On the Utility of Power Spectral Techniques With Feature Selection Techniques for Effective Mental Task Classification in Noninvasive BCI

Akshansh Gupta, Ramesh Kumar Agrawal, Jyoti Singh Kirar, Javier Andreu-Perez, Wei-Ping Ding, Chin-Teng Lin, Fellow, IEEE, and Mukesh Prasad

Abstract—In this paper, classification of mental task-root brain–computer interfaces (BCIs) is being investigated. The mental tasks are dominant area of investigations in BCI, which utmost interest as these system can be augmented life of people having severe disabilities. The performance of BCI model primarily depends on the construction of features from brain, electroencephalography (EEG), signal, and the size of feature vector, which are obtained through multiple channels. The availability of training samples to features are minimal for mental task classification. The feature selection is used to increase the ratio for the mental task classification by getting rid of irrelevant and superfluous features. This paper suggests an approach to augment the performance of a learning algorithm for the mental task classification on the utility of power spectral density (PSD) using feature selection. This paper also deals a comparative analysis of multivariate and univariate feature selection for mental task classification. After applying the above stated method, the findings demonstrate substantial improvements in the performance of learning model for mental task classification. Moreover, the efficacy of the proposed approach is endorsed by carrying out a robust ranking algorithm and Friedman’s statistical test for finding the best combinations and compare various combinations of PSD and feature selection methods.

Index Terms—Brain–computer interface (BCI), feature extraction, feature selection, mental tasks classification, power spectral density (PSD).

I. INTRODUCTION

A BRAIN–COMPUTER interface (BCI) [1], [2] is a message transmission framework, through which an individual can communicate for necessities by his or her brain signals, even absence of normal pathway of the computer system and a very effective device for the person with severe motor impairment [3], [4]. It is a pragmatic area, which has focused to the design and invent of neuron-rooted means to endue solutions for disease prediction, communication, and control [5]–[7]. On the ground of acquisition of the brain signal, BCI is broadly divided in three categories in [8] and [9], viz, invasive, semi-invasive [electrocorticography (ECoG)], and noninvasive [electroencephalography (EEG)]. Economically nature [10] and calibre to capture brain signals in a noninvasive fashion, EEG is a mostly preferred technique to acquire brain activity for BCI systems [7], [11]. BCI systems can be used as a Response to mental tasks system [12], which is perceived to be more practical for locomotive patients. The basic assumption of this type of system is that mental activities lead to produce task-originated patterns. The BCI system’s success depends on the precision of classification assorted mental tasks. These tasks require extractions of discriminative features from the raw EEG signal to distinguish different mental tasks [13].

In previous studies, the researchers have utilized plenty approaches of feature extraction to better model of the EEG signal for the classification process in the BCI domain, for example, band power [14], amplitude values of EEG signals [15], power spectral density (PSD) [16]–[19], autoregressive (AR) and adaptive AR (AAR) parameters [20], and time-frequency and inverse model-based features [21]–[23]. Wavelet transform (WT) [24], [25] and empirical mode decomposition (EMD) [26]–[32] have been used to decompose non-stationary and nonlinear EEG signals into smaller frequency components. However, both WT and EMD methods provide low-frequency resolution and may not handle efficiently different overlapping frequency bands [33], [34] present in the EEG. On the other hand, power spectral analysis provides high-frequency resolution. The recording of EEG data occurs from multiple sensors/channels. Hence, the EEG data contains huge number of features but the recording session of the person is usually very small in number. That produces a small number of data samples. Hence, it suffers the curse of dimensionality as the ratio of features and sample is very small [35].
To conquer the situation, reduction of the dimension using feature selection is suggested in [36]. In spite of that no in-depth study has ever been conducted about how to use power spectral features effectively with combination feature selection techniques in BCI the applications.

The contributions of this paper provide answers to the following questions.

1) Whether extraction of features using power spectral techniques helps in mental task classification.

2) Whether further reduction in dimensionality of features using feature selection approaches improves the classification performance or not.

3) Is multivariate feature selection approach better than univariate feature selection approach?

4) Which conjunction of feature extraction and selection method performs best for mental task classification?

Thus, this paper proposes a procedure of the determination of a compact collection of features from the EEG signal in the two-phase approach. The first phase elaborates about the extraction of PSD features from the EEG signal using three different approaches. In the second level, a set of vital features is sorted by filter feature selection approach, both multivariate and univariate. To investigate the performance of different combinations of PSD methods and feature selection methods, experiments are conducted on an open EEG data [7] source. In order to rank and compare multiple combinations of PSD and feature selection methods, Ranking method and Friedman’s statistical test were also performed.

The rest of this paper is organized as follows. The power spectral estimation approach has been discussed briefly in Section II. The proposed approach to obtain minimal subset of relevant and nonredundant of the PSD features using multivariate feature selection methods is included in Section III. The descriptive information, data, and method results are presented in Section IV and finally conclusions and future work are discussed in Section V.

