A Hierarchical Meta-model for Multi-Class Mental Task Based Brain-Computer Interfaces

Akshansh Gupta[a,*], R. K. Agrawal[a], Jyoti Singh Kirar[a], Baljeet Kaur[b], Weiping Ding[c], Chin-Teng Lin[d], Andreu Perez, Javier[e], Mukesh Prasad[d]

[a]*School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi.*
[b]*Hansraj College, University of Delhi, New Delhi, India*
[c]*School of Computer Science and Technology, Nantong University, Nantong, China*
[d]*Faculty of Engineering and Information Technology, University of Technology Sydney, Australia*
[e] *School of Computer Science and Electronic Engineering, University of Essex*

Abstract

In the last few years, many research works have been suggested on Brain-Computer Interface (BCI), which assists severely physically disabled persons to communicate directly with the help of electroencephalogram (EEG) signal, generated by the thought process of the brain. Thought generation inside the brain is a dynamic process, and plenty thoughts occur within a small time window. Thus, there is a need for a BCI device that can distinguish these various ideas simultaneously. In this research work, our previous binary-class mental task classification has been extended to the multi-class mental task problem. The present work proposed a novel feature construction scheme for multi mental task classification. In the proposed method, features are

*Corresponding author
Email addresses: akshanshgupta83@gmail.com (Akshansh Gupta), rkajnu@gmail.com (R. K. Agrawal), kirarjyoti@gmail.com (Jyoti Singh Kirar), baljeetkaur26@hotmail.com (Baljeet Kaur), dwp9988@163.com (Weiping Ding), Chin-Teng.Lin@uts.edu.au (Chin-Teng Lin), javier.andreu@imperial.ac.uk (Andreu Perez, Javier), Mukesh.Prasad@uts.edu.au (Mukesh Prasad)

extracted in two phases. In the first step, the wavelet transform is used to decompose EEG signal. In the second phase, each feature component obtained is represented compactly using eight parameters (statistical and uncertainty measures). After that, a set of relevant and non-redundant features is selected using linear regression, a multivariate feature selection approach. Finally, optimal decision tree based support vector machine (ODT-SVM) classifier is used for multi mental task classification. The performance of the proposed method is evaluated on the publicly available dataset for 3-class, 4-class, and 5-class mental task classification. Experimental results are compared with existing methods, and it is observed that the proposed plan provides better classification accuracy in comparison to the existing methods for 3-class, 4-class, and 5-class mental task classification. The efficacy of the proposed method encourages that the proposed method may be helpful in developing BCI devices for multi-class classification.
Keywords: Brain Computer Interface, Mental Tasks Classification, Feature Extraction, Feature Selection, Support Vector Machine

---

1. Introduction

The human brain has the capability of differentiating multiple courses of action without any difficulty. In previous studies, a significant part of research works contains to distinguish between two different tasks at a given frame of time. There are few research works suggested for multitasking classification [1, 2, 3, 4, 5]. A BCI system, which could be differentiated more than two mental activities at a given time, is known as multi-class mental task based BCI system. The application of multi class BCI system for stroke

rehabilitation [6], multiple rehabilitation targets simultaneously [7].

It becomes harder to classify a test sample of multi-class mental task framework with the increase in the number of mental tasks. The computational complexity of the multi-class mental task is much high in comparison with a binary class mental task with the comparable amount of data.

There are only few BCI models [4, 3, 5] have been discussed to distinguish more than two tasks at a given instance of time frame. The research works [8, 9, 10, 11, 12] have demonstrated that with the employment of feature selection, classification accuracy improves for binary mental task classification. To the best of our knowledge, feature selection has not been suggested in research work related to multi-class mental task classification. This motivated us to investigate feature selection method for multi-mental task classification problem. It has been observed in the research work [10] that the combination of feature extraction using Wavelet transform (WT) and feature selection using Linear Regression (LR) has given the best set of features that enhance the performance of the classifier for the binary mental task classification. Therefore in this paper, we have used the same combination to extract and find the set of relevant and non-redundant features for the multi-class mental task classification problem. Optimal decision tree (ODT) based multi-class SVM is utilized as a multi-class classifier to build the decision model. The overall flow diagram of the suggested model has been shown in Figure 1.

The major contributions of this paper include:

1. The proposed method utilized Optimal decision tree based on amalgamation with support vector machine (SVM) to build decision model to distinguish multiple mental tasks.

2. To provide a combination of more robust feature selection and feature extraction method that can select a reduced subset of relevant and non-redundant features for multi-class classification.

The paper is structured as : In section 2, the state of art of multi-class BCI is given. Section 3 contains the brief description of feature extraction. Discussion of dimension reduction using LR is given in section 4. An optimal decision tree based support vector machine (ODT-SVM) is explained in section 5. Experimental data and the related discussion are given in section 6, and finally, section 7 draws the conclusion.

2. Related Works

For multi-class BCI, the majority of the research works have been carried out for two categories: sensory-motor activity [1, 2]and response to the mental task [3, 4, 5]. One of the most elegant methods for the identification of sensory-motor rhythms is the method of common spatial patterns (CSPs) proposed by [13]. The extension of CSP to multi-class CSP has been done on the basis of pairwise classification and voting mechanism [14]. Composition Kernal Support Vector Machine (CKS) based CSP (CKSCSP) method is used to determine a compact set of relevant electrodes for motor imagery based BCI [15]. Fuzzy techniques have also been used to discriminate motor imagery pattern using more straightforward features such as phase synchrony [16]. For the steady-state visual evoked potentials (SSVEPs), a spatiotemporal feature from the EEG signals are extracted using multivariate linear regression [17]. In the work of [18], to classify voluntary hand movement direction, regularized wavelet-common spatial pattern, Reg-W-CSP, a method

has been employed for extracting features from EEG signal. As the concern of response to the mental task category, [3] have utilized three type of power of spectral density methods viz Wiener-Khinchine (WK) with Parzen smoothing window, WK with Tukey window smoothing and 6^{th} order autoregressive model to extract features for 3-class mental task classification. Fuzzy ARTMAP classifier has been utilized for three class mental task classification. Welch periodogram has been making use of extracting power spectrum features from the EEG signal, and the different number of a frequency band is calculated with the help of asymmetric ratio for the multi-class mental task [4]. Fisher Discriminant Analysis (FDA) and Mahalanobis distance based classifier have been utilized in their work to build decision model. Wavelet packet entropy and Granger causality have been used for extracting the feature from the EEG signal, and the extracted features were used to build learning model using multiple kernels support vector machine [5].

Among the research work on BCI models for multi-class tasks classification, the commonly used classifiers such as the artificial neural network(ANN), K-Nearest Neighbour (K-NN), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM)have been used for classification. ANN is a natural multi-class classifier but requires extensive computation time and memory. Also in neural networks, the number of hidden layers and the number of nodes in layers has to be tuned to achieve better performance [19, 20]. In K-NN classifier, no model is learned from training data, and decision for the new test sample is determined based on the class label of a majority of training samples, which are nearest to the test sample. In this method, determining optimal choice of K nearest neighbors to achieve better performance

is very time-consuming.

