancy and maximum relevance

1 Introduction

Medicine and the computing experts have been working together since the past few years, in various fields, to assist patients in many areas of disease diagnosis and treatment. Physically challenged individuals suffering from locomotor syndrome, amyotrophic lateral sclerosis, head trauma, severe cerebral palsy or multiple disorders, etc., are unable to exhibit their intent and hence are dependent on others for seemingly basic tasks. Their physical condition restricts them from operating any electronic device smoothly and freely. The brain–computer interface (BCI) strives to improve the quality of life of such individuals by assisting, augmenting or repairing human cognitive or motor sensory function (Kauhanen et al. 2006; Pfurtscheller et al. 1998). The BCI is one of the thrust areas which has contributed to the development of techniques for providing solutions for brain-related disease prediction and communication control (Anderson et al. 1998; Babiloni et al. 2000; Keirn and Aunon 1990; Wolpaw et al. 2002). This is possible because the neurons in the brain are an active moderator with the outside world, even if speech and motor movements are compromised (Freeman 1999; Graitmann et al. 2003). Efforts by the researchers have been directed to explore the possibility of the brain activities to be successfully taken over by the BCI.

Communicated by V. Loia.

A. Gupta (✉) · R. K. Agrawal
School of Computer and Systems Sciences, Jawaharlal
Nehru University, New Delhi 110067, India
e-mail: akshanshgupta83@gmail.com

R. K. Agrawal
e-mail: rkajnu@gmail.com

B. Kaur
Department of Computer Science, Hansraj College, University
of Delhi, Delhi 110007, India
e-mail: baljeetkaur26@hotmail.com

To monitor the activities of the brain, techniques such as electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), magneto encephalography (MEG) and positron emission tomography (PET) are proposed (Kauhanen et al. 2006; Kronegg et al. 2007). EEG is the preferred technique due to its low cost and ability to record brain signals with non-invasive measures (Kauhanen et al. 2006). EEG has a high temporal resolution and the users can be trained to generate EEG signals that can be translated into their intent. Wolpaw et al. (2002) grouped BCI systems into five major categories, i.e. slow cortical potential, visual evoked potential, P300 evoked potential, sensorimotor activity and activity of the neural cell. Further, Bashashati et al. (2007) introduced two more categories to it, namely multiple neuro-mechanisms and response to mental tasks. The category, response to mental tasks, does not involve any muscular movement. The EEG patterns for these mental tasks are distinguishable from each other due to their frequency variations and are an effective basis to recognize the users' intent. We have focused on the response to mental task category, as it is one of the key areas that can help people with acute physical disorders to benefit in real time to communicate with their immediate environment.

The success of such BCI systems depends on the classification accuracy of the detection of mental tasks, which in turn requires effective extraction and better representation of EEG features related to the mental tasks. The features considered should be highly discriminative which can distinguish different mental tasks in real time. To obtain such a feature set, feature extraction and/or feature selection techniques are suggested in literature. The auto-regressive (AR) technique and the adaptive auto-regressive (AAR) (Penny et al. 2000; Pfurtscheller et al. 1998) technique are two popular feature extraction methods used to extract features in the BCI systems. However, the primary issue with AR modelling is that the accuracy of the spectral estimate is highly dependent on the selected model order. An insufficient model order tends to blur the spectrum, whereas an overly large order may create artificial peaks in the spectrum. Also, these methods assume linearity, Gaussian behaviour and minimum-phase within EEG (Anderson et al. 1998; Basseville and Benveniste 1983; Freeman 1999; Graimann et al. 2003; Pfurtscheller et al. 1998).

In fact, the frequency spectrum of the EEG signal is observed to vary over time, indicating that the EEG signal is a non-stationary signal. As a consequence, such a feature extraction method should be chosen which can model the non-stationary effect in the signal for better representation. Also there is a high variability of the EEG signals across sessions. Hence, there is a need to determine appropriate features which can adapt to the dynamic nature of the EEG signal (Li 2004). However, we have concentrated on the off-line setting.

The wavelet transform (WT) (Mallat 1989; Daubechies 1990) is an effective technique that can be used, which allows analysis of both time and frequency contents of the signal. However, WT uses some fixed basis independent of the processed signal, which makes it non-adaptive. Another successful method for feature extraction, empirical mode decomposition (EMD) (Huang et al. 1998), represents the non-linear and non-stationary signal in terms of modes, that correspond to the underlying signal. EMD is a data-driven approach that does not use a fixed set of basis functions, but is self-adaptive according to the processed signal. It decomposes a signal into finite, well-defined, low-frequency and high-frequency components known as intrinsic mode functions (IMFs) or modes.

The WT and the EMD are two methods which help in the analysis of both time and frequency content of the obtained EEG signal. The EMD method has been used to extract representative data for BCI (Diez et al. 2009; Kaleem et al. 2010) for mental task classification. EEG signals have been analysed with the WT in the fields of motor imagery and epileptic seizures (Bostanov 2004; Ocak 2009; Hsu and Sun 2009; Cvetkovic et al. 2008), brain disorders (Hazarika et al. 1997), classification of human emotion (Murugappan et al. 2010), and non-motor imagery (Cabrera et al. 2010).

The feature vector obtained to represent the EEG signal of a mental task using a feature extraction technique is too large, whereas the number of available training samples is small. Also for mental task classification, EEG signals are captured from more than one channel. The features constructed from multiple channels are concatenated which results in a high-dimensional feature vector. Hence, it suffers from the curse-of-dimensionality which arises due to availability of small sample size and high-dimensional feature vector (Bellman 1961; Hastie et al. 2009). Consequently, it is essential to further reduce the dimension of the feature vector. In literature, feature selection methods, namely the filter method and the wrapper method (Guyon and Elisseeff 2003; Kohavi and John 1997) are suggested to reduce the dimension of the data by removing the irrelevant and redundant features. The reduced feature vector size provides better generalization of the learning system. In addition, the memory requirement and the computation time also decreases which allows learning in real time. The BCI community has employed two categories of feature selection methods: (i) univariate, a filter approach (Diez et al. 2009; Koprinska 2009; Mosquera et al. 2010; Rodríguez-Bermúdez et al. 2013; Cabrera et al. 2010) and (ii) wrapper approach (Keirn and Aunon 1990; Dias et al. 2010; Lakany and Conway 2007; Bhattacharyya et al. 2014; Corralejo et al. 2011; Garrett et al. 2003; Rejer and Lorenz 2013).

The univariate filter method evaluates the relevance of each feature individually based on the statistical characteristics of data and assigns it a score. Features are ranked based

on their score. It is a simple and fast method as it does not involve any classifier. However, these univariate methods do not consider redundancy among selected features, which may degrade the performance of the learning algorithm. Wrapper methods, on the other hand, are computationally more expensive since a classifier must be trained for each candidate subset to find features better suited to the predetermined learning algorithm. In literature, researchers have observed that many a times, the computationally intensive wrapper methods do not outperform the simple filter methods (Haurly et al. 2011). Moreover, the wrapper approach is computationally feasible only for low-dimensional data.

In literature, the multivariate filter methods (Devijver and Kittler 1982) have been found suitable for high-dimensional data to determine a subset of relevant and non-redundant features. These methods take less computation time in comparison to the wrapper approach. To the best of our knowledge, multivariate filter methods have not been explored for mental task classification. This motivated us to investigate some commonly used multivariate filter methods.

In this paper, we have worked on the mental task classification for six subjects. First, the features from the raw signal are extracted using the WT (EMD) and compactly represented in terms of the statistical parameters. Further, empirical comparison of six popular multivariate filter methods based on different criterion: distance measure, causal effect and mutual information, is carried out. The performance is evaluated in terms of the classification accuracy and the computation time. To evaluate and compare the performance of these methods, experiments are performed on a publicly available EEG data acquired by Keirn and Aunon (1990).¹ Different combinations of the feature extraction and the feature selection methods are ranked. Statistical tests are also carried out to strengthen the experimental observations.

The rest of the paper is structured as follows: related work is included in Sect. 2. Brief discussion on WT and EMD is included in Sect. 3. A brief overview of different multivariate filter methods used in our study is presented in Sect. 4. Experimental data and results with statistical analysis are discussed in Sect. 5. Section 6 includes conclusion and future work.

