CHAPTER 4

4 Feature Selection using Regularized Neighbourhood Component Analysis to Enhance the Classification Performance of Motor Imagery Signals.

4.1 Introduction

Brain-computer interfaces (BCI) are a type of assistive technology that helps assist people with neuromuscular disorders. The BCI system establishes a direct communication pathway between the electrophysiological signals originated from the brain and the external devices, for instance, a robotic arm, prosthetic device, wheelchair, etc. [207]. It excludes the involvement of nerves and muscles which are damaged or not fully functional to perform movements. Electroencephalogram (EEG) is an electrophysiological signal generated from the brain of the human. In practice, to develop a BCI system, EEG signals of a subject are recorded while performing a specific task. These signals are originated from the different brain locations depending on the mental task performed by the subject, and thus these signals are converted into command signals for executing the control action of the external devices. Among many BCIs, Motor imagery (MI) based BCIs pick up the biopotential signals originated from the sensorimotor cortex area of the brain while a person imagines about the motor movements [208]. Event-related desynchronization (ERD) and event-related synchronization (ERS) patterns are observed in the mu rhythms (8-13 Hz) and beta rhythms (13-25 Hz) of the brain activity during motor imagery task [14].

Despite many feature selection techniques used in MI signal processing as discussed in section 2.4.2, there still exists a scope of reducing the high dimensionality of EEG training data. Yang et al. [128] proposed a new feature weighting approach based on neighbourhood component analysis (NCA) which optimizes the nearest neighbour classifier performance to address the issue of the high dimensionality of the training data. It is notable that feature selection techniques can enhance the accuracy of the prediction model for the MI dataset. In this study, we have proposed a method to regularize the conventional NCA method and investigated the performance of regularized NCA (RNCA) as a feature selection technique on two public MI datasets. The effectiveness of the proposed feature selection RNCA algorithm is compared with other feature selection algorithms such as PCA, GA, and ReliefF. The RNCA is relatively a new approach and to the best of our knowledge has not been applied to MI dataset. In our approach, the fundamental steps followed to differentiate the mental tasks include various statistical, phase and frequency features extraction, selection of the best subset of features using the RNCA and classification using a support vector machine (SVM) classifier.

The subsequent sections in this chapter are presented as follows. Section 4.2 describes the methods and materials used in this work such as dataset descriptions, feature extraction techniques, and various feature selection algorithms including the proposed RNCA feature selection method. Further, results and discussion are presented in section 4.3 and 4.4, respectively. Finally, the conclusion is drawn in section 4.5.

4.2 Methods and Materials

4.2.1 EEG Dataset and paradigm

In this work, two EEG datasets provided by BCI completion II and IV are used. Brief descriptions of the used datasets are as follows.

Dataset 1**:** The first dataset used in this work has been provided by BCI Competition II (Dataset III) [15]. Department of Medical Informatics, Institute for Biomedical Engineering, Graz University of Technology, (GertPfurtscheller, AloisSchlögl) experimented aiming to provide continuous control over a BCI- feedback system. This dataset has been recorded from a 25 years old female while performing a motor imagery task of 2 classes (left-hand or right-hand movement) with a feedback session. The dataset contains three channels EEG from channel locations C3, CZ, and C4, sampled at 128Hz and bandpass filtered between 0.5 and 30 Hz. The subject was sitting comfortably in an armchair and was instructed to control a feedback bar by imagining about the right-hand or left-hand movement according to the displayed cue. Total 140 labelled training trials and 140 unlabelled test trails were recorded. Every trail consists of 9s EEG recording with initial 2s as a rest period, at the time, t=2s an acoustic sound indicating the start of the trial occurred for 1s. At t=3s, the subject started to see an arrow as a cue indicating either to the right or the left and the subject performed the motor imagery task by moving the feedback bar in the direction of the cue by imagining the right-hand or left-hand movement.

Dataset 2**:** BCI competition IV (Dataset 2b) [16] provided the second dataset used in this work. Dataset 2b was recorded by the Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, (Robert Leeb, Clemens Brunner, Gernot -Müller-Putz, AloisSchlögl, GertPfurtscheller). This dataset consists of 3 bipolar EEG channels signals acquired from 9 subjects for two classes (Left and right-hand motor imagery). Bandpass filtering is applied between 0.5 Hz and 100 Hz with a notch filter at 50 Hz to remove the power line noise.