On the other hand, LDA and SVM were originally designed for binary classification problems. To extend these to multi-class problems different combining schemes are suggested in the literature. One of the straightforward extensions is to combine several binary classifiers. It has been shown in work [21] that classifier, which is formed by many different binary classifiers, is almost as effective as all-together classifiers when the underlying binary classifiers are well-tuned. To combine the binary classifiers, two schemes are commonly used: (i) One-versus-One (OvO) and (ii) One-versus-All (OvA).

In OvO, a multi-class problem is split into a set of different binary class problems which consists of all possible combinations of binary classes, i.e., for a k-class problem, there exist $\frac{k(k-1)}{2}$ binary class problems. For each pair of binary combination, there exists one classifier for discriminating the two classes [22]. Prediction of output class can be obtained by aggregation of the output of different binary models. In literature there are many protocols for aggregation techniques, such as voting strategy, max wins rule, weighted voting strategy, pairwise coupling [23] and a learning valued preference for classification based on fuzzy performance modeling [24, 25].
In OvA scheme, there are k classifiers for a given k class problem, one for each class. In each classifier one class (positive) of data is classified against rest of classes (negative) of data. Final decision can be obtained from maximum confidence level of these classifiers.

The construction of multi-class SVM is an on-going research issue. In traditional OvO SVM approach, there are some regions which cannot be

classified (as shown in Figure 2). To resolve issues of the unclassifiable regions, in some literature, Decision Directed Acyclic Graph (DDAG) SVM [26] has been proposed which is based on Decision Tree (DT) based SVM (DT-SVM). In the work [26], it has been shown in the literature, that generalization ability of a given classifier deteriorates on an existence of unclassified regions. The unclassified regions appear, and the performance of the classifier degrades when the number of classes is more than two in a classification problem. Motivated by the research work [26], we have used decision tree based multi-class SVM for classification in this work.

3. Feature Extraction

Features are extracted from the EEG signal in two steps: (1) EEG signal is decomposed by Wavelet Transform (WT) and (2) phase statistical, and uncertainty parameters are calculated from each decomposed signal to represent the signal more compactly. A brief description of WT and the parameters is discussed below.

3.1. Wavelet Transform(WT)

Wavelet analysis is a multi-resolution mathematical tool which provides both spectral and temporal information of the signal. Wavelet transform of discrete signals is known as discrete wavelet transform (DWT). DWT analyses a signal by decomposing it into an approximation component (low band-pass) and detail components (high band-pass). It employs a scaling function to generate approximation component and a wavelet function to find detail components that encode the difference between two adjacent approximations.

Scaling and Wavelet Function

Consider $\{\varphi_{j,k}(m)\}$, a set of expansion functions, which consists of binary scaling and translations of function $\varphi(m)$ and is given by:

$$\varphi_{j,k}(m) = 2^{j/2}\varphi(2^j m - k) \tag{1}$$

Since the shape of $\varphi_{j,k}(m)$ varies with j,$\varphi(m)$ is known as scaling function. Subspace is spanned over k for any j is expressed as:

$$V_j = \underset{k}{span}\,\{\varphi_{j,k}(m)\} \tag{2}$$

For a given scaling function that follows necessary condition of multi-resolution, a function $\psi(m)$, which is spaned the difference between any two adjacent scaling subspaces V_j and V_{j+1}, can be defined and is called wavelet or mother wavelet function. The set $\{\psi_{j,k}(m)\}$ of wavelets is expressed as:

$$\psi_{j,k}(m) = 2^{j/2}\psi(2^j m - k) \tag{3}$$

Similar to V_j space, there exists a space W_j which can be obtained as:

$$W_j = \underset{k}{span}\,\{\psi_{j,k}(m)\} \tag{4}$$

The scaling function and wavelet function subspaces are related as:

$$\mathbf{V}_{j+1=}\mathbf{V}_j \oplus W_j \tag{5}$$

where $\oplus$ denotes ring sum.

DWT in One Dimension

A discrete 1D signal, $x(m)$, can be expanded in terms of scaling function $\phi(m)$ and wavelet function $\psi(m)$ [27] as:

$$x(m) = \frac{1}{\sqrt{M}} \sum_{n} W_{\varphi}(P,n)\varphi_{P,n}(m) + \frac{1}{\sqrt{M}} \sum_{p=1}^{P} \sum_{n} W_{\psi}(p,n)\psi_{p,n}(m) \tag{6}$$

where P denotes the level of decomposition, W_{ϕ} is scaling or approximation coefficients, W_{ψ} is known as wavelet or detail coefficients, $n = \left\{0, 1, 2, ..., \frac{M}{2^P} - 1\right\}$ and $\frac{1}{\sqrt{M}}$ is the normalize factor which imposed total energy change [28]. $x(m)$, $\phi_{p,n}(m)$ and $\psi_{p,n}(m)$ are functions of discrete variable $m = \{0, 1, 2, ..., M-1\}$. For a given one dimension signal $x(m)$,the signal is decomposed into a set of sub-band with help of sub-band coding as shown in Figure 3. Here g and h are high pass and low pass filter respectively and A1 and D1 are approximation and detailed coefficients at level 1 respectively. Decomposition is done by down sampling and synthesis can be done with help of up sampling.

3.2. Parametric Feature Vector Formulation

The concrete characterization of the EEG signal is carried out with the help of following statistical parameters of the Wavelet coefficients of the signal. Some of these parameters depict linear virtue of the EEG signal and other are representive of non-linear properties of the signal [12, 9, 10].

Mean

The first order moment of central tendency is known as mean. If there are n observations $(x_1, x_2 \ldots, x_n)$ then mean is given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

Root Mean Square (RMS)

The sinusoidal property of the signal can be expressed in terms root mean square values, denoting that signal has many positive and negative peaks. The value of RMS is considered quite informative because it presents power of the signal. It is given by:

$$rms\,(x_1, x_2, \ldots x_n) = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \ldots + x_n^2)} \tag{8}$$

Variance

The spreadness of the data around mean, a second order moment of the central tendency measure, is known as variance. The variance of the data is given by:

$$var(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{9}$$

The square root of variance has been known as standard deviation which is given by:

$$\sigma = \sqrt{var(x_1, x_2, \ldots x_n)} \tag{10}$$

Skewness

Asymmetry of distribution with respect to mean value of the signal of can be quantified by third order moment of statistics, known as Skewness .