2 Related work

The univariate (ranking) method, one of the filter methods, and the wrapper methods are commonly used to select relevant features in the BCI systems. Diez et al. (2009) suggested a univariate method based on the Wilks' lambda to determine a relevant set of features. Koprinska (2009) investigated five different univariate filter methods: information gain, correla-

tion, reliefF, consistency and 1r ranking (IRR) on BCI data. The mutual information (Mosquera et al. 2010), is used as a relevance criteria to determine a subset of relevant features. Recently, the research work (Rodríguez-Bermúdez et al. 2013) used feature ranking using the least angle regression and the Wilcoxon rank sum test to select a set of relevant features. It is noted that the reduced number of relevant features obtained, using univariate methods, significantly improves the accuracy.

Among the wrapper methods, the seminal work by Keirn and Aunon (1990) used a combination of forward sequential feature selection and an exhaustive search for a pair of features to obtain a subset of relevant and non-redundant features for the mental task classification. Dias et al. (2010) used a sequential forward selection algorithm that includes features to the subset sequentially for task discrimination in BCI. In the area of non-invasive BCIs, a wrapper method based on SVM is used to guide intention of movement (Lakany and Conway 2007). In recent works in BCI, genetic algorithm (GA) (Bhattacharyya et al. 2014; Corralejo et al. 2011; Rejer and Lorenz 2013) is used to determine a subset of relevant and non-redundant features. GA for BCI based on finger movement experiments (Garrett et al. 2003) is also suggested. It is pointed out that the relevant subset of features selected might not be the same in different runs of GA as GA is a stochastic method. Also, the wrapper methods are computationally expensive as they involve the classifier at each stage of forming the feature subset, which is not encouraged for high-dimension data.

On the other hand, the multivariate filter method, which is not investigated in BCI, overcomes the limitations of both the univariate filter method and the wrapper method.

3 Feature extraction

The features from an EEG signal are extracted in two steps: in the first step, EEG signal is decomposed by EMD/WT and statistical parameters are computed to represent the signal more compactly in the second step. WT, EMD and the statistical measures used are discussed briefly below.

3.1 Wavelet transform (WT)

Wavelet is a multi-resolution mathematical tool used to analyse a signal in both spatial and frequency domain simultaneously with the use of variable-sized windows. Discrete wavelet transform (DWT) decomposes a 1-D signal, $f(x)$, in terms of a shifted and dilated mother wavelet $\varphi(x)$ and scaling function $\phi(x)$, given by Mallat (1989):

$$f(x) = \sum_l s_{j_0,l} \phi_{j_0,l}(x) + \sum_{j=j_0}^{\infty} \sum_l d_{j,l} \varphi_{j,l}(x), \quad (1)$$

¹ http://www.cs.colostate.edu/eeg/main/data/1989_Keirn_and_Aunon.

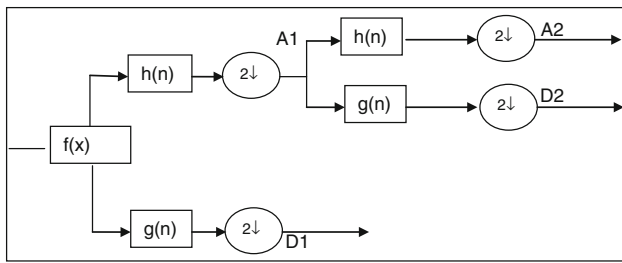


Fig. 1 Decomposition of a signal into approximation and detail components

where j_0 is an arbitrary starting scale, $l = 0, 1, 2, \dots, 2^j - 1$, $s_{j,l}$ and $d_{j,l}$ are scaling and wavelet coefficients, respectively.

The wavelet coefficients $d_{j,l}$ and the scaling coefficients $s_{j,l}$ are computed, respectively, by

$$d_{j,l}(x) = \langle f(x), \varphi_{j,l}(x) \rangle, \quad l = 0, 1, 2, \dots, 2^j - 1 \quad (2)$$

$$s_{j,l}(x) = \langle f(x), \phi_{j,l}(x) \rangle, \quad l = 0, 1, 2, \dots, 2^j - 1 \quad (3)$$

The scaling basis functions and wavelet basis functions at scale j are denoted in terms of scaling basis functions and wavelet basis functions at scale $j + 1$, respectively, as given below:

$$\vartheta(x)_{j,k} = \sum_n h(n)\vartheta(x)_{j+1,n} \quad (4)$$

$$\varphi(x)_{j,k} = \sum_n g(n)\varphi(x)_{j+1,n}, \quad (5)$$

where $g(n)$ is the discrete mother wavelet which is high-pass in nature; $h(n)$ is mirror version of $g(n)$ and is low-pass in nature.

Wavelets provide a simple hierarchical framework for the multiresolution analysis of the signal. The discrete time-domain signal (Fig. 1) is decomposed using DWT by successive low-pass and high-pass filtering. In the figure, the signal is denoted by the sequence $f(x)$. At each level, the high-pass filter produces detail information; while the low-pass filter associated with scaling function produces coarse or approximate information. The filtering and decimation process is continued until the desired level is reached. The maximum number of levels depend upon the length of the signal. The DWT of the original signal is then obtained by concatenating all the coefficients starting from the last level of decomposition.

3.2 Empirical mode decomposition (EMD)

Under the assumption that any signal is composed of a series of different intrinsic oscillation modes, the EMD can be used to decompose an incoming signal into its different intrinsic mode functions (IMF). An IMF is a continuous func-

tion that satisfies the following conditions (Huang et al. 1998):

1. The number of maxima and the number of minima are either equal, or differ at most by one.
2. The mean value of the envelope defined by the local maxima and the local minima is zero.

Given the incoming signal $f(x)$, the algorithm of EMD is based on a sifting process that can be summarized as Huang et al. (1998):

1. For a signal, $f(x)$, identify all local maxima and minima.
2. Calculate the upper envelope by connecting all the local maxima points of the signal using a cubic spline.
3. Repeat the same for the local minima points of the signal to produce the lower envelope.
4. Compute the mean of both envelopes, say m_1 .
5. Update the signal, $f(x) = f(x) - m_1$
6. Repeat the steps 1 to 5, and consider $f(x)$ as the input signal, until $f(x)$ can be considered an IMF as per the definition stated above.
7. The residue r_1 is obtained by subtracting the first IMF (IMF₁) from $f(x)$, i.e. $r_1 = f(x) - \text{IMF}_1$. The residual of this step becomes the signal $f(x)$ for the next iteration.
8. Iterate steps 1 to 7 on the residual r_j ; $j = 1, 2, 3, \dots, n$ in order to obtain all the IMFs of the signal.

The procedure terminates when the residual r_j is either a constant value or a function with a single maxima (minima). The result of the EMD process produces n IMFs and a residue signal r_n . The original signal $f(x)$ can be reconstructed in terms of the n extracted IMFs and the residue as:

$$f(x) = \sum_{j=1}^n \text{IMF}_j + r_n \quad (6)$$

In order to obtain the IMFs of the signal, publicly available EMD toolbox for Matlab[®]2 is utilized. The lower-order IMFs capture the faster oscillation modes of the signal, whereas the higher-order IMFs capture the slower oscillation modes. During the process of decomposing the signal using EMD, it has been found that most of the segments decompose into 4 or more IMFs. Hence, to maintain consistency, we have used four levels of IMF decomposition to form the feature vector of the same size from a given sample. Since in this work, comparison between WT and EMD is done, three levels of decomposition are performed for WT also, to have the same size of the feature vector throughout the experiments.

² <http://perso.enslyon.fr/patrick.flandrin/emd.html>.

3.3 Formation of the feature vector

The following statistical measures or parameters are used to represent the EMD and the wavelet coefficients to form a reduced feature vector. Some of these parameters represent linear characteristics of EEG signal and others represent non-linear properties of EEG (Diez et al. 2009; Gupta and Agrawal 2012): (i) root mean square (RMS); (ii) variance; (iii) Shannon entropy; (iv) Lempel–Ziv complexity measure (LZ); (v) central frequency (50 % of spectrum energy); (vi) maximum frequency (95 % of spectrum energy); (vii) skewness and (viii) Kurtosis.

Root mean square (RMS) and variance represent statistical measure of numerical values of varying quantity and dispersion of the signal, respectively. Entropy measures the average amount of information from a signal (Shannon 1948). Lempel and Ziv (1976) quantifies the complexity of a signal by analysing its spatial–temporal patterns. The central and the maximum frequencies are used as descriptors of the band-width of each component of the EEG signal. Skewness measures the degree of signal asymmetry around the mean and Kurtosis measures the sharpness of the signal.