4.2.2 Preprocessing.

The randomness and non-stationary nature of EEG time series emerge the need of preprocessing step, which helps in reduction of the noise and identification of the main components present in the EEG signal. Temporal characteristics of the EEG signal suggest that it varies significantly between different subjects and also between different training sessions with the same subject. To avoid such issues, time-frequency analysis of the EEG signal is performed to extract features in time and frequency domain at the same time. To accomplish this, we have selected dual-tree complex wavelet transform (DTCWT) to decompose the EEG signal. DTCWT is an enhancement to a discrete wavelet transform (DWT). However, discrete wavelet transforms are widely used to deal with non-stationary signals like EEG, problems using DWT like aliasing and power losses at the transaction bands can be resolved using DTCWT [209].

In EEG signal, ocular artifacts are more prominent up to 10 Hz. To remove ocular artifacts, an adaptive threshold presented by Zikov et al. [206] is applied to the coefficients of the sub-bands with decomposition level below 10Hz. In DTCWT, two DWTs work in parallel to compute the real and imaginary part of the transform. Figure 4.1 shows the analysis and synthesis filter banks of DTCWT to analyze the EEG signal in the wavelet domain. In our method, the level of decomposition was selected to 4, and the signal is decomposed into four details, D1-D4, and one approximation A5.

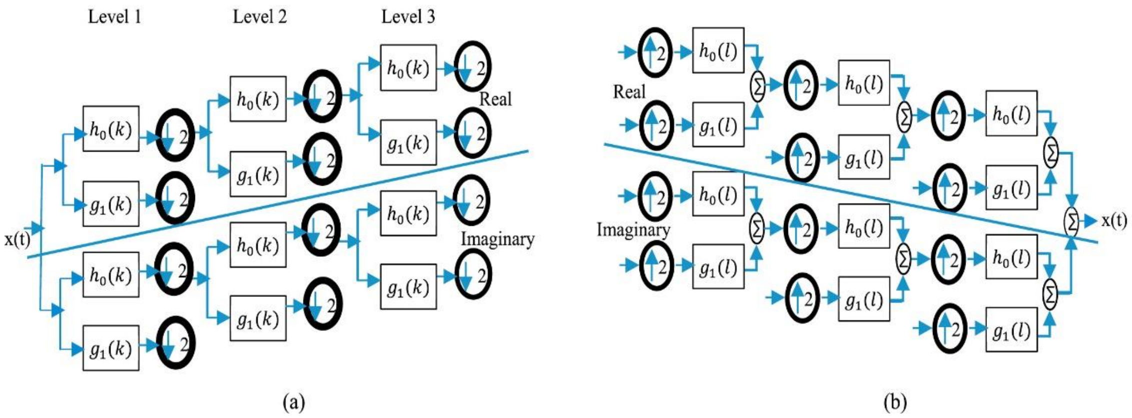


Figure 4.1 DTCWT (a) Analysis filter bank of DTCWT and (b) synthesis filter bank of DTCWT.

4.2.3 Feature Extraction

The EEG signal decomposed into frequency sub-bands after preprocessing results in increased dimensionality. Feature extraction is applied to reduce the high dimensionality and improve the differentiable capacity of the dataset between the different motor imagery classes. In other words, feature extraction is to provide the primary information carried by the raw EEG signal that can easily distinguish between different mental states. The feature space is generated that contains the highlighted information.

In this method, instead of using a wider frequency band we have used three frequency sub-bands to analysis the ERS/ERD patterns which are most prominent in mu and beta rhythms. From the preprocessing step wavelet coefficients of three decompositions, D2-D4 are selected considering the frequency range of interest 4-30 Hz. The frequency range of D2 and D3 are 4-8Hz and 8-15 Hz, respectively. Thus, it gives only the mu rhythms of the brain activity and filters the other frequency components. However, for beta activity, we have used sub-band D4 ranging between 15 Hz and 30 Hz.