II. Feature Extraction Using Power Spectral Density

The PSD is a calculation of an average power associated with any random sequence [37], which can be catalogued into three categories: 1) nonparametric; 2) parametric; and 3) subspace. The nonparametric methods are robust and simple to compute. Periodogram-based estimation, Bartlett window, Welch window, and Blackman and Tuckey method are examples of nonparametric category. However, they do not provide the necessary frequency resolution due to their inability to extrapolate the finite length sequence for data points exceeding the signal length. Another, drawback of this approach is spectral leakage [38]. To overcome the drawback of nonparametric methods, parametric method is suggested. The estimation of PSDs values from a given signal in parametric approaches are carried out by assuming that output of the linear system is driven by white noise and then parameters of the system are calculated. Examples are the Yule-Walker AR method [39], the Burg method [16], covariance and modified covariance, etc. The commonly used parametric linear system model is the all pole model, which consists of a filter with all zeros at the origin and occurs in the z-plane. The output produced by such a filter using white noise as input is an AR process. Thus, these spectral estimation methods are also sometimes known as AR methods. The AR methods tend to aptly describe data spectrum that is “peaky,” the data having PSDs value large at certain frequencies, e.g., speech data. Smoother estimates of the PSD are produced by parametric methods than nonparametric methods, however, it is subject to error if the order of model chosen incorrectly. The subspace methods are often used when signal-to-noise ratio (SNR) is low. The PSDs values are obtained concerning eigen-decomposition of autocorrelation matrix. For line spectra or spectra having sinusoidal nature, the subspace methods are better choice and also effective in the recognition of sinusoids mixed in noise. However, the subspace methods suffer from the following: the method in all probability does not generate true PSD estimates; it does not store power which is required for processing between the time and frequency domains; and it flunks in getting back the autocorrelation series by computing the inverse Fourier transform of the frequency estimate.

For a given stationary random signal \( x_m \), the PSD \( P_{xx} \) is mathematically related to the autocorrelation sequence by the Fourier transform, which regarding normalized frequency \( f_s \) is given by

\[
P_{xx}(f) = \frac{1}{f_s} \sum_{m=-\infty}^{\infty} R_{xx}(m) e^{-j2\pi mf_s} \tag{1}
\]

where \( f_s \) denotes the sampling frequency. The Fourier transform of the autocorrelation of the signal also gives the PSD. Using the inverse discrete-time Fourier transform from the PSD, the correlation sequence is derived as follows:

\[
R_{xx} = \int_{-\pi}^{\pi} P_{xx}(\omega) e^{-j\omega m} d\omega = \int_{-f_s/2}^{f_s/2} P_{xx}(f) e^{-j2\pi mf_s} df. \tag{2}
\]

The average power of the sequence \( x_n \) over the entire Nyquist interval is represented by

\[
R_{xx}(0) = \int_{-\pi}^{\pi} P_{xx}(\omega) d\omega = \int_{-f_s/2}^{f_s/2} P_{xx}(f) df. \tag{3}
\]

For a particular frequency band \([\omega_1, \omega_2]\), \((0 \leq \omega_1 \leq \omega_2 \leq \pi)\), the average power of a signal is given by

\[
P_{[\omega_1, \omega_1]} = \int_{\omega_1}^{\omega_2} P_{xx}(\omega) d\omega \tag{4}
\]

where \( P_{xx}(\omega) \) represents the power content of a signal in an extremely small frequency band, which known as the power spectral density.

A. Welch Method

The Welch method falls under nonparametric approach. For a finite-time duration random signal \( x_n \) of \( N \) interval length, PSD values are estimated with the help of a periodogram which is the squared modulus of the discrete Fourier transform of the signal and is given by

\[
P_{xx}(f) = \frac{1}{N} |X(f)|^2. \tag{5}
\]

Here, \( f \) corresponds to the frequency of the sequence and \( X(f) \) is the Fourier transform of the signal. A periodogram gives asymptotically nonbiased estimate of power spectrum.
In the Welch method, $N$ length signal is divided into $K$ overlapped segments each of length $M$. The $i$th segment is given by

$$x_i(n) = x(n + iD).$$  \tag{6}$$

Here, $n = 0, \ldots, N - 1$, $i = 0, \ldots, K - 1$ and $D$ is the overlap segment. For this, a windowed segment periodogram is given by

$$P_{XX}(f) = \frac{1}{MU} \sum_{i=0}^{N-1} w(n)x_i(n)e^{-j2\pi fn}$$ \tag{7}

where $w(n)$ and $U$ denote the window function and the power of the window function, respectively, and defined as follows:

$$U = \frac{1}{M} \sum_{n=0}^{M-1} w^2(n).$$ \tag{8}

The average of $K$ periodograms depicts the Welch power spectrum and is given by

$$P^W_{XX} = \frac{1}{K} \sum_{i=0}^{K-1} P_{XX}(f).$$ \tag{9}

B. Burg Method

The Burg method [37] is a parametric method of spectral analysis. The PSDs values can be obtained by finding $p$th order coefficients of an AR process. A $p$th order real-valued AR signal $x(n)$ (with zero mean) at point $n$ is given by [19]

$$x(n) = - \sum_{m=1}^{p} a_m x(n - m) + e(n)$$ \tag{10}

where $a_m$ and $e(n)$ represent the AR coefficient of $x(n - m)$ and the error term at point $n$ independent of past terms, respectively. The Burg algorithm test to find the AR coefficient by applying more data points and minimizes the both forward and backward prediction errors in the least squares sense [19], with the AR coefficients constrained to satisfy the Levinson–Durbin recursion. It provides high resolution for short data records. After finding AR coefficients by Burg algorithm, PSD value $S(f)$ at frequency $f$ is given by

$$S(f) = \frac{S_c(f)}{|1 + \sum_{j=1}^{p} a_j e^{-j2\pi fT}|^2}$$ \tag{11}

where $T$ $S_c(f)$ represent the sampling period and spectrum of error sequence which should be flat, i.e., independent of frequency, respectively. One of the foremost concern in AR modeling is the choice of order $p$. To determine $p$, several criteria, such as final prediction error (FPE) [40], minimum description length [41], Akaike information criterion (AIC) [42], and AR transfer function [43], have been proposed in the literature. Among all these criteria, AIC is the most commonly used and defined as follows:

$$\text{AIC}(p) = \ln \sigma^2_{ep} + \frac{2p}{n}$$ \tag{12}

where $\sigma^2_{ep}$ defines an estimated variance in the linear prediction error. From Table 1, it can be observed that AIC value is minimum for order 5 or 6. This paper adopts $p = 6$ for experiments, which also suggested by Kerin and Aunon [7].