It is a pure number which denotes bending nature of the signal around mean value of the signal on either side. It is defined as:

$$skewness(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{\sigma}\right)^3 \tag{11}$$

Kurtosis

Relative spikeness or flatness of signal with respect to the normal distributed signal can be known by the fourth order statistics, Kurtosis. It can be represented as

$$Kur\,(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{\sigma}\right)^4 \tag{12}$$

Lempel-Ziv Complexity

This complexity was first introduced by [29]. It quantifies the characteristics of degree of order or disorder and development of spatio-temporal patterns of the signal. It gives number of distinct patterns in a given finite sequence and reflects the rate of occurrence of new symbols in the pattern. Its value lies between 0 and 1; 0 indicates pure static and 1 represents randomness. If $L(n)$ is the length of encoded n observations then LZ complexity is given by:

$$C_{LZ} = \frac{L\,(n)}{n} \tag{13}$$

Central and Maximum frequency

These values also show how much frequency content is centralized over the signal and the maximum frequency present in the signal. The frequency content can be analysed by discrete Fourier transform of the signal, and is given as

$$X\,(f) = \sum_{n=-\infty}^{\infty} x\,[n]\,e^{-j2\pi fn} \tag{14}$$

Shannon Entropy

It quantifies how much uncertainty is possessed by the signal, i.e. randomness of signal. Higher entropy also means more randomness is present in the signal. If p_i is the probability associated with variable x_i in a set of n observations then entropy is also expressed as:

$$H(x) = -\sum_i p_i \log_2 (p_i) \tag{15}$$

4. Feature Selection

It can be observed in the research work [11] that improved classification accuracy is achieved with the set of highly related features obtained with the use of univariate feature selection modalities in comparison to the learning model developed without taking advantage of feature selection modalities. But univariate method ignores the correlation among the features while determining relevant features. Hence the performance of the learning model may degrade by using redundant or correlated features.

In this work, relevant and non-redundant subset of features are determined by utilizing commonly used multivariate filter method namely Linear regression [30] .

4.1. Linear Regression

In literature, Regression analysis is used as a well-established statistical method that determines the relationship of independent variable over dependent variable. The target variable is considered as the dependent variable and the features affected by these target variables are determined. This

method can capture the linear dependices of a response variable with two or more explanatory variables . Multiple regression model has been adopted to determines the causal effect of multiple features to the target variable. A multiple regression model with k independent varying quantities $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k$ and a output response y is given by [30]:

$$y_i = \beta_0 + \beta_1 f_{i1} + ... + \beta_k f_{ik} + \zeta_i, i = 1, 2, ..., n \tag{16}$$

where $\beta_0, \beta_1, ..., \beta_k$ are constants estimated by class label y and observed values of $\mathbf{X}$. The sum of squared error (SSE), a sum of the squared residuals can be given by:

$$SSE = \sum_{i=1}^{n} (y_i - y_i^p)^2 \tag{17}$$

where y_i and y_i^p are target and predicted values respectively. The lower value of SSE demonstrates better regression model. The total sum of squares (SSTO) is given by:

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{18}$$

where $\bar{y}$ is the average value of $y_i, i = 1, 2, ..., n$. The criterion value J_{LR} is given as:

$$J_{LR} = 1 - \frac{SSE}{SSTO} \tag{19}$$

The value of J_{LR} lies between 0 and 1. In a linear regression analysis, the feature for which the value of J_{LR} is higher is selected.

4.2. Sequential Forward Feature Selection

To find out a relevant and non-redundant subset of features using linear regression, various sub-optimal search methods are suggested in literature.

A greedy approach based sequential forward feature selection search method has been used in this work. It is faster approach with a time complexity of $O(d^2)$ where d is the dimension of feature vector. The outline of the algorithm using linear regression is given in algorithm 1.

The output of this algorithm is a set of non-redundant and relevant features

Algorithm 1: Sequential Forward Feature Selection

1 Given a set of d features, $R = \{\mathbf{f}_1, \mathbf{f}_2, \ldots \mathbf{f}_d\}$;

2 Initialization: $S = \emptyset$ // Initial empty set of relevant and non-redundant features;

3 The single best feature is selected which optimizes a criterion function, $J(.)$ $\boldsymbol{f}_i = \text{optimum}_i\, J(\boldsymbol{f})$; $S = S \cup \boldsymbol{f}_k$; $R = R - \boldsymbol{f_k}$;

4 Sets of features are formed using one of the remaining features from the set R and the already selected set of features, S. Compute $\boldsymbol{f}_j = \text{optimum}_i\ J(S \cup \boldsymbol{f}_i)$; $S = S \cup \boldsymbol{f}_j$; $R = R - \boldsymbol{f_j}$;

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

having predefined cardinality of the set.

5. Decision Tree-based Multi-Class Support Vector Machine (SVM)

Different variants of Support Vector Machine has been widely used to classify biological data such as structure magnetics resonance image for Alzheimer [31], for preclinical diagnosis of brain-related diseases [32] and in many other areas like tea-category system [33].

Decision tree based classifiers decompose a large problem into many smaller sub-problems, and hence it is efficient in handling the massive problems. To

solve the multi-class problem using decision tree based classifiers, results from all these sub-problems are integrated. In decision tree-based classification method such as SVM Binary Decision Tree (SVM-BDT) [26], binary class classification has been extended to multi-class classification using different ensemble strategies. This classifier partitions the classes into two sets at each node. In SVM-BDT, there is a problem of determining the structure of the tree, i.e. how to partition the data into two groups, as shown in Figure 4 for multi-class classification. To handle this problem, clustering algorithms have been utilized for the hierarchical design of binary decision subtasks using SVMs in which Euclidean distance (ED) has been used as the class separability measure for creating two disjoint groups of patterns. The classification accuracy of the classifier depends on the clusters so generated. In the SVM-BDT, two disjoint groups g_1 and g_2 are formed by dividing the classes into two groups. The SVM classifier is trained at the root node of the decision tree using these two groups. The left and right subtrees of the decision tree consists of classes from the first and second clustering groups respectively. The process is repeated until only one class is left in each group as a leaf. However, the disadvantage of SVM-BDT is the higher time complexity associated with the clustering phase.

In the research work of [34], to separate one class from others, a hyperplane is determined in training phase. If the separated classes contained more than one class, the hyperplane is determined to separate the classes at the node that connected to top node. The training process is continued until data corresponding to only one class is left in the separated group. Thus the problem of the unclassifiable region can be solved in OvA SVM scheme. In

their method, the class separability is measured regarding Euclidean distance (ED) between class centres.

The ED measure, utilized in the building process of decision tree OvA and SVM-BDT, does not consider scatteredness of the given class pattern. Hence, it may not be the appropriate choice to measure class separability between two different classes of patterns. To determine better variability within a class, a statistical distance measure is utilized by pattern recognition community which constitutes a natural concept of measuring class separability. Among these statistical distance measures, entropy-based [35] statistical measure, known as information gain (IG) is utilized by [36, 37].