4 Feature selection

The feature vector from each channel obtained encloses all the features constructed with the above statistical parameters. The final feature vector obtained after concatenation of features from six channels is large, i.e. each feature vector contains 192 parameters (4 IMFs or 4 wavelet filters \times 8 parameters \times 6 channels). Hence, feature selection is carried out to exclude noisy, irrelevant, and redundant features. The feature selection methods can be divided into two broad categories: filter and wrapper category (Guyon and Elisseeff 2003; Kohavi and John 1997). The statistical characteristics of data are employed for feature selection in filter methods. This approach involves less computation overhead. The filter methods independently measure the relevance of features without encompassing any classifier. The selected features obtained with the help of filter approach may not be the most relevant set of features for a learning algorithm. This is because the learning algorithm is not involved during the feature selection phase. Alternatively, the wrapper methods select features that are better suited to a given learning algorithm, resulting in better performance. However, wrapper methods are computationally more costly since the classifier needs to be trained for each candidate subset.

Filter methodologies are further classified into two categories (Guyon and Elisseeff 2003): (i) univariate (ranking) and (ii) multivariate (feature subset). Feature ranking uses scoring function for measuring the relevance of each feature individually for classification. These methods are

quite simple and efficient. But the limitation associated with feature ranking methods is that the correlation among the selected features is not considered. Hence the subset of features may be highly redundant, which may degrade the performance of the learning algorithm. On the other hand, multivariate methods consider that subset of features which are relevant to the class and non-redundant among themselves. Among research community (Bhattacharyya 1943; Kullback and Liebler 1951; Devijver and Kittler 1982; Park et al. 2007; Peng et al. 2005; Groissboeck et al. 2004; Sakar et al. 2012), the most widely used multivariate filter methods are Euclidean distance, Bhattacharyya distance measures Kullback–Leibler distance, ratio of scatter matrices, linear regression and mRMR. These multivariate filter methods overcome the limitations of both the ranking and the wrapper methods. These six multivariate methods consider different criterions such as the distance measure, the cluster criterion, regression and mutual information. Brief discussion of these techniques is given below. The given five mental tasks classification problem is formulated as a set of ten different two-class classification problem (Keirn and Aunon 1990). The following discussion is based on two classes, C_1 and C_2 .

4.1 Euclidean distance (ED)

Euclidean distance is a simple distance measure, based on the Pythagorean Theorem, which computes the distance between the data points of two classes. Euclidean distance between two classes C_1 and C_2 corresponding to the inclusion of K features is given by:

$$J_{K,E} = \sqrt{(\mu_K^1 - \mu_K^2)(\mu_K^1 - \mu_K^2)^T} \quad (7)$$

where μ_K^1 and μ_K^2 are the mean vectors of class C_1 and C_2 , respectively.

It is simple, fast and considers information related to the compactness and overlap between classes. However, it assumes that the data are distributed about the sample mean in a spherical manner. It is not suitable for datasets that have low signal-to-noise ratio and negative spikes. Also, it works only for continuous and quantitative data. Moreover, it does not consider the data spread/variation.

4.2 Bhattacharyya distance (BD)

The class conditional p.d.f., $p(\mathbf{X}_k|C_i)$ of k dimensional sample $\mathbf{X}_k = [x_1, x_2, \dots, x_k]$ for a given class C_i where $i = 1, 2$ is given by:

$$p(\mathbf{X}_k|C_i) = \frac{1}{(2\pi)^{d/2}} \left| \Sigma_k^i \right|^{1/2} \times \exp \left[-\frac{1}{2} (\mathbf{X}_k - \mu_k^i)^T (\Sigma_k^i)^{-1} (\mathbf{X}_k - \mu_k^i) \right], \quad (8)$$

where μ_k^i is the mean vector and Σ_k^i is the covariance matrix for the class C_i . For multivariate normal distribution for two classes, Bhattacharyya distance (BD) measure is given by Bhattacharyya (1943):

$$J_B = \frac{1}{8}(\mu_k^2 - \mu_k^1)^T \left[(\Sigma_k^1 + \Sigma_k^2)/2 \right]^{-1} (\mu_k^2 - \mu_k^1) + \frac{1}{2} \log \frac{|\Sigma_k^1 + \Sigma_k^2|}{(|\Sigma_k^1|^{1/2} |\Sigma_k^2|^{1/2})} \tag{9}$$

However, it suffers from the problem of singularity when covariance becomes very small.

4.3 Kullback divergence (KD)

Kullback divergence is a statistical measure that has a direct relation with the Bayes error. It gives the minimum achievable error for adopting a particular feature. The KD measure for class C_1 and C_2 is given by:

$$J_k = \frac{1}{2}((\mu_k^2 - \mu_k^1)^T ((\Sigma_k^2)^{-1} - (\Sigma_k^1)^{-1})(\mu_k^2 - \mu_k^1) + \frac{1}{2} tr((\Sigma_k^2)^{-1}(\Sigma_k^1) + (\Sigma_k^1)^{-1}(\Sigma_k^2) - 2I), \tag{10}$$

where $\mu_k^1, \mu_k^2, \Sigma_k^1$ and Σ_k^2 are the mean vectors and covariance matrices of class C_1 and C_2 , respectively.

Both the KD and BD measures assume that the data follows the Gaussian distribution. Any deviation from this assumption may not provide better results. They also suffer from the problem of singularity when the covariance between classes becomes very small.

4.4 Ratio of scatter matrices (SR)

In literature, a simple criterion based on the dissemination of features in high-dimensional space is recommended, which is the trace of ratio of scatter matrices. The criterion selects those features that are well clustered around their class mean and the features of the different classes are well separated. To this end, the following scatter matrices are defined.

Within-class scatter, S_w , and between-class scatter matrices, S_b , are, respectively, given by

$$S_w = \frac{1}{N} \sum_{i=1}^c \sum_{x \in i} (x - \mu_i)(x - \mu_i)^T \tag{11}$$

$$S_b = \sum_{i=1}^c (x - \mu_i)(x - \mu_i)^T, \tag{12}$$

where μ_i, P_i and N are the mean vector, prior probability of the i th class data and the total number of data samples, respectively.

From these definitions of the scatter matrices, it is straightforward to observe that the criterion

$$J_{SR} = \frac{\text{trace}(S_b)}{\text{trace}(S_w)} \tag{13}$$

takes large values when samples of the selected feature space are well clustered around their mean, within each class, and the clusters of the different classes are well separated. Also, the criteria J_{SR} has the advantage of being invariant under linear transformation. The main advantage of this criterion is that it is independent of external parameters and assumptions of any probability density function. Also it selects features simultaneously taking care of within-class and between-class spread of data values for features. In a similar approach, leave-one-feature-out (Lughofer 2011), difference between-class and within-class variances are updated in an incremental mode by discarding features stepwise.

4.5 Linear regression (LR)

Regression analysis is another well-established statistical method in literature, which investigates the causal effect of the independent variables upon the dependent variable. The class label is used as the dependent variable (target) and the features that affect this target are sought. The linear regression method attempts to model the relationship between two or more explanatory variables and a response variable, by fitting a linear equation to the observed data. Since many features can affect the class, therefore the multiple regression model is more appropriate. A multiple regression model with a target variable, y and multiple variables, X is given by:

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \zeta_i, \quad i = 1, 2, \dots, n, \tag{14}$$

where $\beta_0, \beta_1, \dots, \beta_k$ are constants estimated by observed values of X and the class label y and is estimated by normal distribution having mean zero and a variance σ^2 .

The sum of squares error (SSE) which is sum of the squared residuals is given by

$$SSE = \sum_{i=1}^n (y_i - y_i^p)^2, \tag{15}$$

where y and y^p are observed and predicted values, respectively. A large value of SSE means that the regression is predicted poorly. The total sum of squares is given by

$$SSTO = \sum_{i=1}^n (y_i - \bar{y})^2, \tag{16}$$

where \bar{y} is the average of y_i . In a regression model, the choice of features which best explains the class label depends on the value of R^2 , given by

$$J_{LR} = R^2 = 1 - \frac{SSE}{SSTO} \tag{17}$$

The value of R^2 lies between 0 and 1. A feature with a large value of R^2 (approaching 1) is considered good for distinguishing the classes. However, it considers a linear relationship between data and class labels, which may not be the case every time. Also, it considers that the data are independent, i.e. there is no correlation in the data.