C. Multiple Signal Classification

The multiple signal classification (MUSIC) is an orthogonal subspace decomposition method based on Pisarenko idea [44], which allows the estimation of low SNR frequency components. This method is used to lower the effect of noise in the analyzed signal and finds the optimal frequency resolution in a dynamic signal [45]. Subspace method assumes that any discrete-time signal $s[n]$ is representable in the form of $m$ complex sinusoids with a noise $p[n]$ such that

$$s[n] = \sum_{i=1}^{m} A_i e^{j2\pi f_i n} + p[n], \quad n = 0, 1, 2, \ldots, N - 1$$ \tag{13}

where $A_i = |A_i|e^{j\theta_i}$ is a magnitude of $i$th complex sinusoid; $m, N, f_i, \theta_i$ are the frequency signal dimension order, number of sample data, and frequency and phase of $i$th complex sinusoid, respectively.

The autocorrelation matrix $R$ of signal $s[n]$ is given by

$$R = \sum_{i=1}^{m} |A_i|^2 p(f_i) P(f_i) + \sigma^2 I$$ \tag{14}

where

$$p(f_i) = [1, e^{j2\pi f_i}, e^{j4\pi f_i}, \ldots, e^{j2\pi(N-1)f_i}]^T$$

and $\sigma^2, H$, and $I$ denote the variance of white noise signal, Hermitian transpose, and the identity matrix, respectively. Therefore, it can be observed that $R$ is a composition of sum of signal and noise autocorrelation matrices such that

$$R = R_s + \sigma^2 I.$$ \tag{15}

Pisarenko has noticed that variance of noise acts with the smallest eigenvalues of $R$. The orthogonality of the signal and noise subspace is given as

$$p(f_i)^H v(m + 1) = 0,$$ \tag{16}

where $v(m + 1)$ is the eigenvector of noise in matrix $R$ with dimension of $(m + 1) \times (m + 1)$. The estimation of PSD by Pisarenko is defined as

$$P_{\text{Pisarenko}} = \frac{1}{|p(f)|^2}.$$ \tag{17}

PSD estimation by MUSIC gives better performance than Pisarenko due to addition of averaging of extra noise eigenvectors ($k = m + 1, m + 2, \ldots, M$). The estimation of PSDs by MUSIC is given by

$$P_{\text{MUSIC}}(f) = \frac{1}{\sum_{k=m+1}^{M} |p(f)|^2}.$$ \tag{18}

where $p(f)^H v(k) = 0$ for $k = 1, \ldots, m$ using orthogonality of the signal and noise subspace. These PSD values have major

<table>
<thead>
<tr>
<th>Task</th>
<th>Order</th>
<th>$5$</th>
<th>$6$</th>
<th>$7$</th>
<th>$8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>-1.012</td>
<td>-1.0117</td>
<td>-1.0109</td>
<td>-1.0106</td>
</tr>
<tr>
<td>Count</td>
<td></td>
<td>-1.2841</td>
<td>-1.2851</td>
<td>-1.2847</td>
<td>-1.2842</td>
</tr>
<tr>
<td>Letter</td>
<td></td>
<td>-1.2574</td>
<td>-1.2595</td>
<td>-1.2589</td>
<td>-1.2585</td>
</tr>
<tr>
<td>Math</td>
<td></td>
<td>-1.2783</td>
<td>-1.2772</td>
<td>-1.2762</td>
<td>-1.2768</td>
</tr>
<tr>
<td>Rot</td>
<td></td>
<td>-1.177</td>
<td>-1.176</td>
<td>-1.175</td>
<td>-1.1758</td>
</tr>
</tbody>
</table>
peaks at the principal components only. The performance of MUSIC depends on the dimension of the autocorrelation matrix \((M \leq N)\).

### III. Proposed Feature Selection Approach

The number of PSD values obtained using one of the given three methods from multiple channels would be large. In general, the available number of training samples is relatively small, which leads to curse of dimensionality problem [35]. In order to subdue the curse of dimensionality problem, there is a need to determine a minimal set of pertinent features, which can improve classification accuracy of a learning system. This paper proposes an approach to find a minimal subset of relevant feature using multivariate feature selection methods.

Feature selection method [36], [46] is one of the widely used approaches to determine relevant features. In spite of many researches have been done in different areas with the feature selection, however, there is not much work carried out in the domain of mental task classification. The filter and the wrapper approaches are the two major approaches of feature selection techniques. In filter approach, the step of selecting optimal features set is considered as one of the preprocessing steps of just before applying any machine learning algorithm. The filter approach adopts only inherent properties of the features and does not consider any virtue of learning algorithm. Therefore, it may not select the optimal feature set for the learning algorithm. The wrapper approach [46] finds an optimal features subset, which is compatible with the given learning algorithm. In the wrapper approach, given classifier requires to be trained for each feature of set of all features separately, which is more computationally costly than filter approach.