For a given attribute, reduction in measurement of impurity of the partitions set as compared to whole set of samples is obtained by this measure. Thus IG provides information regarding how a given attribute is related to whole system. IG can be given as:

$$IG\,(C|A) = H(C) - H(C|A) \tag{20}$$

where $IG(C|A)$ is information gain of class C for attribute A, $H(C)$ is entropy of the given data and $H(C|A)$ is the conditional entropy of the data for given attribute A. Conditional entropy can be measured as:

$$H(C|A) = \sum_{j=1}^{|A|} p(a_j)(-\sum_{i=1}^{n} p(C_i|a_j)\log p(C_i|a_j)) \tag{21}$$

The entropy of whole data is defined as:

$$H(C) = -\sum_{i=1}^{n} p(C_i)\log p(C_i) \tag{22}$$

where $p(C_i)$ is the probability of class i, $p(a_j)$ is the probability of value a_j of attribute, $|A|$ is the total number of different values attribute A can take

and n is the number of classes. IG can be generally used to determine the class separability between different groups of data points as the less overlap or more distance between two different groups of data points will have higher value of IG. The selection of optimal SVM model can be determined on the basis of higher values of IG which signifies better separability between two different patterns of data class. IG for a given independent binary SVM containing n_i elements of C_i and n_j elements of $C_{i\neq j}$, can be calculated as

$$IG(i) = H(C_i, C_{i\neq j}) - [p(C_i)H(t_p, f_p) + p(C_{i\neq j})H(t_n, f_n)] \tag{23}$$

where

$$H(x, y) = -\left(|x|\, log\frac{|x|}{|x|\,|y|} + |y| \log \frac{|y|}{|x|\,|y|}\right) \tag{24}$$

$p(C_i) = \frac{n_i}{n_i+n_j}$ and $p(C_{i\neq j}) = \frac{n_j}{n_i+n_j}$

t_p, f_p, t_n, and f_n stand for number of true positive, true negative, false positive and false negative samples in data respectively.

The outline of OvA ODT-SVM is given in algorithm 2.

6. Experimental Set-up and Result

6.1. Dataset

For the mental task classification, we have accessed publicly available data to carried out this experiment. [38], which has been summarized in Table 1 in terms of number trials of five different mental tasks performed by seven subjects age group between 20 to 48. To our best of knowledge, no other data available for the response of mental task type of BCI. We have utilized all trials of data and discarded Subject 4 because it is having some trails of

Algorithm 2: OvA ODT-SVM

Input: Whole Data;

Construct the initial list of n class data $C_1, C_2 \dots C_n$;

Calculate information for between class C_i and class C_j given $i \neq j$ and $i = 1 \dots n$;

Calculate $H(t_p, f_p)$, $H(f_n, t_n)$, $p(C_i)$ and $p(C_{(j \neq i)})$;

Compute information gain IG(i) for ith class data;

Identify model i which take maximum IG(i);

If $j \geq 2$ repeat steps $3 - 6$, otherwise terminate;

data. EEG signal was taken from six electrodes in our experiment, placed on the scalp at C3, C4, P3, P4, O1 and O2 referencing to two electrodes placed at electrically linked mastoid, A1, and A2, Each trial is of 10 second

Table 1: Data Description

Subject No	Tasks	Trials
1	Baseline(Relax)-**B**;Letter Composing-**L**; Visual Counting-**C**; Mathematics-**M**; Geometric Rotation-**R**	10
2	Do	5
3	Do	10
4	Do	10
5	Do	15
6	Do	10
7	Do	5

time duration recorded with a sampling frequency of 250 Hz, which resulted in 2500 samples points per trial. More detail about the data can be found in the work of [38].

The data of each task of each subject is disintegrated into half-seconds segments to construct its feature thus, yielding 20 segments (signal) per trial

for each subject.

The feature vector construction is done in two phase. Initially,in the first phase, three level decomposition of signal is experimented with help of wavelet transform using db1 mother wavelet. In the second phase, the signal is characterized as a combination of eight statistical or uncertainty parameters, obtained from each decomposed signal as shown in Table 2.

Table 2: Summery of Features

Number of Channels	Level of Wavelet Decomposition	Number of Coefficients	Number of Extracted Parameters	Total Number of Features
6	3	4 (D1, D2, D3, A3)	8 (Root mean square, Variance, Skewness, Kurtosis, Lempel-Ziv Complexity Central & Maximum Frequency, Shannon Entropy)	$6 \times 4 \times 8 = 192$

6.2. Result

Figure 5 depicts variation of statistical quantities of detailed coefficients D1 among five mental tasks obtained using WT for Channel 1 It can be noted from Figure 5 that there are few features whose values are significantly different for different metal tasks and thus help to distinguish different mental tasks. For some features, there is not any variation in values for different mental tasks. Hence, such features may not be suitable to distinguish two different mental tasks. Similar observations are also be noted for other components and other channels. Thus, to select the relevant and non-redundant features, the linear regression feature selection method is employed using forward feature selection approach. To build decision model, ODT SVM is

used with Gaussian Kernel function. To find optimal choice of regularization constant C and gamma, Grid search method is applied. The performance of the proposed model regarding classification accuracy is compared with classification accuracy achieved with the model learned without feature selection. The average classification accuracy of 10 runs of 10 cross-validations is quoted. In order to validate the efficacy of the proposed method,three type of multi-mental task classification problems viz. three class, four class, and five class have formulated.

6.3. Three Class Problem

Here, we have framed three-class mental tasks problems by simultaneous opting the three different mental tasks from known five mental tasks. There are ten different triplet mental task combinations for forming three class problem given as BCL, BCM, BCR, BLM, BLR, BMR, CLM, CLR, CMR and LMR.

6.4. Four Class Problem

Building up to four mental task classification problems is achieved by opting four tasks simultaneously from the available five mental tasks. There are five different four class mental task problems namely BCLM, BCLR, BCMR, BLMR, and CLMR.

6.5. Five Class Problem

For the building of the five mental task classification problem, we have chosen all five mental tasks in a instant. Thus there is the five-class mental tasks classification problem as BCLMR.

Table 3: Comparison of classification accuracy for combination of WT method, (without and with) LR feature selection method and ODT_SVM, for all six Subjects for ten different triplet mental tasks (3 class problem).

Task	Subject 1		Subject 2		Subject 3		Subject 5		Subject 6		Subject 7	
	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR
BCL	0.65	0.88	0.62	0.88	0.62	0.78	0.52	0.75	0.63	0.75	0.72	0.89
BCM	0.84	0.95	0.74	0.89	0.51	0.72	0.48	0.72	0.81	0.95	0.88	0.96
BCR	0.77	0.93	0.82	0.98	0.52	0.69	0.57	0.76	0.76	0.88	0.90	0.95
BLM	0.82	0.93	0.69	0.83	0.61	0.82	0.53	0.75	0.69	0.83	0.79	0.93
BLR	0.81	0.93	0.82	0.93	0.55	0.71	0.61	0.82	0.66	0.78	0.82	0.96
BMR	0.93	0.99	0.77	0.93	0.52	0.67	0.63	0.79	0.76	0.86	0.77	0.94
CLM	0.78	0.98	0.80	0.91	0.59	0.77	0.45	0.74	0.76	0.88	0.92	1.00
CLR	0.67	0.86	0.87	0.93	0.57	0.67	0.58	0.83	0.73	0.83	0.93	1.00
CMR	0.85	0.94	0.84	0.96	0.43	0.66	0.59	0.78	0.8	0.9	0.82	0.95
LMR	0.91	0.95	0.89	0.99	0.53	0.70	0.63	0.81	0.72	0.84	0.80	0.98

Table 4: Comparison of classification accuracy for combination of WT method, (without and with) LR feature selection method and ODT_ SVM, for all six Subjects for five different quadruplet mental tasks (4 class problem).