4.6 Minimum redundancy–maximum relevance (mRMR)

Minimum redundancy–maximum relevance (mRMR), proposed by Peng et al. (2005) is a filter selection approach, based on mutual information, to find a subset of features that have minimum redundancy among themselves and maximum relevance with the class labels. The mRMR method uses mutual information $I(\mathbf{x}_i, \mathbf{x}_j)$ as a measure of similarity between two discrete variables \mathbf{x}_i and \mathbf{x}_j , given by:

$$I(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k,l} p(\mathbf{x}_{ik}, \mathbf{x}_{jl}) \log \left(\frac{p(\mathbf{x}_{ik}, \mathbf{x}_{jl})}{p(\mathbf{x}_{ik})p(\mathbf{x}_{jl})} \right), \quad (18)$$

where $p(\mathbf{x}_{ik})$, $p(\mathbf{x}_{jl})$ are the marginal probabilities of the k th and the l th samples of discrete variables \mathbf{x}_i and \mathbf{x}_j , respectively, and $p(\mathbf{x}_{ik}, \mathbf{x}_{jl})$ is their joint probability density.

The relevance, REL, between the feature \mathbf{x}_i and the set of class labels can be expressed as:

$$\text{REL} = \frac{1}{|S|} \sum_{\mathbf{x}_i \in S} I(\mathbf{x}_i, c), \quad (19)$$

where c denotes the set of target class labels.

The average redundancy among features in the set S , RED, can be expressed as:

$$\text{RED} = \frac{1}{|S|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \in S} I(\mathbf{x}_i, \mathbf{x}_j), \quad (20)$$

where S denotes the subset of features being considered for selection and $|S|$ denotes the number of features in set S .

Minimum redundancy and maximum relevance can be measured as follows:

$$\begin{aligned} J_{\text{MID}} &= \max_S [\text{REL} - \text{RED}] \\ &= \max_S \left[\frac{1}{|S|} \sum_{\mathbf{x}_i \in S} I(\mathbf{x}_i, c) - \frac{1}{|S|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \in S} I(\mathbf{x}_i, \mathbf{x}_j) \right] \end{aligned} \quad (21)$$

Clearly, the maximum values of J_{MID} can be achieved with minimum redundancy and maximum relevance values.

4.7 Sequential forward feature selection

To determine a subset of relevant and non-redundant features using a given multivariate criterion, there are many suboptimal search methods. We have used sequential forward fea-

ture selection search method which is based on the greedy approach. It is simple and fast (the time complexity is $O(d^2)$). The outline of the algorithm to determine a subset of relevant and non-redundant features using a given multivariate criterion is given below:

Step 1. Initialization: $R = \{\}$ // Initial empty set of relevant and non-redundant features

$S =$ The set of given d input features, $\{x_1, x_2, \dots, x_d\}$

Step 2. The single best feature is selected which optimizes a criterion function, $J(\cdot)$

$x_k = \underset{i}{\text{optimum}} J(x_i);$

$R = R \cup \{x_k\}; S = S - \{x_k\};$

Step 3. Sets of features are formed using one of the remaining features from the set S and the already selected set of features, R . Then the best set is selected.

Compute $x_j = \underset{i}{\text{optimum}} J(R \cup \{x_i\});$

$R = R \cup \{x_j\}; S = S - \{x_j\};$

Step 4. Repeat step 3 until a predefined number of features is selected.

5 Experimental setup and results

The EEG data used in our experiment were acquired by Keirn and Aunon (1990). An Electro-cap elastic electrode cap was used to record the data from positions C3, C4, P3, P4, O1, and O2, defined by the 10–20 system of electrode placement. Data were recorded at a sampling rate of 250 Hz. Originally, seven subjects were used for the purpose of collecting data. For our experiment, the data from six subjects leaving out subject 4 (due of unavailability of proper data) performing five different mental tasks were analysed. The mental tasks observed under this category are: visual counting (referred to as C in the paper), imagined geometrical rotation (R), letter composition (L), mathematics involving non-trivial multiplication (M) and a complete relaxed observation (B). Data were recorded for 10 s during each task and each task was repeated five times per session. Most subjects attended two such sessions recorded on separate weeks, resulting in a total of 10 trials for each task. With a 250-Hz sampling rate, each 10-s trial produced 2,500 samples per channel. These are divided into half-second segments, producing at most 20 segments per trial. Features are extracted from the signal using both EMD and WT. We have used the more commonly used

Table 1 Comparison of average classification accuracy among various combinations of feature extraction and selection methods

Task	Classifier	WFS		ED		BD		KD		SR		LR		mrMR	
		Emd	Wavelet	Emd	Wavelet	Emd	Wavelet	Emd	Wavelet	Emd	Wavelet	Emd	Wavelet	Emd	Wavelet
BaseLine count	LDC	0.69	0.72	0.76	0.73	0.85	0.87	0.65	0.66	0.88	0.89	0.86	0.89	0.79	0.78
	QDC	0.55	0.59	0.71	0.69	0.82	0.85	0.6	0.64	0.85	0.86	0.84	0.86	0.76	0.78
	KNN	0.53	0.54	0.71	0.58	0.85	0.85	0.61	0.66	0.83	0.85	0.81	0.85	0.59	0.55
	SVM	0.55	0.54	0.63	0.62	0.92	0.92	0.7	0.73	0.64	0.6	0.82	0.9	0.62	0.6
BaseLine letter	LDC	0.66	0.69	0.73	0.64	0.79	0.84	0.6	0.61	0.84	0.87	0.8	0.87	0.66	0.66
	QDC	0.54	0.59	0.69	0.65	0.75	0.78	0.59	0.61	0.8	0.84	0.78	0.84	0.67	0.67
	KNN	0.52	0.54	0.74	0.57	0.79	0.81	0.64	0.64	0.77	0.82	0.74	0.82	0.58	0.55
	SVM	0.56	0.56	0.64	0.6	0.87	0.89	0.71	0.71	0.65	0.6	0.84	0.91	0.62	0.6
BaseLine math	LDC	0.79	0.78	0.84	0.75	0.87	0.9	0.65	0.67	0.91	0.92	0.91	0.92	0.78	0.78
	QDC	0.58	0.59	0.8	0.76	0.84	0.87	0.65	0.66	0.88	0.89	0.89	0.89	0.77	0.78
	KNN	0.57	0.61	0.8	0.63	0.85	0.85	0.69	0.68	0.86	0.86	0.87	0.86	0.72	0.66
	SVM	0.59	0.62	0.69	0.68	0.9	0.92	0.77	0.8	0.69	0.72	0.85	0.93	0.69	0.72
BaseLine rot	LDC	0.72	0.76	0.87	0.85	0.9	0.91	0.73	0.71	0.92	0.93	0.92	0.93	0.83	0.86
	QDC	0.57	0.67	0.86	0.85	0.87	0.89	0.7	0.71	0.9	0.9	0.9	0.9	0.82	0.85
	KNN	0.53	0.6	0.84	0.63	0.9	0.89	0.72	0.71	0.89	0.89	0.88	0.89	0.77	0.67
	SVM	0.62	0.58	0.69	0.67	0.91	0.92	0.77	0.76	0.69	0.67	0.92	0.94	0.7	0.67
Count letter	LDC	0.69	0.76	0.76	0.65	0.85	0.88	0.68	0.69	0.89	0.9	0.87	0.9	0.71	0.74
	QDC	0.58	0.62	0.71	0.64	0.8	0.83	0.63	0.65	0.85	0.86	0.84	0.86	0.71	0.77
	KNN	0.5	0.53	0.72	0.56	0.83	0.87	0.67	0.69	0.81	0.82	0.81	0.82	0.55	0.54
	SVM	0.53	0.54	0.57	0.56	0.88	0.91	0.76	0.73	0.56	0.58	0.86	0.9	0.59	0.58
Count math	LDC	0.78	0.79	0.83	0.77	0.88	0.9	0.62	0.65	0.91	0.91	0.91	0.91	0.78	0.81
	QDC	0.58	0.6	0.8	0.77	0.85	0.89	0.61	0.65	0.9	0.91	0.9	0.91	0.79	0.81
	KNN	0.54	0.57	0.74	0.63	0.86	0.92	0.65	0.68	0.88	0.89	0.88	0.89	0.73	0.66
	SVM	0.62	0.61	0.69	0.72	0.92	0.96	0.75	0.76	0.71	0.72	0.93	0.92	0.68	0.72
Count rot	LDC	0.7	0.8	0.81	0.79	0.86	0.88	0.66	0.66	0.88	0.91	0.85	0.91	0.76	0.78
	QDC	0.57	0.64	0.81	0.79	0.83	0.88	0.63	0.66	0.87	0.9	0.84	0.9	0.76	0.78
	KNN	0.5	0.56	0.74	0.61	0.85	0.87	0.66	0.66	0.85	0.84	0.84	0.84	0.65	0.6
	SVM	0.57	0.54	0.64	0.63	0.92	0.89	0.75	0.73	0.62	0.63	0.86	0.88	0.64	0.63
Letter math	LDC	0.78	0.8	0.85	0.78	0.91	0.93	0.66	0.7	0.92	0.93	0.88	0.93	0.75	0.79
	QDC	0.59	0.6	0.83	0.76	0.88	0.9	0.67	0.7	0.91	0.93	0.87	0.93	0.76	0.8
	KNN	0.54	0.57	0.78	0.61	0.85	0.9	0.68	0.71	0.88	0.91	0.87	0.91	0.69	0.65
	SVM	0.6	0.6	0.74	0.71	0.92	0.95	0.76	0.79	0.73	0.71	0.92	0.95	0.67	0.71
Letter ro	LDC	0.76	0.81	0.88	0.84	0.91	0.91	0.73	0.72	0.93	0.92	0.9	0.92	0.74	0.83
	QDC	0.6	0.66	0.86	0.82	0.89	0.89	0.71	0.71	0.91	0.91	0.89	0.91	0.74	0.83
	KNN	0.52	0.59	0.79	0.61	0.88	0.89	0.72	0.73	0.88	0.88	0.87	0.88	0.7	0.74
	SVM	0.58	0.58	0.66	0.62	0.94	0.93	0.77	0.8	0.65	0.63	0.89	0.91	0.68	0.63
Math rot	LDC	0.75	0.74	0.8	0.74	0.86	0.87	0.67	0.68	0.9	0.9	0.9	0.9	0.72	0.75
	QDC	0.6	0.62	0.79	0.75	0.87	0.88	0.66	0.68	0.88	0.88	0.88	0.88	0.73	0.75
	KNN	0.54	0.6	0.79	0.65	0.85	0.87	0.68	0.68	0.83	0.84	0.83	0.84	0.69	0.63
	SVM	0.61	0.6	0.71	0.73	0.88	0.93	0.79	0.79	0.71	0.73	0.92	0.88	0.69	0.73