The filter approach is further partitioned in two categories on the basis of the way of opting features [36], as univariate (single feature ranking) and multivariate (feature subset ranking). The univariate method utilizes a scoring function for measuring relevance of the feature and implementation is very simple. In BCI field, the authors [47]–[50] used univariate filter method, where the performance of learning model usually improves with the help of reduced set of relevant features obtained by the univariate feature selection method. However, the univariate filter method does not capture the correlation among the features. Therefore, there may be many redundant features in the subset of relevant feature which may take down the performance of learning model. The wrapper method [7], [51], [52] is applied to obtain a subset of nonredundant features for the mental task classification. Due to high-dimensionality of feature of EEG data, wrapper approach is not a feasible option for mental task classification as it will become more computationally expensive. Therefore, this paper applies both univariate and multivariate filter feature selection algorithms.

Let us assume, we have a data matrix \(X\) of \(m\) rows, and \(k+1\) columns, with data sample \(x_i, i = 1, 2, \ldots, m\); containing features set \(S = f_1, f_2, \ldots, f_k\) and class label \(C_1, C_2, \ldots, C_n\), where \(n \leq m\).

### A. Univariate Feature Selection

1) Pearson’s Correlation Coefficient: Pearson’s correlation coefficient (CORR) [53], [54] is employed to determine the linear relationship between two variables. The CORR of \(i\)th feature vector \((f_i)\) with the class label vector \((c)\) is given by

\[
\text{CORR}(f_i, c) = \frac{\text{cov}(f_i, c)}{\sigma_f \sigma_c} = \frac{E[(f_i - \mu_f)(c - \bar{c})]}{\sigma_f \sigma_c}
\]

where \(i = 1, 2, \ldots, k\) and \(\sigma_f\) and \(\sigma_c\) represent the standard deviations of feature vector \(f_i\) and \(c\), respectively. \(\text{cov}(f_i, c)\) represents the covariance between \(f_i\) and \(c\), \(\mu_f = (1/k)\sum_{i=1}^{k} X_f\) and \(\bar{c} = (1/k)\sum_{j=1}^{k} C\) are the mean of \(f_i\) and \(c\), respectively.

The range of CORR\((f_i, c)\) falls between \((-1, +1)\). The value nearby to \([1]\), depicts the stronger linear relation among the prescribed variables while zero value implies no correlation between the two variables.

2) Mutual Information: The mutual information (MI) is a feature ranking method on basis of the Shannon entropy, which determines the relationship between two variables. The MI of a feature vector \(i\) and the class vector \(c\) can be calculated as [55]

\[
I(f_i, c) = \sum P(f_i, c) \log \frac{P(f_i, c)}{P(f_i)P(c)}
\]

where \(P(f_i)\) and \(P(c)\), and \(P(f_i, c)\) are the marginal probability distribution functions for random variables \(f_i\), \(c\) and joint probability distribution, respectively. The most extreme estimation of MI demonstrates the higher reliance of the variable on the class label. The advantage of MI is that it can discover even the nonlinear dependency between the attribute and the relating class label vector \(c\).

3) Fisher Discriminant Ratio: The Fisher discriminant ratio (FDR) is an univariate filter feature selection technique, which depends on the statistical virtue of the attributes or features. The FDR \((f_i)\) for \(i\)th feature for two class \(C_1\) and \(C_2\) is given as

\[
\text{The FDR}(f_i) = \frac{\mu_{1(i)} - \mu_{2(i)}}{\sigma^2_{1(i)} + \sigma^2_{2(i)}}
\]

where \(\mu_{1(i)}\) and \(\sigma^2_{1(i)}\) are the mean and deviation of the data of class \(C_1\) for \(i\)th feature, respectively.

4) Wilcoxon’s Rank-Sum Test: The Wilcoxon’s rank-sum test [56] is a nonparametric statistical test, which accomplishes between data of two classes on the basis of median of the samples having no prior knowledge of probability distribution.

The statistical distinctness \(t(f_i)\) of feature \(f_i\) for known two classes \(C_1\) and \(C_2\) using Wilcoxon’s statistics is defined as [57]

\[
t(f_i) = \sum_{l=1}^{N_i} \sum_{m=1}^{N_j} \text{DF}(X_{li} - X_{mi}) \leq 0
\]

where \(N_i\) and \(N_j\) are the number of the data example in class \(C_1\) and \(C_2\), respectively. DF represents the logical discriminative mapping between two classes of data, which defines an estimation of 1 or 0 corresponding to true or false and \(X_{li}\) represents the expression values of \(i\)th feature for \(l\)th sample.
The value of \(t(f_i)\) lies between 0 and \((N_i \times N_j)\). The relevance of the feature is defined as
\[
R(t(f_i)) = \max(t(f_i), N_i \times N_j - t(f_i)).
\] (23)

**B. Multivariate Feature Selection**

The time efficient multivariate filter method selects a subset of features, which are relevant to the class label of data and independent from each other. Thus, it up dues the limitations of both univariate and wrapper approaches. Therefore, this paper opts the most widely utilized multivariate filter methods for dimensionality reduction, such as Bhattacharyya distance (BD) measure [58], ratio of scatter matrices [59], linear regression (LR) [60], and minimum redundancy–maximum relevance (mRMR) [61].

1) **Bhattacharyya Distance**: The BD is used for finding similarity between two continuous or discrete probability distribution. It is a special case of Chernoff distance that provides similarity overlap of the distribution. For multivariate normal probability distribution, Chernoff distance measure is defined as [62]
\[
J_c = \frac{1}{2} \beta (1 - \beta) (\mu_2 - \mu_1)^T (1 - \beta) \Sigma_1 + \beta \Sigma_2)^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \log \frac{[(1 - \beta) \Sigma_1 + \beta \Sigma_2]^2}{|\Sigma_1|^{-\beta}|\Sigma_2|^{1-\beta}}.
\] (24)
where \(\mu_i\) and \(\Sigma_i\) are the mean vector and covariance matrix for class \(C_i\), respectively, where \(i = 1, 2\).