Tasks	Subject 1		Subject 2		Subject 3		Subject 5		Subject 6		Subject 7	
	Without LR	with LR	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR
BCLM	0.70	0.84	0.62	0.77	0.48	0.71	0.40	0.64	0.63	0.77	0.73	0.87
BCLR	0.62	0.85	0.66	0.87	0.47	0.66	0.48	0.68	0.59	0.73	0.72	0.88
BCMR	0.73	0.91	0.71	0.91	0.39	0.60	0.48	0.69	0.68	0.83	0.72	0.88
BLMR	0.81	0.92	0.70	0.85	0.44	0.65	0.51	0.68	0.58	0.73	0.66	0.88
CLMR	0.71	0.84	0.74	0.91	0.44	0.62	0.44	0.71	0.65	0.79	0.76	0.96

Table 3, Table 4, and Table 5 show comparison of classification accuracy of the proposed method with and without feature selection for three class, four class, and five class mental task classification respectively. From these tables, we can observe the following:

1. The classification accuracy varies from subject to subject, for all the three types of multi-class mental task classification.
2. There is an improvement in classification accuracy with the use of feature selection in all multi-class mental task (3-class, 4-class, and 5-class) classification for all subjects.

Table 5: Comparison of classification accuracy for combination of WT method, (without and with) LR feature selection method and ODT_ SVM, for all six Subjects for a combination of five mental tasks (5 class problem).

Tasks	Subject 1		Subject 2		Subject 3		Subject 5		Subject 6		Subject 7	
	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR	Without LR	With LR
BCLMR	0.64	0.79	0.63	0.82	0.39	0.67	0.38	0.63	0.54	0.80	0.63	0.83

3. As the number of classes participating in data increases, the classification accuracy decreases.

From figures 6 to 8 demonstrate computational time for various types of multi mental tasks problem. It can be observed from these figures computational time also decreases after applying feature selection algorithm.

6.6. Comparison with the existing methods

While combiningTable 6, Table 7 and Table 8, it shows comparison of the proposed method with existing methods by [4] and [5]. In these tables A, B and C are the schemes used by [4] based on asymmetry ratio for the expression of different number of frequency band powers using 75-dimensional, 90-dimensional and 42-dimensional feature vector respectively whereas in [5] a multi-kernel SVM has been used for the classification. Also, in Table 8, comparison of the proposed method is shown with three power spectral density methods namely Wiener-Khinchine (WK) with Parzen smoothing window (A1), WK with Tuky window smoothing (B1) and 6^{th} order auto-regressive model (C1), used for feature extraction method and Fuzzy ARTMAP as a classifier [3].

Following observations can be drawn from these tables:

1. It can be noted from Table 6 that the proposed method with feature se-

Table 6: Comparison of the proposed model with existing three methods for three class mental task classification.

Three class mental task classification results									
	The proposed model		Li, et al., 2014	Zhang , et al.,2010			Palaniappan, et al., 2002		
Subjects	Without LR	With LR		A	B	C	A1	B1	C1
Subject1	0.80	0.93	0.74	0.64	0.75	0.71	0.8	0.75	0.82
Subject2	0.79	0.92	0.84	0.47	0.54	0.48	0.74	0.73	0.81
Subject3	0.55	0.72	0.81	0.54	0.59	0.57	0.85	0.84	0.86
Subject5	0.56	0.77	0.80	-	-	-	-	-	-
Subject6	0.73	0.85	0.87	-	-	-	-	-	-
Subject7	0.84	0.96	0.78	-	-	-	-	-	-
Average	0.71	**0.86**	0.81	0.55	0.63	0.59	0.8	0.77	0.83

Table 7: Comparison of the proposed model with existing two methods for four class mental task classification.

Four Class mental task classification results						
	The proposed method		Li, et al., 2014	Zhang , et al.2010		
Subjects	Without LR	With LR		A	B	C
Suject1	0.71	0.87	0.73	0.54	0.67	0.61
Suject2	0.69	0.86	0.78	0.38	0.45	0.38
Suject3	0.44	0.65	0.69	0.45	0.52	0.50
Suject5	0.46	0.68	0.79	-	-	-
Suject6	0.63	0.77	0.71	-	-	-
Suject7	0.72	0.89	0.79	-	-	-
Average	0.61	**0.79**	0.75	0.46	0.55	0.49

Table 8: Comparison of the proposed model with existing two methods for five class mental task classification.

Five class mental task classification results						
	The proposed method		Li, et al., 2014	Zhang , et al.,2010		
Subjects	Without LR	With LR		A	B	C
Subject 1	0.64	**0.79**	0.66	0.48	0.6	0.54
Subject 2	0.63	**0.82**	0.72	0.32	0.4	0.39
Subject 3	0.39	0.67	**0.75**	0.39	0.46	0.44
Subject 5	0.38	0.63	**0.68**	-	-	-
Subject 6	0.54	0.80	**0.85**	-	-	-
Subject 7	0.63	**0.83**	0.75	-	-	-
Average	0.54	**0.76**	0.74	0.39	0.49	0.46

lection performs better concerning average classification accuracy over all subjects and all 10 three mental tasks than all the existing methods by [3, 4, 5]. It can also be observed that the classification accuracy obtained with the proposed method without feature selection(without LR) also outperforms the method suggested by [4].

2. Similarly, for the 4 class problem from Table 7, it can be noted that the proposed method performs better than both the existing methods except for Subject 3 and 5 of [5]. Also, the average classification accuracy obtained with the proposed method without feature selection also outperforms the method suggested by [4].
3. In the comparative results of the five class problem in Table 8, the proposed method performs better for subject 1, 2 and 7, as compared

to the results of [5]. The average classification accuracy of all subjects of the proposed method is better than [5]. Both the with and without feature selection methods(columns SVM and LR_SVM) show improved classification accuracy as compared to the method of [4].

We have also applied a two way analysi by rank [39] and non-parametric statistical test known as Friedman test [40]is shown in table 9. The null hypothesis H_o was that all algorithms perform well. H_o was rejected at significant level α=.05. For k algorithms, each algorithm associates rank rang 1 to k, 1 denotes best and k depicts worst. It endorses our findings.

Table 9: Ranking by Friedman

Method	Mean Ranking
LR_ODT_SVM	1.3
ODT_SVM	4
[5]	1.7
[4] A	4.7
[4] B	3.7
[4] C	5.7

7. Conclusion

The inherent properties of EEG signal, i.e., a small amplitude which is not helpful in distinguishing different mental tasks makes the multi-class classification for BCI a challenging problem. In the proposed work, the multi-class classification for the mental task in BCI is proposed using EEG signals. In the proposed method, two-phase feature extraction is proposed. In the

first phase, the wavelet transform is applied to extract approximate and detailed coefficients, while in the second phase, statistical measures are used to represent the decomposed signal more compactly. A set of relevant and non-redundant features is obtained using LR. Then, Optimal decision tree based SVM is used as a multi-class classifier to build decision model. Experiments are performed on publicly available dataset [38] which contains EEG signal for five mental task classification. The performance of the proposed approach is evaluated for 3-class, 4-class, and 5-class mental task classification. Experimental results are compared with existing methods. It is observed that the proposed method provides better classification accuracy in comparison to the existing methods for 3-class, 4-class, and 5-class mental task classification. It is also observed that the classification accuracy improves with the use of feature selection. The proposed method may be helpful in developing BCI devices for multi-class classification.