Daubechies mother wavelet (db1) in our experiments. During the process of decomposing the signal using EMD, it has been found that most of the segments decompose into 4 or more IMFs. Hence, we have used four levels of IMF decomposition for the sake of uniformity to form the feature vector. In order

to evaluate EMD and WT, both with the same size of the feature vector, we have used three-level wavelet decomposition, to obtain four wavelet transform coefficient vectors to maintain consistency. Each segment of signal is represented in terms of 192 statistics (4 IMFs or 4 WT coefficient vector \times

8 parameters \times 6 channels). Experiments were initially performed without using any feature selection method (WFS) and have been recorded for both EMD and WT. To remove redundancy from the selected pool of features, following six multivariate feature selection techniques were investigated: Euclidean distance (ED), Bhattacharyya distance measure (BD), Kullback–Leibler distance (KD), ratio of scatter matrices (SR), linear regression (LR) and maximum relevance minimum redundancy (mRMR). For all the multivariate filter methods, the top 25 features were incrementally included one by one to develop the decision model, as described by the sequential forward feature selection search method outlined in the Sect. 4.7. The maximum average classification accuracy of 10 runs of 10 cross-validations is quoted. The 12 combinations of the feature extraction and selection methods (referred to as EXT_SEL methods) are compared amongst each other in terms of the average classification accuracies obtained with four well-known classifiers: linear discriminant classifier (LDC), quadratic discriminant classifier (QDC), K nearest neighbour (kNN) and support vector machine (SVM). Table 1 shows the average classification accuracy obtained for the six subjects for the ten distinct combination of binary classes with four different classifiers ($N = 40$). Maximum classification accuracy for each pair of mental tasks is emphasized in bold in Table 1. We can observe the following from Table 1:

- For the combination of Baseline-Count task, the maximum classification accuracy, 92 %, is obtained using BD and SVM with both feature extraction methods.
- For the combinations of (i) Baseline task with letter, (ii) baseline task with math and (iii) baseline task with geometry rotation, maximum accuracy of 91, 93 and 94 %, respectively, is achieved with the combination of WT, LR and SVM.
- For the combinations of (i) counting task with letter, (ii) counting task with math and (iii) math with rotation, maximum accuracy of 91, 96 and 93 %, respectively, is achieved with the combination of WT, BD and SVM.
- For the combinations of letter with math, maximum accuracy of 95 % is achieved with two combinations (a) WT, BD and SVM and (b) WT, LR and SVM.
- For the combination of counting with rotation, maximum accuracy of 91 % is achieved with the combination of SR, LR and LDC.
- For the combination of letter with rotation, maximum accuracy of 94 %, is achieved with the combination of EMD, BD and SVM.

Table 2 and Fig. 2 show average classification accuracy over all subjects, tasks and classifiers for both feature extraction methods: EMD and WT. We can observe the following:

Table 2 Comparison between EMD and WT for all feature selection methods in terms of average accuracy (over all subjects, tasks and classifiers)

Feature selection	WFS	ED	BD	KD	SR	LR	mRMR
EMD	0.60	0.76	0.87	0.68	0.82	0.87	0.71
WT	0.63	0.69	0.89	0.70	0.83	0.89	0.71

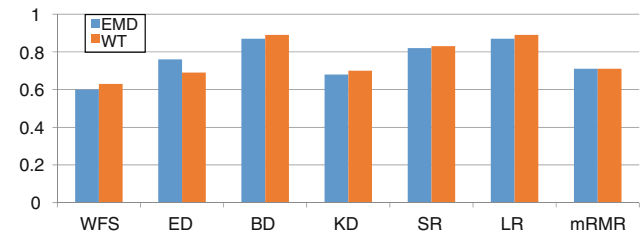


Fig. 2 Variation in average classification accuracy with the choice of feature selection method for EMD and wavelet

- The average classification accuracy of WT is better in comparison to EMD without involving any feature selection.
- For both feature extraction techniques, the classification accuracy improves with the use of feature selection.
- The average classification accuracy of WT is same or better in comparison to EMD for all feature selection techniques except for ED.
- The performance of LR and BD is better compared to other feature selection techniques for both EMD and WT.

Figures 3 and 4 show the average training time and the average testing time over all subjects and tasks for WT-based extraction for without feature selection (WFS) and all six multivariate filter methods. We can observe the following:

- In the comparison of the average training times, all six multivariate methods show improvement in computation times as compared to the time when no filter selection (WFS) is used.
- During training, all the multivariate methods perform equally well when used with LDC and QDC.
- During training, all multivariate methods show the largest improvement with QDC.
- Similarly, it can be seen that while comparing the average testing time, all six multivariate methods show improvement in computation times as compared to WFS.
- The computation time while testing with SVC is the minimal for all the multivariate filter methods.
- The mRMR multivariate method takes lesser time among all multivariate filter methods.

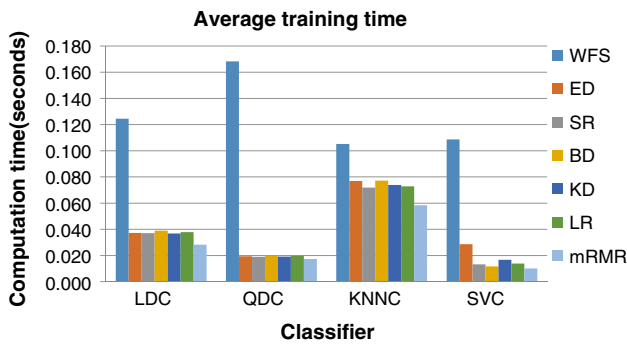


Fig. 3 Comparison of average training time of WFS and multivariate methods on feature vector formed using WT

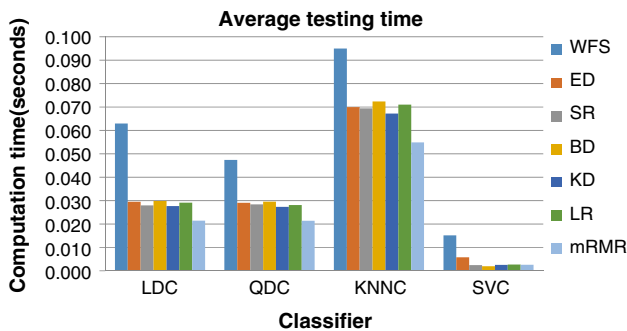


Fig. 4 Comparison of average testing time of WFS and multivariate methods on feature vector formed using WT