When \(\beta\) is \((1/2)\) then this distance is called as BD [58], which is given as
\[
J_B = \frac{1}{8} (\mu_2 - \mu_1)^T (\mu_2 - \mu_1) + \frac{1}{2} \log \frac{(|\Sigma_1| + |\Sigma_2|)^2}{|\Sigma_1|^{|\frac{1}{2}| |\Sigma_2|^{|\frac{1}{2}}}}.
\] (25)

2) **Ratio of Scatter Matrices**: The trace of ratio of scatter matrices [scatter ratio (SR)] is a measure of separability. As the trace of a scatter matrix is equal to the sum of the eigenvalues, which indicates the total variance in the data. The total variance defines how well features cluster around their class mean and how well they separate the class means. The scatter matrices, within-class scatter matrices \(S_w\) and between-class scatter matrices \(S_b\), are defined as
\[
S_w = \sum_{i=1}^{c} P_i E[(x - \mu_i)^T (x - \mu_i)]
\] (26)
\[
S_b = \sum_{i=1}^{c} P_i (\mu_i - \mu_0)^T (\mu_i - \mu_0)
\] (27)
where \(\mu_i, P_i\), and \(\mu_0\) are mean vector of \(i\)th class data, prior probability of \(i\)th class data, and global mean of data samples, respectively.

From the definitions of scatter matrices, the criterion value which has to be maximized is given as
\[
J_{SR} = \frac{\text{trace}(S_b)}{\text{trace}(S_w)}.
\] (28)

When intracluster distance is very small and the intercluster distance is very large, then \(J_{SR}\) takes the high value. The main advantage of this criterion is that it is not subject to any external parameters and assumptions of any probability density function. Also, the measure \(J_{SR}\) under linear transformation has the advantage of being invariant under linear transformation.

3) **Linear Regression**: The LR is a statistical approach, which determines casual link of an independent variable upon a dependent variable. The class label of the data is recognized as the target dependent variable and the feature that affect the target, which known as independent variable. There may be many features, which can affect the class of the data, therefore, in such case multiple regression analysis is more appropriate. A multiple regression model with \(k\) independent features \(f_1, f_2, \ldots, f_k\) and a class variable \(y\) is defined as [60]
\[
y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \xi_i, i = 1, 2, \ldots, n
\] (29)
where \(\beta_0, \beta_1, \ldots, \beta_k\) defines a set of fixed values calculated by the class label \(y\) and observed values of \(X\) and \(\xi_i\) is the error term. The sum of squared error (SSE) is given by
\[
\text{SSE} = \sum_{i=1}^{n} (y_i - y_i^p)^2
\] (30)
where \(y_i\) and \(y_i^p\) are observed and predicated values, respectively. The lower value of SSE depicts preferable regression
model. The total sum of squares (SSTOs) is calculated as
\[
SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2
\] (31)
where \(\bar{y}\) defines the mean value of \(y_i, i = 1, 2, \ldots, n\). The criterion function \(J_{LR}\) is given as
\[
J_{LR} = R^2 = 1 - \frac{SSE}{SSTO}. \tag{32}
\]
The value of \(J_{LR}\) lies between 0 and 1. The higher value of \(J_{LR}\) is selected for feature.

4) Minimum Redundancy–Maximum Relevance: The mRMR [60] is based on MI to discover a subset of features, which have minimum redundancy among themselves and maximum relevance with the class labels. The mRMR uses MI \(I(f_i, f_l)\) as a measure of similarity between two feature vectors \(f_i\) and \(f_l\) is given as
\[
I(f_i, f_l) = \sum_{k,l} p(f_k, f_l) \log \left( \frac{p(f_k, f_l)}{p(f_k)p(f_l)} \right) \tag{33}
\]
where \(p(f_k)\) and \(p(f_l)\) are the marginal probabilities of \(k\)th and \(l\)th features, respectively, and \(p(f_k, f_l)\) is the selected joint probability density. The relevance between the set of features \(S\) and
the target class label vector \( c \), denoted by REL, is expressed as

\[
\text{REL} = \frac{1}{|S|} \sum_{f_i \in S} I(f_i, c). \tag{34}
\]

The average redundancy among features in the set \( S \), denoted by RED, is defined as

\[
\text{RED} = \frac{1}{|S|^2} \sum_{f_i, f_l \in S} I(f_i, f_l). \tag{35}
\]

where \( S \) denotes the subset of features and \( |S| \) denotes the number of features in set \( S \). Minimum redundancy and maximum relevance is measured by

\[
J_{\text{MID}} = \max(f_i) [\text{REL} - \text{RED}]
\]

\[
= \max(f_i) \left[ \frac{1}{|S|} \sum_{f_i \in S} I(f_i, c) - \frac{1}{|S|^2} \sum_{f_i, f_l \in S} I(f_i, f_l) \right]. \tag{36}
\]

Clearly, the maximum values of \( J_{\text{MID}} \) can be achieved with minimum redundancy among features and maximum relevance with target vector.