The proposed framework for multi mental task classification have utilized wavelet transform to decompose the EEG signal. The major drawback of the wavelet transform is that it uses some fixed wavelet function, independent of the signal to be process. In future work, we would like to explore signal adaptive decomposition techniques such as empirical mode decomposition and its variants, and many more. The parametric feature extraction model of the proposed framework uses only statistical, uncertainty, frequency contents and complexity parameters. Moreover, signal may has some memory based property also, therefore we we would like to explore some memory related and some dynamics parameters. Since, the utilized dataset only five class mental activities, as number of mental activities will increase, there would

be highly imbalance class problem for one versus rest. In future, we would like to explore hybridization of one versus one and one versus rest i.e. one versus one versus approach for classification model.

Acknowledgement

Authors of this manuscript express their gratitude to the Council of Scientific and Industrial Research (CSIR), India, for providing financial support to support research work.

References

[1] J. P. Donoghue, Connecting cortex to machines: recent advances in brain interfaces, nature neuroscience 5 (2002) 1085–1088.

[2] D. Wang, D. Miao, G. Blohm, Multi-class motor imagery eeg decoding for brain-computer interfaces, Frontiers in neuroscience 6.

[3] R. Palaniappan, R. Paramesran, S. Nishida, N. Saiwaki, A new brain-computer interface design using fuzzy artmap, Neural Systems and Rehabilitation Engineering, IEEE Transactions on 10 (3) (2002) 140–148.

[4] L. Zhang, W. He, C. He, P. Wang, Improving mental task classification by adding high frequency band information, Journal of medical systems 34 (1) (2010) 51–60.

[5] X. Li, X. Chen, Y. Yan, W. Wei, Z. J. Wang, Classification of eeg signals using a multiple kernel learning support vector machine, Sensors 14 (7) (2014) 12784–12802.

[6] M. Jochumsen, I. Khan Niazi, M. Samran Navid, M. Nabeel Anwar, D. Farina, K. Dremstrup, Online multi-class brain-computer interface for detection and classification of lower limb movement intentions and kinetics for stroke rehabilitation, Brain-Computer Interfaces 2 (4) (2015) 202–210.

[7] L. Tonin, M. Pitteri, R. Leeb, H. Zhang, E. Menegatti, F. Piccione, J. d. R. Millán, Behavioral and cortical effects during attention driven brain-computer interface operations in spatial neglect: A feasibility case study, Frontiers in human neuroscience 11 (2017) 336.

[8] A. Gupta, R. Agrawal, B. Kaur, A three phase approach for mental task classification using eeg, in: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ACM, 2012, pp. 898–904.

[9] A. Gupta, R. Agrawal, Relevant feature selection from eeg signal for mental task classification, in: Advances in Knowledge Discovery and Data Mining, Springer, 2012, pp. 431–442.

[10] A. Gupta, R. Agrawal, B. Kaur, Performance enhancement of mental task classification using eeg signal: a study of multivariate feature selection methods, Soft Computing 19 (10) (2015) 2799–2812.

[11] I. Koprinska, Feature selection for brain-computer interfaces, in: New Frontiers in Applied Data Mining, Springer, 2010, pp. 106–117.

[12] P. F. Diez, V. Mut, E. Laciar, A. Torres, E. Avila, Application of the empirical mode decomposition to the extraction of features from eeg

signals for mental task classification, in: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, IEEE, 2009, pp. 2579–2582.

[13] J. Müller-Gerking, G. Pfurtscheller, H. Flyvbjerg, Designing optimal spatial filters for single-trial eeg classification in a movement task, Clinical neurophysiology 110 (5) (1999) 787–798.

[14] H. Ramoser, J. Muller-Gerking, G. Pfurtscheller, Optimal spatial filtering of single trial eeg during imagined hand movement, Rehabilitation Engineering, IEEE Transactions on 8 (4) (2000) 441–446.

[15] J. S. Kirar, R. Agrawal, Composite kernel support vector machine based performance enhancement of brain computer interface in conjunction with spatial filter, Biomedical Signal Processing and Control 33 (2017) 151–160.

[16] J. Andreu-Perez, F. Cao, H. Hagras, G.-Z. Yang, A self-adaptive online brain machine interface of a humanoid robot through a general type-2 fuzzy inference system, IEEE Transactions on Fuzzy Systems.

[17] H. Wang, Y. Zhang, N. R. Waytowich, D. J. Krusienski, G. Zhou, J. Jin, X. Wang, A. Cichocki, Discriminative feature extraction via multivariate linear regression for ssvep-based bci, IEEE Transactions on Neural Systems and Rehabilitation Engineering 24 (5) (2016) 532–541.

[18] N. Robinson, C. Guan, A. Vinod, K. K. Ang, K. P. Tee, Multi-class eeg classification of voluntary hand movement directions, Journal of neural engineering 10 (5) (2013) 056018.

[19] J. Kihoro, R. Otieno, C. Wafula, Seasonal time series forecasting: a comparative study of arima and ann models, AJST 5 (2).

[20] G. Zhang, B. E. Patuwo, M. Y. Hu, Forecasting with artificial neural networks:: The state of the art, International journal of forecasting 14 (1) (1998) 35–62.

[21] R. Rifkin, A. Klautau, In defense of one-vs-all classification, The Journal of Machine Learning Research 5 (2004) 101–141.

[22] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, Pattern Recognition 44 (8) (2011) 1761–1776.

[23] T. Hastie, R. Tibshirani, et al., Classification by pairwise coupling, The annals of statistics 26 (2) (1998) 451–471.

[24] J. C. Huhn, E. Hullermeier, Fr3: a fuzzy rule learner for inducing reliable classifiers, Fuzzy Systems, IEEE Transactions on 17 (1) (2009) 138–149.

[25] E. Hüllermeier, K. Brinker, Learning valued preference structures for solving classification problems, Fuzzy Sets and Systems 159 (18) (2008) 2337–2352.

[26] G. Madzarov, D. Gjorgjevikj, I. Chorbev, et al., A multi-class svm classifier utilizing binary decision tree., Informatica (Slovenia) 33 (2) (2009) 225–233.

[27] R. C. Guido, A note on a practical relationship between filter coefficients and scaling and wavelet functions of discrete wavelet transforms, Applied Mathematics Letters 24 (7) (2011) 1257–1259.

[28] A. Abbate, P. K. Das, C. M. DeCusatis, Wavelets and subbands, Springer, 2002.

[29] A. Lempel, J. Ziv, On the complexity of finite sequences, Information Theory, IEEE Transactions on 22 (1) (1976) 75–81.

[30] H.-S. Park, S.-H. Yoo, S.-B. Cho, Forward selection method with regression analysis for optimal gene selection in cancer classification, International Journal of Computer Mathematics 84 (5) (2007) 653–667.

[31] Y. Zhang, S. Wang, Z. Dong, Classification of alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree, Progress In Electromagnetics Research 144 (2014) 171–184.