Parameters used to create feature space are the combination of various statistical, frequency and phase information features of the EEG signal for each frequency sub-band. Parameters used in this work are as follows:

a) Statistical features: Discriminatory information from the wavelet coefficients of the raw EEG signal can be defined and calculated using statistical analysis [210,211]. In our approach, five statistical features are evaluated including mean absolute values (MAV), standard deviation, variance, sample entropy, and root mean square (RMS) values of the coefficients of the details D2-D4 for channel location C3 and C4.

b) Frequency features: Power spectral density (PSD) of the EEG signal describes the power carried by the signal as a function of frequency [212]. EEG signal of two channels C3 and C4 is filtered using DTCWT in three separate frequency bands ranging between 4 and 30Hz. PSD of each subband is evaluated.

c) Phase features: Phase locking value is a method to measure the instantaneous phase relationship between two signals [213]. We use EEG signal from Channel location

CZ as a reference signal, and phase relationship of C3 and C4 is calculated with respect to reference CZ. The PLV is calculated for all three subbands.

Considering three subbands and signal from two-channel locations with seven features extracted, the outcome dimension of the feature space is $3 \times 2 \times 7 = 42$ features. Hence, the complete training set is a $42 \times N$ matrix where N represents the total number of trails. Table 4 .1 lists the brief description of all the features.

Table 4.1 Description of different features extracted using statistical, frequency and phase analysis.

Feature Index	Feature description
1-6	MAV of the Details D2-D4 for two EEG channels ($3 \times 2 = 6$)
7-12	The standard deviation of the Details D2-D4 for two EEG channels ($3 \times 2 = 6$)
13-18	The variance of the Details D2-D4 for two EEG channels ($3 \times 2 = 6$)
19-24	Sample entropy of the Details D2-D4 for two EEG channels ($3 \times 2 = 6$)
25-30	Root mean square (RMS) of the Details D2-D4 for two EEG channels ($3 \times 2 = 6$)
31-36	PSD of the Details D2-D4 for two EEG channels ($3 \times 2 = 6$)
37-42	Phase locking value of the Details D2-D4 for two EEG channels ($3 \times 2 = 6$)

4.2.4 Feature Selection

Feature selection is performed to convert the m-dimensional feature vector to a lower p-dimensional feature vector by rejecting the redundant features. Also, the feature selection procedure plays a vital role in reducing the amount of data used for learning of the classifier. As a result, the execution speed of the classifier increases. Feature selection is said to be adequately performed if it enhances the generalization performance. Generally, feature selection methods are categorized into two groups: the wrapper approach and the filter approach. Wrapper methods evaluate the best subset of features by calculating the weights of all the features using a learning algorithm. On the contrary, the filter approach is a rank-based feature selection method that utilizes the predefined parameters to select the best features. In this section, four feature selection methods and the proposed algorithm has been explained in detail.

4.2.4.1 Genetic Algorithm

The theory of natural evolution given by Charles Darwin's inspired the Genetic Algorithms (GAs) [214]. GAs are computational models which replicate the natural process of selecting the fittest individuals to reproduce the best offspring of the next generation. As a feature selection technique, it works to preserve the critical information carried by the features. GAs solve the feature selection problem considering a chromosome-like data structure where a population of chromosomes is chosen, and each chromosome is encoded as an array of binary bits. The length of the array is taken equivalent to the size of the features set in the problem. Thus, each bit in the array represents one particular feature. Features with bit value ‘high' are selected for classification, and features with bit value ‘low’ are rejected. Hence, each chromosome represents one feature subset. Then, the fitness value of every chromosome from the population is measured as the kappa coefficient and classification accuracy using a learning algorithm. In our approach, population size was equal to the length of the features set, and tournament selection method was used with the elitist size set to two [215]. Also, arithmetic crossover function was applied to every generation for the creation of next-generation offspring. The generation size was varied during the experiment.

4.2.4.2 Principal Component Analysis

Principal component analysis (PCA) is an unsupervised feature selection technique that can linearly transform a higher dimensional feature space into a lower dimensional feature space using the statistical approach [216]. Some of the variables in the original feature space are correlated with one another, and there exists some redundancy. These correlated variables are linearly combined to a smaller number of principal components that preserve a maximum amount of variance in the variables. Principal components are orthogonal to one another to avoid the redundancy.

For feature selection, we have evaluated the contributions of individual features to the principal components. Let the size of the training set S is $n - by - p$, where rows of S correspond to trail observations and columns correspond to features. In PCA, S is transformed to a coefficient matrix of size $p - by - p$ in which each column represents coefficients of one principal component and columns are arranged in descending order of features variance [217]. To evaluate each feature contribution, the original data set S is multiplied with the PCA coefficient matrix, this will project the original data on the principal component vector space. In our method, we calculated the mean and variance of the columns of the projected data matrix to read the contribution of each feature to principal components. Thus, features with larger absolute mean compared with variance are selected for classification.