5.1 Ranking of various combinations of feature extraction and selection methods

In order to study the relative performances of the feature selection techniques in combination with a feature extraction method, a robust ranking mechanism is employed as suggested by Adhikari and Agrawal (2012). It ranks combination of a feature extraction method and a feature selection method according to the percentage gain (p) in classification accuracy with reference to the maximum classification accuracy achieved among two feature extraction methods without any feature selection (WFS). A mathematical description of this ranking procedure is as follows:

Let n_s, n_c and n_t denote the number of combinations of feature extraction and feature selection techniques, number of classifiers and number of different combination of tasks, respectively. Let $a_{ct}^s (c = 1, 2, \dots, n_c; t = 1, 2, \dots, n_t)$ be the average classification accuracy of the s th combination

of feature extraction and feature selection technique with the c th classifier for the t th task combination and a_{ct}^m be the maximum average accuracy among two feature extraction methods without feature selection with the c th classifier for the t th tasks combination. The percentage gain in accuracy is computed as follows:

$$p_{ct}^s = \left(\frac{a_{ct}^s - a_{ct}^m}{a_{ct}^m} \right) \times 100 \tag{22}$$

Then the average (over all the classifiers and task combination) percentage gain in accuracy for the s th technique is given by:

$$p^s = \frac{1}{n_c n_t} \sum_{c=1}^{n_c} \sum_{t=1}^{n_t} p_{ct}^s, \quad \forall s = 1, 2, \dots, n_s \tag{23}$$

Finally the rank, r^s , of each of the s th combination is assigned in such a way that:

$$r^a \leq r^b \quad \text{if} \quad p^a \geq p^b \tag{24}$$

Table 3 presents the percentage gain in the accuracy of the different combinations of the feature extraction and the feature selection methods. Figure 5 shows the percentage gain in accuracy of different combinations of the feature extraction and the feature selection methods and their corresponding ranks. From Table 3 and Fig. 5, we can observe that the combination of WT and LR performs the best in terms of percentage gain in accuracy; whereas, the combination of EMD with KD performs the worst.

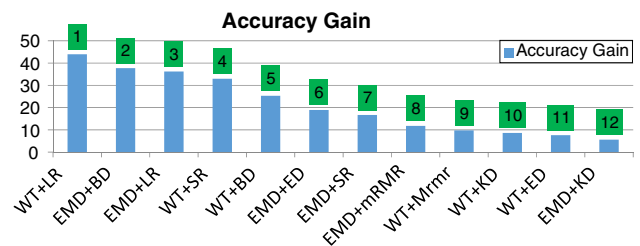


Fig. 5 Ranking of different combinations of feature extraction and selection methods

Table 3 Percentage gain in accuracy along with ranks for all combination of feature extraction and feature selection methods

Method	WT +LR	EMD +BD	EMD +LR	WT +SR	WT +BD	EMD +ED	EMD +SR	EMD +mRMR	WT +mRMR	WT +KD	WT +ED	EMD +KD
Accuracy gain	43.97	37.73	36.2	32.92	25.31	18.94	16.66	11.78	9.75	8.62	7.63	5.58
Rank	1	2	3	4	5	6	7	8	9	10	11	12

Table 4 Wilcoxon pair-wise table

Comparison	R^+	R^-	Z value	p value	Comparison	R^+	R^-	Z value	p value
WT+KD vs. WT+LR	0	820	-7.7043	1.32E-14	EMD+SR vs. WT+mRMR	799	21	-4.7292	2.25E-06
EMD+KD vs. WT+LR	0	820	-7.7011	1.35E-14	EMD+ED vs. EMD+KD	699	121	-4.392	1.12E-05
WT+BD vs. EMD+KD	820	0	-7.6914	1.46E-14	EMD+ED vs. WT+KD	654.5	165.5	-3.9736	7.08E-05
WT+LR vs. EMD+mRMR	820	0	-7.6719	1.69E-14	WT+SR vs. EMD+ED	783	37	-3.9586	7.54E-05
WT+BD vs. WT+KD	820	0	-7.6657	1.78E-14	EMD+SR vs. EMD+ED	789.5	30.5	-3.6281	2.85E-04
WT+BD vs. EMD+mRMR	820	0	-7.6286	2.37E-14	EMD+SR vs. WT+LR	49	771	-3.5599	3.71E-04
EMD+BD vs. EMD+KD	820	0	-7.6245	2.45E-14	EMD+ED vs. WT+ED	798	22	-3.4288	6.06E-04
EMD+KD vs. EMD+LR	0	820	-7.612	2.70E-14	EMD+SR vs. WT+BD	241.5	578.5	-2.9983	0.0027
WT+KD vs. EMD+LR	0	820	-7.596	3.05E-14	EMD+ED vs. EMD+mRMR	754.5	65.5	-2.997	0.0027
EMD+BD vs. WT+KD	820	0	-7.5797	3.46E-14	EMD+SR vs. WT+LR	62	758	-2.8691	0.0041
WT+ED vs. WT+LR	0	820	-7.5517	4.30E-14	EMD+BD vs. WT+LR	97.5	722.5	-2.8646	0.0042
WT+LR vs. WT+mRMR	820	0	-7.5384	4.76E-14	WT+SR vs. WT+LR	232.5	587.5	-2.5803	0.0099
WT+ED vs. WT+BD	0	820	-7.5182	5.55E-14	EMD+BD vs. WT+BD	53.5	766.5	-2.407	0.0161
EMD+BD vs. EMD+mRMR	820	0	-7.4799	7.44E-14	WT+BD vs. EMD+LR	692.5	127.5	-2.3	0.0214
WT+BD vs. WT+mRMR	820	0	-7.4576	8.81E-14	EMD+ED vs. WT+mRMR	703.5	116.5	-2.2834	0.0224
EMD+LR vs. EMD+mRMR	820	0	-7.4387	1.02E-13	WT+BD vs. WT+SR	349	471	-2.0228	0.0431
WT+ED vs. EMD+BD	0	820	-7.2615	3.83E-13	EMD+KD vs. EMD+mRMR	295	525	-1.8595	0.063
EMD+ED vs. WT+LR	0	820	-7.1927	6.35E-13	EMD+KD vs. WT+mRMR	281.5	538.5	-1.6863	0.0917
WT+ED vs. EMD+LR	0	820	-7.1888	6.53E-13	EMD+SR vs. EMD+LR	408.5	411.5	-1.2545	0.2097
EMD+BD vs. WT+mRMR	820	0	-7.0975	1.27E-12	EMD+SR vs. EMD+BD	415	405	-1.1883	0.2347
EMD+ED vs. WT+BD	0	820	-7.058	1.69E-12	WT+KD vs. EMD+mRMR	359	461	-1.1667	0.2433
EMD+LR vs. WT+mRMR	820	0	-7.0561	1.71E-12	WT+ED vs. WT+mRMR	183.5	636.5	-1.1367	0.2556
EMD+ED vs. EMD+LR	0.5	819.5	-6.2683	3.65E-10	EMD+KD vs. WT+KD	145.5	674.5	-1.1049	0.2692
EMD+ED vs. EMD+BD	0	820	-6.2464	4.20E-10	WT+ED vs. EMD+mRMR	277.5	542.5	-1.0402	0.2983
EMD+SR vs. EMD+KD	754.5	65.5	-5.442	5.27E-08	WT+KD vs. WT+mRMR	331	489	-0.9881	0.3231
WT+SR vs. EMD+KD	759	61	-5.3158	1.06E-07	WT+ED vs. WT+KD	397	423	-0.7566	0.4493
EMD+SR vs. WT+ED	807	13	-5.2492	1.53E-07	WT+BD vs. WT+LR	297	523	-0.7397	0.4595
EMD+SR vs. WT+KD	757	63	-5.1516	2.58E-07	EMD+SR vs. WT+SR	212	608	-0.7378	0.4606
WT+SR vs. WT+ED	801	19	-5.1472	2.64E-07	EMD+mRMR vs. WT+mRMR	357	463	-0.4142	0.6787
WT+SR vs. WT+KD	760	60	-5.1446	2.68E-07	WT+SR vs. EMD+BD	439.5	380.5	-0.357	0.7211
WT+SR vs. EMD+mRMR	794.5	25.5	-4.9684	6.75E-07	WT+SR vs. EMD+LR	443.5	376.5	-0.3521	0.7247
EMD+SR vs. EMD+mRMR	798.5	21.5	-4.9025	9.46E-07	EMD+BD vs. EMD+LR	385.5	434.5	-0.0483	0.9615
WT+SR vs. WT+mRMR	792.5	27.5	-4.869	1.12E-06	WT+ED vs. EMD+KD	462.5	357.5	-0.0482	0.9616

5.2 Comparison among different feature extraction-selection (Ext_Sel) combinations

To compare the performance of the various feature selection methods in conjunction with the two feature extraction methods, Wilcoxon test and Friedman statistical tests are used, which are based on the research work of [Demsar \(2006\)](#) and [Derrac et al. \(2011\)](#).