IV. RESULTS AND DISCUSSION

A. Data

The proposed framework uses public available mental task classification dataset [7], where total seven subjects participated in the recording of EEG signals; however, subject 4 is ignored due to incomplete information. Subjects are instructed to perform five mental tasks: 1) baseline (relax: B) 2) the mental letter composing task (L); 3) the non trivial mathematical task (M); 4) the visualizing counting of numbers written on a blackboard task (C); and 5) the geometric figure rotation task (R). Each EEG recording consists of the five trials of each of five mental tasks. The EEG signals are recorded from C3, C4, P3, P4, O1, and O2 electrode position with A1 and A2 as the reference electrode shown in
Fig. 6. FDR score for a pair of baseline task and count task for features extracted using MUSIC.

Fig. 7. Comparison of different combination of univariate methods and PSD methods in terms of classification accuracy.

Fig. 1. Each trial is recorded for 10 s with the sampling rate of 250 per second, which resulted in 2500 samples points per trial.

The proposed framework for mental task classification is shown in Fig. 2, which consists of four steps: 1) segmentation; 2) feature extraction; 3) feature selection; and 4) classification (to distinguish two different mental tasks). The proposed framework adopts filter feature selection technique to enhance the performance of learning algorithm for the classification of the mental tasks.

B. Feature Formation

For feature vector formulation, each trial data is preprocessed by decomposing into half-second segments, generating 20 fragments per trial for each subject. The extraction of features is carried out from each signal using three different PSD approaches, such as Welch, Burg, and MUSIC, separately. A total of 52 PSD values are obtained from each channel. Combining PSD values of all six channels, each signal represents 312 PSD values. The PSD values obtained for different tasks using Burg (parametric approach) for all six channels are shown in Fig. 3, which shows the extracting features from Burg PSD approach are effective in distinguishing different mental tasks. It can also be observed that PSD values at some frequency values differ considerably among different mental tasks (e.g., frequency range of 6–9 Hz for channel C3, 6–13 Hz for channel C4, 6–13 Hz for channel P3, 6–16 Hz for channel P4, 6–9 Hz for channel O1, and 16–19 Hz for channel O2). This difference in PSD values can help in distinguishing different mental tasks. While PSD values at some frequency values take similar values (e.g., frequency values above 15 Hz for C3, above 17 Hz for channel C4, above 13 Hz for channel O1, above 30 Hz for channel O2, above 20 Hz for channel P3, and above 22 Hz for channel P4) and cannot help in distinguishing different mental tasks. Similar observations are also noted for Welch and MUSIC methods. This means that all features (PSD values) are not relevant for mental task classification.

C. Application of Uni-Variate Feature Selection

To determine relevant features that can distinguish different mental tasks, four different univariate methods: Correlation (Cor); FDR; MI; and Wilcoxon’s rank-sum test (Ranksum) are investigated in this paper. FDR score corresponding to features obtained from each of the three PSD approaches
to distinguish Baseline task from Count Task is shown in Figs. 4–6. From Figs. 4–6, it can be seen that FDR score corresponding to few features is very high and less for others. This means that some features are more relevant than others. Similar observations are also noted for other univariate methods and other pairs of tasks. For all univariate feature selection methods, the top 25 ranked features are incrementally added to develop the decision model using forward feature selection approach. The comparison of different methods is reported in terms of maximum average classification accuracy for top features of ten runs of tenfold cross-validations. The three well-known classifiers: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and support vector machine (SVM) are used for mental task classification problem. Fig. 7 shows a comparison of all combinations of three PSD approaches and four univariate methods with each of the three PSD approaches without any feature selection in terms of average classification accuracy (over six subjects for all combination of tasks). Few observations can be seen from Fig. 7 as follows:

1) In general, the classification accuracy of all three PSD approaches improve with the use of univariate feature selection method with all three classifiers.

2) Among all the combinations of PSD approaches, univariate methods, and classifiers, the maximum classification accuracy is achieved with the combination of Burg, FDR, and SVM.

3) Among four univariate feature selection methods, maximum classification accuracy is achieved with FDR.

D. Application of Multivariate Feature Selection

Fig. 8 shows a color map of correlation values among top 20 relevant features obtained using the combination of FDR and Burg method to distinguish Baseline task from Count Task. It can be noted that some of the correlation values take a high value which depicts that such features are correlated (redundant) among themselves. The similar observations are also noted for other combinations of PSD approaches and univariate methods for another pair of tasks. This observation suggests the need to determine a subset of relevant and nonredundant features to further improve the performance of mental task classification. This paper uses four well known multivariate methods: LR, BD, SR, and mRMR to obtain minimal subset of nonredundant and relevant features using forward feature selection approach. Fig. 9 shows a comparison of all combinations of three PSD approaches and four multivariate methods with the combination of PSD approaches and FDR (best performing univariate method) in terms of average classification accuracy. Few observations can be seen from Fig. 9 as follows.
Among all the combinations of PSD approaches, multivariate feature selection methods, and classifiers, the maximum classification accuracy is achieved with the combination of Burg, LR, and LDA.

2) The performance of all combinations of PSD approaches and multivariate methods is better in comparison to the combination of PSD approaches and FDR for LDA and QDA in terms of classification accuracy.

3) The performance of MUSIC is worst among three PSD approaches with univariate as well as multivariate feature selection methods.

E. Relational Rankings

To investigate the relational rank of both univariate and multivariate methods, combination of feature selection and extraction techniques has been adopted. A robust ranking approach [63] has been utilized on the ground of percentage gain in classification accuracy with respect to without applying any feature selection method [64]. Fig. 10 shows 24 combinations of FS-FXT methods, which are the feature selection and extraction methods. These methods are compared on the basis of percentage gain in accuracy of the different combination of feature selection and extraction methods and their corresponding ranks. From Fig. 10, it can be observed that the combination of multivariate feature selection with all three feature extraction is ranked better in comparison to the combination of univariate feature selection and all three feature extraction methods except one combination (BD-MUSIC). Among all the combinations of selection and extraction methods, the combination of LR and Burg is best, whereas the team of MUSIC and Ranksum performs the worst.