4.2.4.3 ReliefF

As a filter approach, ReliefF [125] is an effective algorithm to solve multiclass data problems. The algorithm randomly selects a sample x from the training set and searches for k nearest neighbor samples of the same class and k nearest neighbor samples of the non-similar classes. Using the Euclidean distance, closest nearest neighbor samples from each class are selected. A relevant weight is assigned to each feature by comparing interclass distance and intraclass distance from neighbor samples. This procedure is repeatedly performed on each feature sample, and each feature is assigned a weight. The features selected as best subset have weights larger to a predefined threshold.

4.2.5 Neighbourhood component analysis

Neighbourhood component analysis (NCA) is a learning algorithm to measure the Mahalanobis distance used in the KNN classification algorithm [218]. NCA as a feature selection technique is a feature weighting scheme to select the best subset of features maximizing an objective function that evaluates the average leave one out classification accuracy over the training data [128]. The algorithm works to assess a weighting vector w corresponds to the feature vector x_i optimizing the nearest neighbor learning classifier. In NCA framework, like 1-nearest neighbor classifier a reference sample point x_j is selected for the sample x_i from all the samples. The probability P_{ij} of x_j being chosen as a reference point for x_i from all the samples is higher depending on the closeness in the distance between the two samples. This distance can be measured by a weighted distance D_w defined as

$$D_w(x_i, x_j) = \sum_{m=1}^{r} w_m^2 |x_{im} - x_{jm}| \tag{4.1}$$

Where w_m is the assigned weight of the mth feature. The relation between the probability P_{ij} and weighted distance D_w can be established by introducing a kernel function k which returns large values for small D_w. The P_{ij} can be defined as

$$P_{ij} = \frac{k(D_w(x_i, x_j))}{\sum_{j=1, j \neq i}^{n} k(D_w(x_i, x_j))} \tag{4.2}$$

Also, if $i = j$ then Pii = 0. Where kernel function k is defined as $k(z) = exp(-\frac{z}{\sigma})$ and the parameter σ is the kernel width that affects the probability of a sample x_j to be selected as a reference point. Now, the probability of x_i being correctly classified can be written as

$$P_i = \sum_{j=1, j\neq i}^{n} P_{ij} Y_{ij} \tag{4.3}$$

Where Y_{ij} indicates one only if $y_i = y_j$. Hence, the average leave one out classification accuracy is the summation of P_i over all the trails divided by the total number of trails, and that can be seen as an objective function which needs to be maximized. However, the objective defined by Equation (4.3) is prone to overfitting. A term regularization parameter λ is introduced in the final objective function to avoid overfitting of the NCA model [128]. Thus, the objective function can be expressed as

$$A = \sum_{i=1}^{n} P_i - \lambda \sum_{m=1}^{r} w_m^2 \tag{4.4}$$

The objective defined in Equation. (4.4) is known as regularized NCA (RNCA). The target of RNCA is to maximize the objective function A. To perform this; A can be solved using the conjugate gradient approach. If A is limited to be a diagonal matrix, then its diagonal values provide the weight of each feature. Based on the weights outcome best subset of features is selected.

4.2.6 Proposed Method: RNCA as Feature Selection

This work assesses the performance of feature selection on BCI dataset using RNCA. The idea of this method is to select the subset of features which gives the maximum classification accuracy or minimum generalization error. Further, this study presents a technique to regularize the NCA model of Equation (4.4) for feature selection of MI data. Our approach contains the following steps:

Step1: To apply feature selection on BCI dataset we begin with considering the training set $S = \{(x_i, y_i), i = 1,2, .. N\}$ where, x_i are the feature vectors, N is the number of trails, $y_i \in \{1,2, .. C\}$ defines its class label, and C is the number of class. In this work, x_i comprise 42 features explained in section 3 for two motor imagery classes.

Step2: Perform the 5-fold cross-validation on the training set S and evaluate the generalization error, err defined as

$$err = \frac{1}{n}\sum_{i=1}^{N} I(k_i \neq t_i) \tag{4.5}$$

Where k_i represents the predicted class, t_i is the true class, $I(x)$ returns 1 when k_i is not equal to t_i else it gives 0.