5.2.1 Statistical test based on pair-wise comparisons

To compare the results pair-wise, the signed rank test proposed by [Wilcoxon \(1945\)](#) is used, that aims to detect signif-

icant differences between each pair of Ext_Sel combination under consideration. For each of the i th case of the total N cases, let d_i denote the difference in the classification accuracies of the two Ext_Sel combinations. The differences are ranked according to the absolute values, and average ranks are assigned in case of ties. R^+ is the sum of the ranks of those cases where the first Ext_Sel combination outperforms the second combination and R^- is the sum of the ranks of those cases where the second Ext_Sel combination performs better. $T = \min(R^+, R^-)$. The Wilcoxon statistic, is given by

$$z = \frac{T - \frac{1}{4}N(N + 1)}{\sqrt{\frac{1}{24}N(N + 1)(2N + 1)}} \tag{25}$$

Table 5 Average ranking of the combination algorithms using Friedman statistic

Algorithm	Rank	Algorithm	Rank
WT+LR	2.23	EMD+ED	7.61
WT+BD	3.08	WT+KD	9.08
WT+SR	4.08	WT+mRMR	9.24
EMD+LR	4.26	EMD+mRMR	9.27
EMD+BD	4.41	EMD+KD	9.64
EMD+SR	4.93	WT+ED	10.16

The statistic is distributed normally. A level of significance α is chosen to determine the level at which the hypothesis may be rejected. The significance of a result may also be represented by the p value: smaller the p value, stronger is the evidence against the null hypotheses. Choosing the significance level, $\alpha = 0.05$, the null hypothesis is rejected if $|z| > 1.96$ (equivalently if $p < 0.05$).

Table 4 shows the result obtained for all possible comparisons among the 12 Ext_Sel combinations. R^+ , R^- , z value and p value are tabulated for comparison. The statistically better performing Ext_Sel combination is shown in bold.

5.2.2 Statistical test based on comparisons with the control method

The null hypothesis assumes that each of the k Ext_Sel combinations are equivalent in terms of their performance. A comparison of multiple algorithms can be accomplished after ranking them according to their classification accuracy.

For each case, rank ranging from 1 to k is associated with every Ext_Sel combination. Rank value 1 and k denotes the best and worst result, respectively. Let this rank be denoted by r_i^j ($1 \leq i \leq N$, $1 \leq j \leq k$). For each Ext_Sel combination, j , let R^j denote the average of ranks over the N experimental observations. The ranks computed are given in Table 5 for the 12 Ext_Sel combinations. In this case, the best performing Ext_Sel combination has the least rank value of 2.23 for WT (extraction) combined with LR (selection).

The statistical hypothesis test used is the one proposed by Iman and Davenport (1980), which is based on the following statistic:

$$F_{ID} = \frac{(N - 1)\chi_F^2}{N(k - 1) - \chi_F^2} \tag{26}$$

which is distributed according to an F distribution with $k - 1$ and $(k - 1)(N - 1)$ degrees of freedom, where χ_F^2 is the Friedman statistics (1937) given by $\frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right]$. The p value computed by Iman and Davenport statistic is 2.22 E-16 which suggests the significant difference among

Table 6 Adjusted p values of comparison with the control method (WT+LR) obtained using Friedman statistic

Algorithm	Unadjusted p	pHolm	pHoch	pHomm
WT+ED	8.38E-23	9.22E-22	9.22E-22	9.22E-22
EMD+KD	4.37E-20	4.37E-19	4.37E-19	4.37E-19
EMD+mRMR	2.95E-18	2.65E-17	2.65E-17	2.36E-17
WT+mRMR	3.87E-18	3.10E-17	3.10E-17	3.10E-17
WT+KD	1.96E-17	1.37E-16	1.37E-16	1.37E-16
EMD+ED	2.61E-11	1.57E-10	1.57E-10	1.57E-10
EMD+SR	8.58E-04	0.004289	0.004289	0.004289
EMD+BD	0.006981	0.027923	0.027923	0.02403
EMD+LR	0.012015	0.036045	0.036045	0.033988
WT+SR	0.022659	0.045318	0.045318	0.045318
WT+BD	0.291748	0.291748	0.291748	0.291748

the Ext_Sel combinations considered in our study, hence rejecting the null hypothesis.

The performance of an Ext_Sel combination is studied with respect to a control method, i.e. the one that emerges with the lowest rank (best performer). The test statistic for comparing the m th combination with n th combination, z is given as:

$$z = (R_m - R_n) / \sqrt{\frac{k(k+1)}{6N}} \tag{27}$$

where R_m and R_n are the average ranks of the combinations being compared. However, these p values so obtained are not suitable for comparison with the control method. Instead, adjusted p values (Derrac et al. 2011) are computed that take into account the error accumulated and provide the correct correlation. For this, a set of post hoc procedures is defined and adjusted p values are computed to be used in the analysis. For pair-wise comparisons, the widely used post hoc methods to obtain adjusted p values are (Derrac et al. 2011): Bonferroni–Dunn, Holm, Hochberg and Hommel procedures. Table 6 shows adjusted p values for the aforementioned procedures. The values in Table 6 represent the p value when pair-wise comparison with control method (WT+LR) is conducted. The bold values suggest significant difference of WT+LR with all combinations except with WT+BD, at the significance level of 0.05. This emphasizes that WT+LR combination performs better than all other combinations.

6 Conclusion

Brain–computer interface assists physically challenged people to communicate with the help of their brain electroencephalogram (EEG) signal. Wavelet transform and empirical mode decomposition are better choices to handle non-linear and non-stationary brain signal to extract relevant features.

Statistical parameters are used to compactly represent the extracted features. However, features from multiple channels generate a large size of feature vector, but the available number of samples is small. Under such a situation, the performance of the learning model may degrade, hence dimensionality reduction is required. In this paper, we have investigated and compared six well-known multivariate filter methods to determine a minimal subset of relevant and non-redundant features. Experimental results demonstrate improvement in the classification accuracy and reduction in the computation time of the learning model with the use of feature selection. It is also noted that the performance of linear regression is better compared to the other multivariate filter methods. We have defined a performance measure to rank various combinations of feature extraction and feature selection methods. It is found that wavelet in combination with linear regression performs the best. Further, Wilcoxon pair-wise test and Friedman test are also carried out to observe significant difference among various combinations. Both statistical tests reaffirm that the performance of wavelet transform in combination with linear regression is significantly different from majority of the combinations.

Experiments in this study were carried out with Daubechies mother wavelet which is fixed and independent of signal. In future, we would like to develop some adaptive hybrid approach of EMD and wavelet transform to further improve the performance of mental task classification. The EEG signal is non-stationary and there is a high variability of the EEG signals across sessions. There is a need to determine an appropriate subset of features which can adapt to the dynamic nature of the EEG signal. Hence, it will be interesting to apply on-line feature selection methods to the EEG signals, changing the feature ranking lists over time.

Acknowledgments The first author expresses his gratitude to the Council of Scientific and Industrial Research (CSIR), India, for the obtained financial support in performing this research work. We also thank the reviewers for the constructive and valuable review of our paper that has helped us to further strengthen the overall quality of the paper.