F. Friedman Statistical Test

In order to compare the statistically significant difference evolving in various combinations of the feature selection and the PSD methods, Friedman on-parametric statistical test is adopted. From Table II, it can be noted that almost (11 out of 12) all combinations of multivariate feature selection with PSD methods obtained better rank than the combination of univariate feature selection method and PSD methods. Also, the SEL-EXT pair performance is examined with respect to a control method, i.e., the one that emerges with the lowest rank (combination of LR and Burg). In the comparison of the control method with other 23 combinations of feature selection and extraction method, adjusted p-values [65] is computed for consideration of accumulated error and to provide the correct correlation. The adjusted p-values show whether the control method having any statistical difference when compared with the other remaining methods. Table III
This paper presented the evaluations of the combination of three different PSD approaches with four well-known univariate and multivariate filter feature selection methods. The experimental findings demonstrate that the multivariate feature selection algorithms endure more distinguishable feature set for the mental task classification compared with univariate feature selection approach. In general, it is observed that the multivariate filter feature selection methods outperforms the univariate filter feature selection methods. The combination of Burg method, LR, and LDA achieved maximum classification accuracy among all other combinations. Also, in most of the cases, multivariate feature selection approach works better than univariate feature selection approach with the conjunction of PSD approach for mental task classification.

In the future, an individual extraction of spectral density of different brain frequency will be considered. Also, the approach of comparisons and investigations will be extended from binary mental task classification to the multiclass mental task classification.

REFERENCES


TABLE III
ADJUSTED p-VALUES FOR THE HOMMEL PROCEDURE

<table>
<thead>
<tr>
<th>Different Combination</th>
<th>unadjusted p</th>
<th>p Hommel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranksum + Welch</td>
<td>5.43E-12</td>
<td>1.29E-10</td>
</tr>
<tr>
<td>MI + MUSIC</td>
<td>1.37E-10</td>
<td>3.01E-9</td>
</tr>
<tr>
<td>Ranksum + Burg</td>
<td>1.70E-09</td>
<td>3.57E-08</td>
</tr>
<tr>
<td>FDR + Welch</td>
<td>4.91E-09</td>
<td>9.82E-08</td>
</tr>
<tr>
<td>MI + Burg</td>
<td>4.84E-08</td>
<td>8.07E-07</td>
</tr>
<tr>
<td>BD + MUSIC</td>
<td>5.85E-08</td>
<td>9.95E-07</td>
</tr>
<tr>
<td>Ranksum + MUSIC</td>
<td>9.91E-08</td>
<td>1.68E-06</td>
</tr>
<tr>
<td>CORR + Burg</td>
<td>2.77E-07</td>
<td>4.43E-06</td>
</tr>
<tr>
<td>MI + Welch</td>
<td>9.53E-06</td>
<td>1.43E-04</td>
</tr>
<tr>
<td>CORR + MUSIC</td>
<td>2.11E-05</td>
<td>2.95E-04</td>
</tr>
<tr>
<td>FDR + MUSIC</td>
<td>1.00E-04</td>
<td>0.001305</td>
</tr>
<tr>
<td>FDR + Burg</td>
<td>6.37E-04</td>
<td>0.007647</td>
</tr>
<tr>
<td>CORR + Welch</td>
<td>0.0020477</td>
<td>0.020477</td>
</tr>
<tr>
<td>mMKR + MUSIC</td>
<td>0.0031092</td>
<td>0.027983</td>
</tr>
<tr>
<td>LR + MUSIC</td>
<td>0.0048515</td>
<td>0.037211</td>
</tr>
<tr>
<td>SR + MUSIC</td>
<td>0.00799</td>
<td>0.0832</td>
</tr>
<tr>
<td>BD + Burg</td>
<td>0.8446384</td>
<td>0.312469</td>
</tr>
<tr>
<td>mMKR + Welch</td>
<td>0.1591641</td>
<td>0.637466</td>
</tr>
<tr>
<td>SR + Burg</td>
<td>0.2059032</td>
<td>0.823613</td>
</tr>
<tr>
<td>mMKR + Burg</td>
<td>0.2549452</td>
<td>0.91187</td>
</tr>
<tr>
<td>BD + Welch</td>
<td>0.5270893</td>
<td>0.91187</td>
</tr>
<tr>
<td>LR + Welch</td>
<td>0.837144</td>
<td>0.91187</td>
</tr>
<tr>
<td>SR + Welch</td>
<td>0.9118703</td>
<td>0.91187</td>
</tr>
</tbody>
</table>
Akshansh Gupta received the master’s and Ph.D. degrees in computer science and technology from the School of Computer and Systems Sciences, Jawaharlal Nehru University (JNU), New Delhi, India, in 2010 and 2015, respectively. He is currently a Department of Science and Technology (DST) funded Post-Doctoral Research Fellow as a Principle Investigator under the scheme of cognition Science Research Initiative with the DST, Ministry of Science and Technology, Government of India, with project entitled “Identification of EEG Signature for Different Mental States” with the School of Computational Integrative and Science, JNU. He is also a Co-PI of ICPS Program, DST, Government of India, on a consultancy project, “Development of Machine Learning Algorithms for Automated Classification Based on Advanced Signal Decomposition of EEG Signals.” His current research interests include pattern recognition, machine learning, data mining signal processing, brain–computer interface, cognitive science, and Internet of Things.
Ramesh Kumar Agrawal received the M.Tech. degree in computer science from the Indian Institute of Technology Delhi, New Delhi, India, and the Ph.D. degree in computational physics from the University of Delhi, New Delhi.