Step 2 is executed to check if feature selection is required. Now fit the NCA model defined by Equation (4.4) keeping regularization parameter λ equals to zero and again calculate the generalization error. If the value of generalization error after fitting the NCA model is less than that of before fitting the model, then there is a requirement of feature selection.

Step 3: Now, we tune the regularization parameter λ to obtain the minimum classification loss. To perform this, generate a uniformly distributed array λ_{val} of length L. Then, fit the NCA model for each λ and store the estimated generalization error in an array.

Step 4: Step 2 is executed repeatedly for all folds and all values of λ. Simultaneously, average classification loss is calculated from all the folds for each λ value. Subsequently, the value of λ corresponds to minimum average classification loss is selected as best lambda λ_{Best}.

Step 5: Using the λ_{Best} value, the NCA model is performed on complete data and feature weights of each feature are evaluated. Features with weights higher than the 5% of the maximum feature weight are selected to classify the data.

Step 6: Next, train the SVM classifier using the updated training set selected in the previous step. Calculate the evaluation parameters such as confusion metrics, kappa value, and classification accuracy.

Note: Classification accuracy (CA) is equal to $1 - err$. Therefore, this algorithm works fine if we search for a subset with maximum CA instead of minimum classification error. The complete scheme of the proposed method has been summarized in Figure 4.2.

4.2.7 Classification

This work has implemented a support vector machine (SVM) as a classifier to perform the classification task on the two MI datasets used. SVM Classifier creates a discriminant hyperplane to improve the generalization capabilities by maximizing the margin between the classes. SVM is relatively a fast algorithm and capable of dealing with a large dataset [219]. Since the MI datasets used in this work have two classes (right-hand and left-hand motor imagery), SVM model creates a hyperplane to differentiate the two classes in a way that the gap between them is globally maximized.

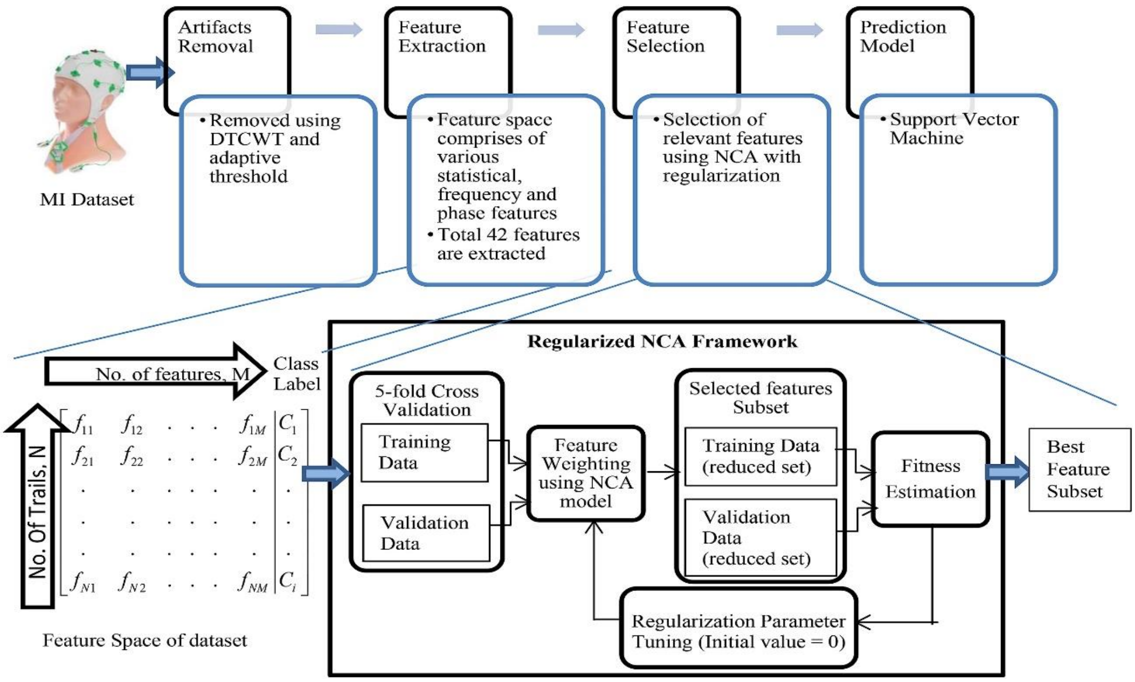


Figure 4.2 Workflow elucidates the proposed method used to reduce the high dimensionality of MI Dataset.