References

- Adhikari R, Agrawal RK (2012) Performance evaluation of weights selection schemes for linear combination of multiple forecasts. *Artif Intell Rev*. doi:10.1007/s10462-012-9361-z
- Anderson CW, Stolz EA, Shamsunder S (1998) Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Trans Biomed Eng* 45(3):277–286
- Babiloni F, Cincotti F, Lazzarini L, Millan J, Mourino J, Varsta M, Heikkonen J, Bianchi L, Marciani MG (2000) Linear classification of low-resolution EEG patterns produced by imagined hand movements. *IEEE Trans Rehab Eng* 8(2):184–186
- Bashashati A, Faourechhi M, Ward RK, Brich GE (2007) A survey of signal processing algorithms in brain computer interface based on electrical brain signals. *J Neural Eng* 4:R32–R57
- Basseville M, Benveniste A (1983) Sequential segmentation of non-stationary digital signals using spectral analysis. *Inf Sci* 29(1):57–73
- Bellman RE (1961) *Adaptive control processes: a guided tour*. Princeton University Press, Princeton
- Bhattacharyya A (1943) On a measure of divergence between two statistical populations defined by their probability distributions. *Bull Calcutta Math Soc* 35:99–110
- Bhattacharyya S, Sengupta A, Chakraborti T, Konar A, Tibarewala D (2014) Automatic feature selection of motor imagery EEG signals using differential evolution and learning automata. *Med Biol Eng Comput* 52(2):131–139
- Bostanov V (2004) BCI competition 2003-data sets Ib and Iib: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Trans Biomed Eng* 51(6):1057–1061
- Cabrera A, Farina D, Dremstrup K (2010) Comparison of feature selection and classification methods for a brain-computer interface driven by non-motor imagery. *Med Biol Eng Comput* 0140–0118(48):123–132
- Corralejo R, Hornero R, Alvarez D (2011) Feature selection using a genetic algorithm in a motor imagery-based brain computer interface. In: *Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp 7703–7706
- Cvetkovic D, Übeyli ED, Cosic I (2008) Wavelet transform feature extraction from human PPG, ECG, and EEG signal responses to ELF PEMF exposures: A pilot study. *Digit Signal Process* 18(5):861–874
- Daubechies I (1990) The wavelet transform. Time–frequency localization and signal analysis. *IEEE Trans Inf Theory* 3(5):961–1005
- Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
- Derrac J, Garcia S, Molina D, Herrera F (2011) A practical tutorial on the use of nonparametric statistical tests as a methodology or comparing evolutionary and swarm intelligence algorithms. *Swarm Evol Comput* 1:3–18
- Devijver PA, Kittler J (1982) *Pattern recognition: a statistical approach*. PHI
- Dias N, Kamrunnahar M, Mendes P, Schiff S, Correia J (2010) Feature selection on movement imagery discrimination and attention detection. *Med Biol Eng Comput* 48:331–341
- Diez PF, Mut V, Lacier E (2009) Application of the empirical mode decomposition to the extraction of features from EEG signals for mental task classification. In: *31st Annual International Conference of the IEEE EMBS Minneapolis*, pp 2579–2582
- Freeman WJ (1999) Comparison of brain models for active vs. passive perception. *Inf Sci* 116:97–107
- Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 32(1937):674–701
- Garrett D, Peterson D, Anderson C, Thaut M (2003) Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Trans Neural Syst Rehab Eng* 11(2):141–144
- Graimann B, Huggins JE, Schlogl A, Levine SP, Pfurtscheller G (2003) Detection of movement-related desynchronization patterns in ongoing single-channel electrocardiogram. *IEEE Trans Neural Syst Rehab Eng* 11(3):276–281
- Groissboeck W, Lughofer E, Klement EP (2004) A comparison of variable selection methods with the main focus on orthogonalization. *Adv Soft Comput* 479–486
- Gupta A, Agrawal RK (2012) Relevant feature selection from EEG signal for mental task classification. In: *Pacific-Asia conference on knowledge discovery and data mining (PAKDD)*, in part II. *Lecture Notes in Computer Science*, vol 7302, pp 431–442
- Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182

- Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, New York
- Haury AC, Gestraud P, Vert JP (2011) The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One* 6(12). doi:[10.1371/journal.pone.0028210](https://doi.org/10.1371/journal.pone.0028210)
- Hazarika N, Chen JZ, Tsoi AC, Sergejew A (1997) Classification of EEG signals using the wavelet transform. *Signal Process* 59(1):61–72
- Hsu WY, Sun YN (2009) EEG-based motor imagery analysis using weighted wavelet transform features. *J Neurosci Methods* 176(2):310–318
- Huang NE, Shen Z, Long SR, Wu ML, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc R Soc Lond A* 454:903–995
- Iman R, Davenport J (1980) Approximations of the critical region of the Friedman statistic. *Commun Stat* 9(1980):571–595
- Kaleem MF, Sugavaneswaran L, Guergachi A, Krishnan S (2010) Application of empirical mode decomposition and teager energy operator to EEG signals for mental task classification. In: Annual international conference of the engineering in medicine and biology society (EMBC). IEEE Press, New York, pp 4590–4593
- Kauhanen L, Nykopp T, Lehtonen J, Jylanki P, Heikkonen J, Rantanen P, Alaranta H, Sams M (2006) EEG and MEG brain-computer interface for tetraplegic patients. *IEEE Trans Neural Syst Rehab Eng* 14(2):190–193
- Keirn ZA, Aunon JI (1990) A new mode of communication between man and his surroundings. *IEEE Trans Biomed Eng* 37(12):1209–1214
- Kohavi R, John G (1997) Wrapper for feature subset selection. *Artif Intell* 97(1–2):273–324
- Koprinska I (2009) Feature selection for brain-computer interfaces. In: International workshop on new frontiers in applied data mining (PAKDD), LNCS, vol 5669, pp 106–117
- Kronegg J, Chanel G, Voloshynovskiy S, Pun T (2007) EEG-based synchronized brain-computer interfaces: a model for optimizing the number of mental tasks. *IEEE Trans Neural Syst Rehab Eng* 15(1):50–58
- Kullback S, Liebler RA (1951) On information and sufficiency. *Ann Math Stat* 22:79–86
- Lakany H, Conway BA (2007) Understanding intention of movement from electroencephalograms. *Expert Syst* 24:295–304
- Lempel A, Ziv J (1976) On the complexity of finite sequences. *IEEE Trans Inf Theory* 22:75–81
- Li Y (2004) On incremental and robust subspace learning. *Pattern Recognit* 37(7):1509–1518
- Lughofer E (2011) On-line incremental feature weighting in evolving fuzzy classifiers. *Fuzzy Sets Syst* 163(1):1–23
- Mallat GS (1989) A theory for multi-resolution signal decomposition the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11(7):674–693
- Mosquera C, Verleysen M, Navia Vazquez A (2010) EEG feature selection using mutual information and support vector machine: a comparative analysis. In: 32nd annual international IEEE EMBC conference, pp 4946–4949
- Murugappan M, Ramachandran N, Sazali Y (2010) Classification of human emotion from EEG using discrete wavelet transform. *J Biomed Sci Eng* 3(4):390–396
- Ocak H (2009) Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. *Expert Syst Appl* 36(2):2027–2036
- Park HS, Yoo SHY, Cho SB (2007) Forward selection method with regression analysis for optimal gene selection in cancer classification. *Int J Comput Math* 84(5):653–668
- Peng H, Loung F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency. Max-relevance and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27(8):1226–1238
- Penny WD, Roberts SJ, Curran EA, Stokes MJ (2000) Eeg-based communication: a pattern recognition approach. *IEEE Trans Rehab Eng* 8(2):214–215
- Pfurtscheller G, Neuper C, Schlogl A, Lugger K (1998) Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. *IEEE Trans Rehab Eng* 6(3):316–325
- Rajer I, Lorenz K (2013) Genetic algorithm and forward method for feature selection in EEG feature space. *J Theor Appl Comput Sci* 7(2):72–82
- Rodríguez-Bermúdez G, García-Laencina PJ, Roca-Dorda J (2013) Efficient automatic selection and combination of EEG features in least squares classifiers for motor imagery brain-computer interfaces. *Int J Neural Syst* 23(4). doi:[10.1142/S0129065713500159](https://doi.org/10.1142/S0129065713500159)
- Sakar OC, Kursun O, Gurgen F (2012) A feature selection method based on kernel canonical correlation analysis and the minimum redundancy-maximum relevance filter method. *Expert Syst Appl* 39:3432–3437
- Shannon CE (1948) A mathematical theory of communication. *AT T Tech J* 27(379–423):623–656
- Wilcoxon F (1945) Individual comparisons by ranking methods. *Biometrics* 1:80–83
- Wolpaw RJ, Birbaumer N, McFarland JD, Pfurtscheller G, Vaughan MT (2002) Brain-computer interfaces for communication and control. *Clin Neurophysiol* 767–791