He is currently a Professor with the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. His current research interests include classification, feature extraction, and selection for pattern recognition problems in domains of image processing, security, and bioinformatics.

Jyoti Singh Kirar received the Ph.D. degree in computer science from Jawaharlal Nehru University, New Delhi, India.

She is currently an Assistant Professor with Shiv Nadar University, Greater Noida, India. Her current research interests include brain–computer interfaces, machine learning, signal processing, and artificial intelligence.

Javier Andreu-Perez received the Ph.D. degree in intelligent systems from the School of Computer Science and Communications, Lancaster University, Lancaster, U.K., in 2012.

He was a Research Associate of Machine Learning with the Department of Computing, Imperial College London, London, U.K. He is currently a University Lecturer of Artificial Intelligence and Machine Learning with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. (Top 20 Research U.K. and Emerging Technologies, such as Machine Learning, Signal Processing, and Artificial Intelligence).

Dr. Andreu-Perez is the Chair of the prestigious international task forces, such as the European Union and U.K. Research Councils, in addition to successful research collaborations with the industry. His current research interests include the development and investigation of new methods for artificial intelligence, machine learning, and their applications in neuroinformatics, neuroengineering, biomedical sensing, health informatics, and wearable computing.

Dr. Andreu-Perez is the Chair of the prestigious international task forces, such as the European Union and U.K. Research Councils, in addition to successful research collaborations with the industry. His current research interests include the development and investigation of new methods for artificial intelligence, machine learning, and their applications in neuroinformatics, neuroengineering, biomedical sensing, health informatics, and wearable computing.

Wei-Ping Ding received the B.S. degree in computer science and technology from Nantong University, Nantong, China, in 2002, the M.S. degree in software engineering from Soochow University, Suzhou, China, in 2005, and the Ph.D. degree in computer application from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013.

He was a Visiting Researcher with the Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, AB, Canada, with the financial support “Jiangsu Government Scholarship for Overseas Studies” in 2011. In 2014, he was a Post-Doctoral Researcher with the Brain Research Center, National Chiao Tung University, with Prof. C.-T. Lin, Hsinchu, Taiwan. He is currently an Associate Professor with the School of Computer Science and Technology, Nantong University. He has authored or coauthored over 60 papers in journals and conference proceedings. His current research interests include co-evolutionary algorithms, granular computing, data mining, and machine learning and their applications in medicine.

Dr. Ding was a recipient of the National Natural Science Young Foundation of China in 2013 and the Excellent-Young Teacher of Jiangsu Province sponsored by Qing Lan Project, Jiangsu Province, China, in 2014. He is a member of Association of Computing Machinery, IEEE Computer Society, and China Computer Federation.

Chin-Teng Lin (F’05) received the B.S. degree in electrical engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1986, and the master’s and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1989 and 1992, respectively.

He is currently a Chair Professor with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia, a Chair Professor of Electrical and Computer Engineering, NCTU, an International Faculty Member with the University of California at San Diego, San Diego, CA, USA, and an Honorary Professorship with the University of Nottingham, Nottingham, U.K. He has coauthored a book entitled Neural Fuzzy Systems (Prentice-Hall) and has authored the book entitled Neural Fuzzy Control Systems With Structure and Parameter Learning (World Scientific). He has published over 200 journal papers (Total Citation: 20 155, H-index: 53, and 10-index: 373) in the areas of neural networks, fuzzy systems, multimedia hardware/software, and cognitive neuro-engineering, including approximately 101 IEEE journal papers.

Dr. Lin has been the Editor-in-Chief of the IEEE Transactions on Fuzzy Systems since 2011. He also served on the Board of Governors at IEEE Circuits and Systems (CAS) Society from 2005 to 2008, IEEE Systems, Man, and Cybernetics Society from 2003 to 2005, IEEE Computational Intelligence Society (CIS) from 2008 to 2010, and the Chair of IEEE Taipei Section from 2009 to 2010. He is a Distinguished Lecturer of IEEE CAS Society from 2003 to 2005, and CIS Society from 2015 to 2017. He served as the Deputy Editor-in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART II: EXPRESS BRIEFS from 2006 to 2008. He was the Program Chair of IEEE International Conference on Systems, Man, and Cybernetics in 2005 and the General Chair of the 2011 IEEE International Conference on Fuzzy Systems. He was elevated to be an IEEE Fellow for his contributions to biologically inspired information systems in 2005, and was elevated to an International Fuzzy Systems Association Fellow in 2012.

Mukesh Prasad received the master’s degree in computer application from Jawaharlal Nehru University, New Delhi, India, in 2009, and the Ph.D. degree in computer science from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2015.

He was also a Post-Doctoral Fellow with NCTU until 2015. He was a Principle Engineer (Research and Development) with Taiwan Semiconductor Manufacturing Company, Hsinchu, until 2017. He is currently a Lecturer with the School of Software, University of Technology Sydney, Ultimo, NSW, Australia. He has published papers in international journal and conferences, including with IEEE, ACM, Elsevier, and Springer. His current research interests include machine learning, pattern recognition, fuzzy systems, neural networks, artificial intelligence, and brain–computer interface.