[32] Y. Zhang, Z. Dong, S. Wang, G. Ji, J. Yang, Preclinical diagnosis of magnetic resonance (mr) brain images via discrete wavelet packet transform with tsallis entropy and generalized eigenvalue proximal support vector machine (gepsvm), Entropy 17 (4) (2015) 1795–1813.

[33] S. Wang, X. Yang, Y. Zhang, P. Phillips, J. Yang, T.-F. Yuan, Identification of green, oolong and black teas in china via wavelet packet entropy and fuzzy support vector machine, Entropy 17 (10) (2015) 6663–6682.

[34] F. Takahashi, S. Abe, Decision-tree-based multiclass support vector machines, in: Neural Information Processing, 2002. ICONIP'02. Proceed-

ings of the 9th International Conference on, Vol. 3, IEEE, 2002, pp. 1418–1422.

[35] C. E. Shannon, W. Weaver, The mathematical theory of communication (urbana, il (1949).

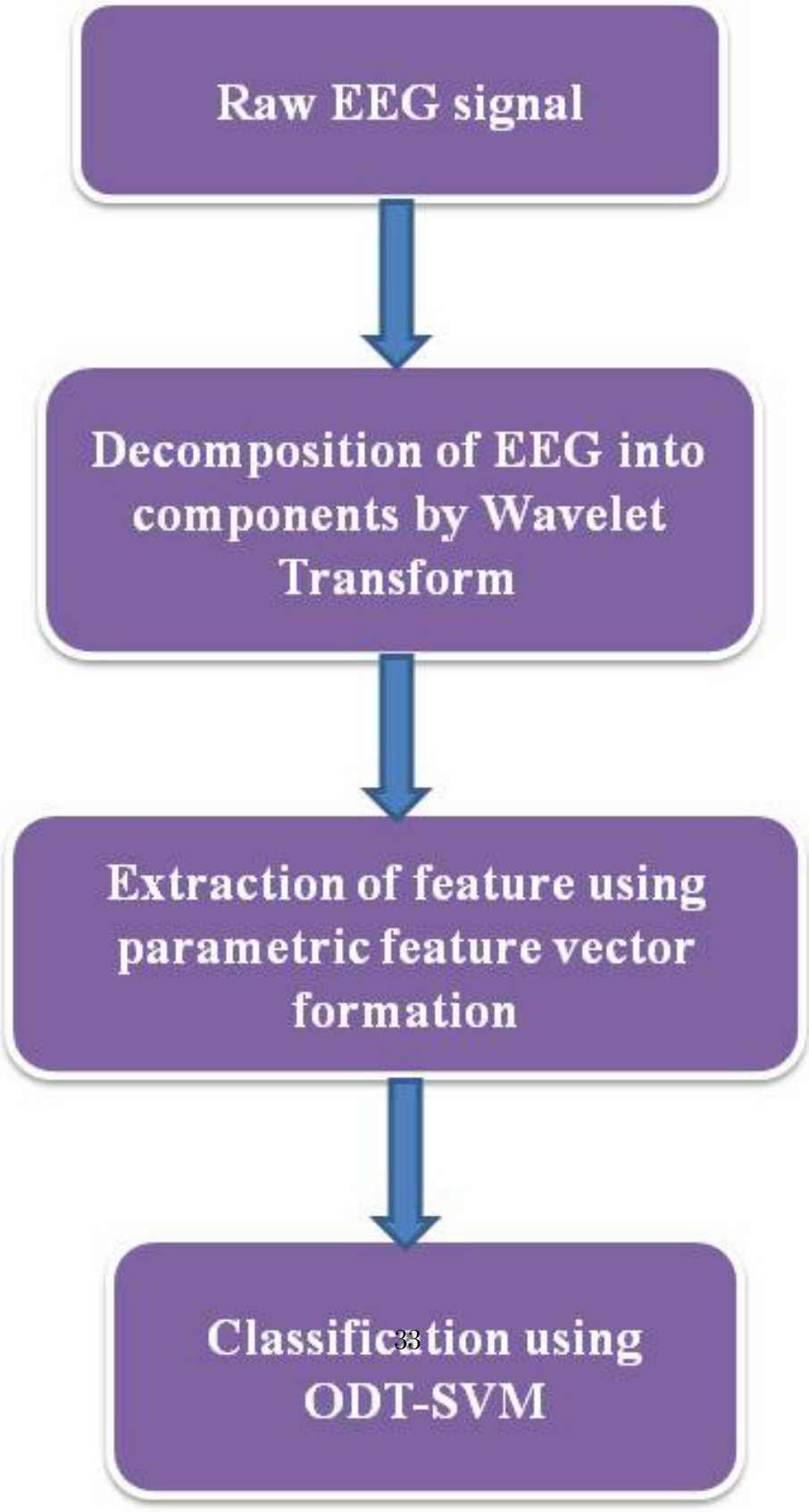
[36] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.

[37] M. Bala, R. Agrawal, Optimal decision tree based multi-class support vector machine, Informatica: An International Journal of Computing and Informatics 35 (2) (2011) 197–209.

[38] Z. A. Keirn, J. I. Aunon, A new mode of communication between man and his surroundings, Biomedical Engineering, IEEE Transactions on 37 (12) (1990) 1209–1214.

[39] J. Derrac, S. García, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm and Evolutionary Computation 1 (1) (2011) 3–18.

[40] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (200) (1937) 675–701.


Raw EEG signal
Decomposition of EEG into components by Wavelet Transform
Extraction of feature using parametric feature vector formation
Classification using ODT-SVM

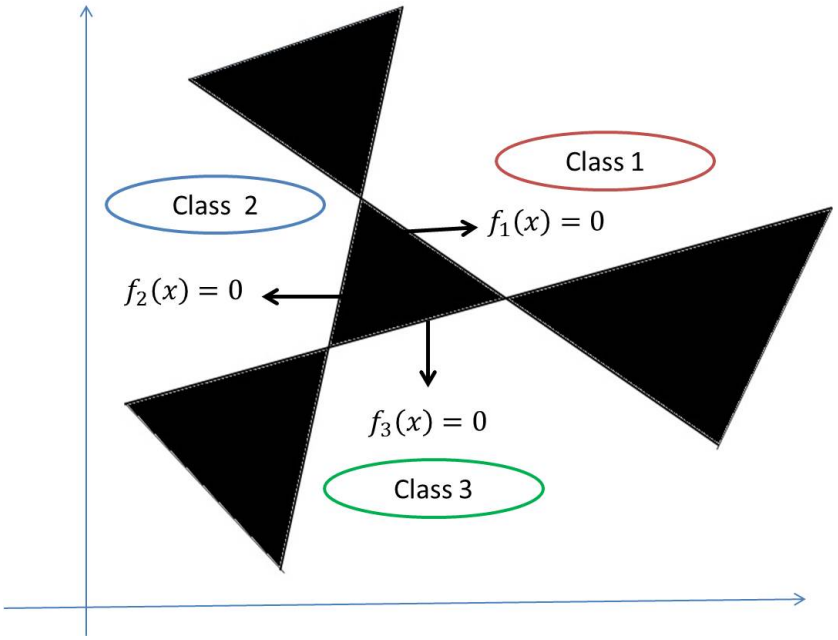


Figure 2: The existence of unclassifiable regions (in black).