4.3 Results

This section elucidates the feature selection, and classification performances of the proposed method in comparison with the baseline methods explained in this chapter. Algorithms used in this work were developed and implemented in a computer having 12GB RAM and Intel Core i7 (@ 3.4 GHz) processor using the 64-bit version of Matlab R2018a software and applied on the two different BCI datasets explained in section II. Dataset 1 has EEG data from one subject for one training session. Whereas, dataset 2 consists of EEG data from 9 different subjects for two training sessions of each subject. Hence, the EEG recording of 19 training sessions has been used in this work.

4.3.1 Feature selection results:

The selection of a particular feature is based on the weight of that feature calculated by the feature selection algorithm. To represent the results of feature selection methods we have taken data of subject id B0101T of BCI competition IV dataset2b. Specifically, preceded by preprocessing and feature extraction using DTCWT, a feature space of dimension $42 \times N$ features, is generated. Figure 4.3(a) illustrates the weights evaluated for each feature using ReliefF. It shows the feature weights in the descending order, and

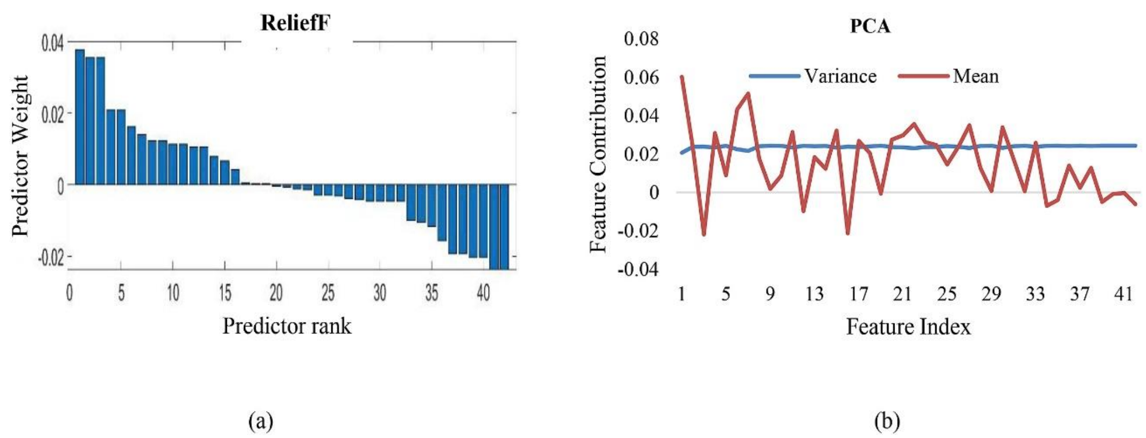


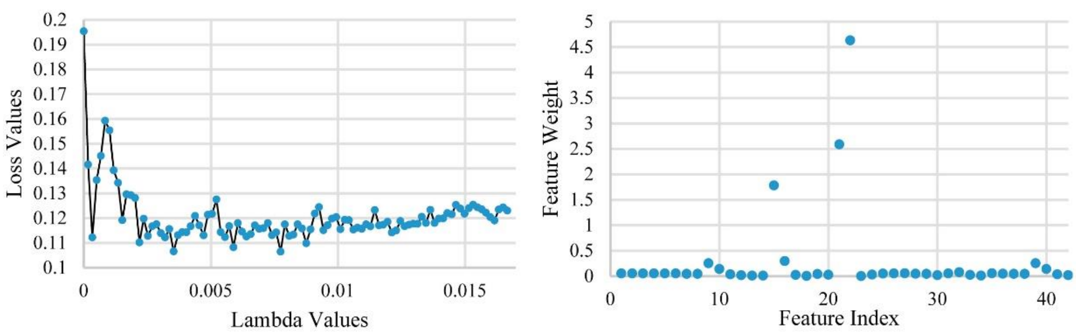
Figure 4.3 Weights assigned to different features using (a) ReliefF algorithm and (b) Principal Component Analysis (PCA).

accordingly, a rank is provided to each feature, i.e., the feature with the most significant weight has been assigned rank 1 and so on. For classification, we have only selected the features with positive weights.

In PCA analysis, the contribution of each feature to principal components is evaluated by projecting the dataset on another coordinate, i.e., principal axis. Then, the absolute mean and variance of each feature in principal co-ordinate are calculated and plotted as shown in Figure 4.3(b). Features with absolute mean more than the variance are selected for classification.