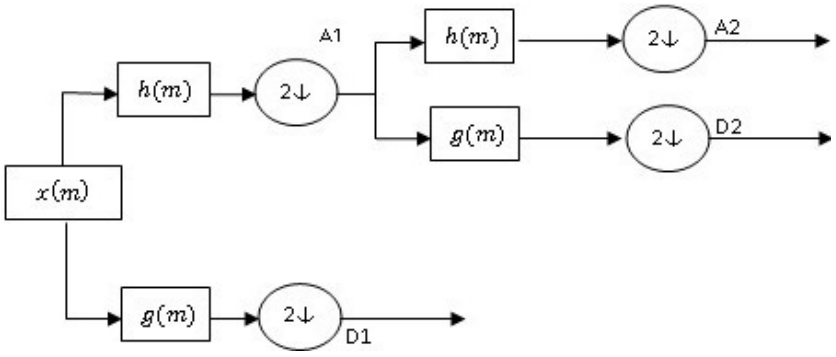


Figure 3: 2-level Decomposition of a signal into approximation and detail components

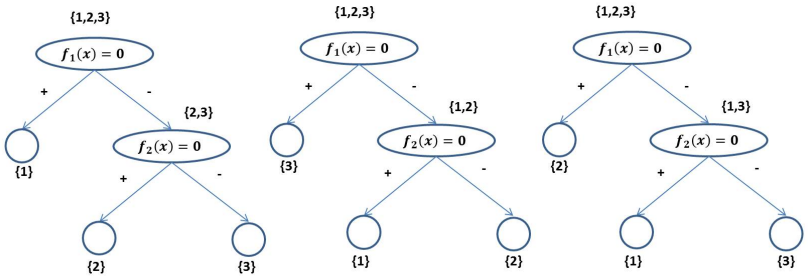


Figure 4: The three possible binary decision tree for three class problem in OvA scheme.

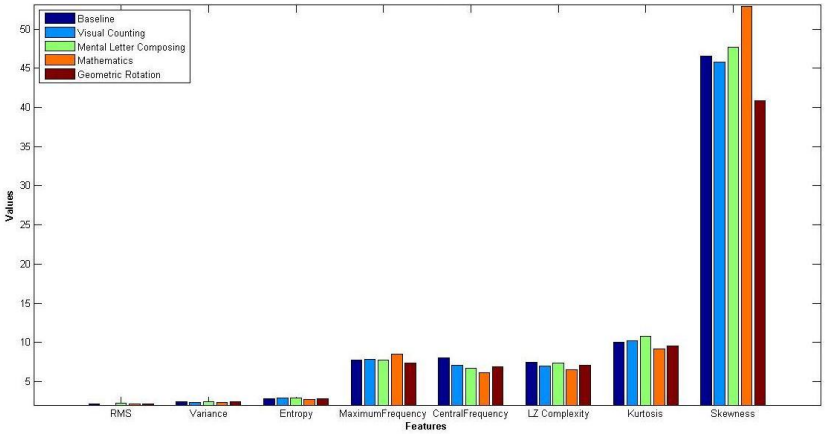


Figure 5: Variation of statistical quantities of detailed coefficients D1 among five mental tasks obtained using WT for Channel 1.

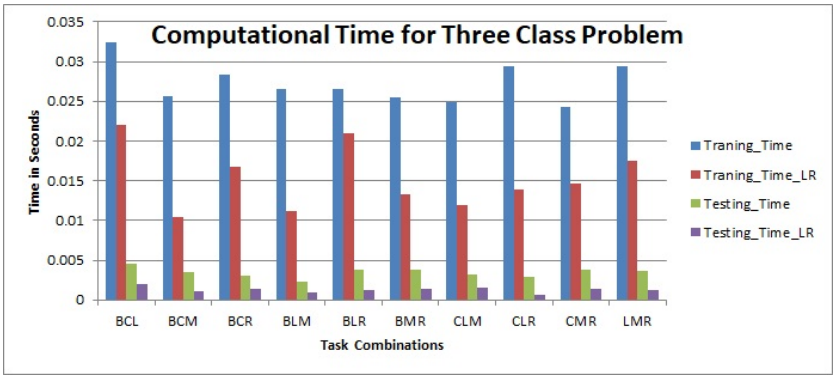


Figure 6: Computational Time for Three Class Problem.

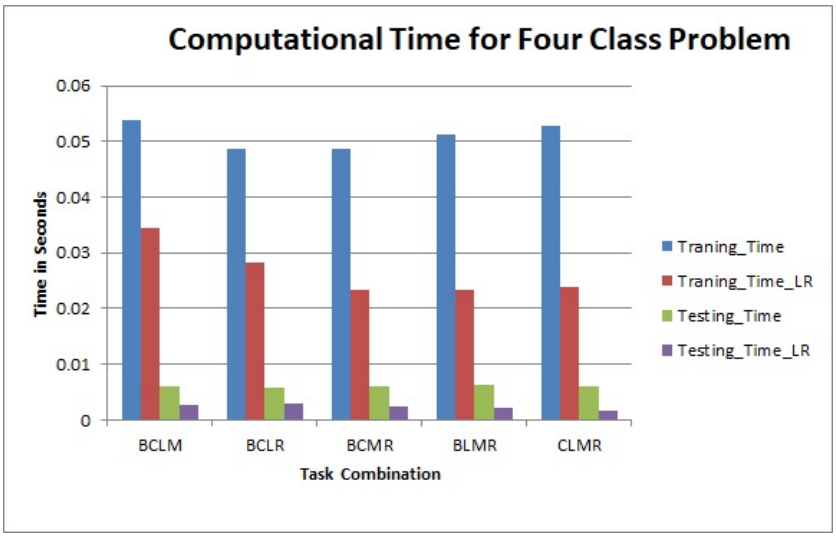


Figure 7: Computational Time for Four Class Problem.

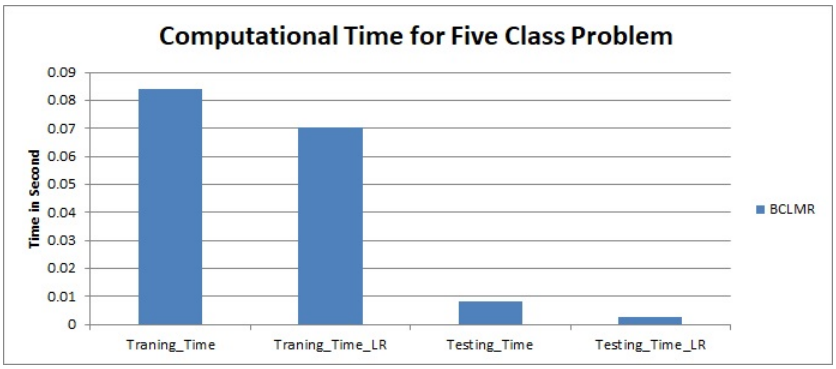


Figure 8: Computational Time for Five-Class Problem.