Performance of RNCA as feature selection is shown in Figure 4.4. First, the best regularization parameter was determined at 0.0077 with average classification loss of 0.105. Using this value of regularization parameter weights of all the features were evaluated. Subsequently, a threshold of 5% of the maximum weight is set to select the features. It is notable that only six features have weights significantly more than 5% of the maximum weight and hence, these features are selected for classification.

Further, the Figures 4.3 and 4.4 show the feature selection procedure applied to EEG data of only one training session; the same methodology was applied to all the EEG data of 19



(a)

(b)

Figure 4.4 Regularized Neighbourhood component analysis as feature selection (a) Estimation of the regularization parameter λ_{Best} at minimum loss value. (b) feature weights calculated using λ_{Best}=0.0077.

training sessions. Figure 4.5 presents the comparison of the average number of features selected using RNCA with ReleifF, PCA and GA. Although all the algorithms significantly eliminate the irrelevant features, it is notable that RNCA outperforms the other algorithms for most of the training sessions data with an average of 6.6±1.88 features selected. The number of features selected (FS) using various algorithms for each training session is listed in Table 4.2. However, the reduced number of features improves the execution speed of the classifier; it is also crucial that the selected subset of features must enhance the classification performance. The following section presents the classification results.

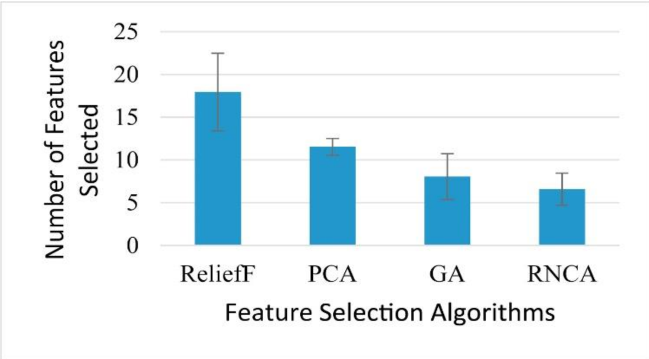


Figure 4.5 Comparison of the average number of features selected between four feature selection algorithms: ReleifF, PCA, GA, and RNCA for 19 different training sessions of 2 motor imagery datasets from BCI competition II (dataset III) and IV (dataset 2b). It also indicates the standard deviation from the mean value.

4.3.2 Classification Performance:

To compare the effectiveness of different feature selection approaches concerning classification performance, two evolution criteria [32]; classification accuracy and kappa coefficient has been chosen, which can be defined as

$$classification\ accuracy = p_0 = \frac{\sum_{i=1}^{C} n_{ii}}{\sum_{i=1}^{C}\sum_{j=1}^{C} n_{ij}} \tag{6}$$

Where n_{ii} and n_{ij} represent the elements of the confusion matrix and indicate how many times class i has been predicted as class j. If $i = j$ then a true class is predicted by the classifier. C is the number of class, which is 2 for our case.

$$\text{and, } kappa\ cofficient = \frac{p_0 - p_e}{1 - p_e} \qquad (7)$$

Where p_e is the expected accuracy turns out to be 0.5 for a two-class problem. SVM is used as a classifier, and 5-fold cross-validation has been applied to divide the data into training and test sets. The comparison of classification accuracy, kappa coefficient and the number of features selected using feature selection techniques such as RNCA, PCA, ReliefF, and GA has been listed in Table 4.2. It is observable that the learning-based feature selection methods such as GA and RNCA perform better than rank- based feature selection methods such as ReliefF and PCA in terms of classification performance. Also, in learning based feature selection approaches, RNCA obtained better classification performance than the GA. The average classification accuracy and kappa achieved by RNCA was 80.7% and 0.615 which is better than that produced by GA 78.9% and 0.579. Best classification accuracy, 99.2% was made by subject id B0401T using RNCA. As seen from Table 4.2, by comparing the classification accuracy, and kappa coefficient, RNCA achieves the highest values for all subjects, except for subject IDs BCI1, B0302T, B0702T, B0801T, B0902T.

Besides classification accuracy and kappa coefficient, the confusion matrix is further analyzed in detail and parameters such as precision, recall or sensitivity, specificity, and F1-score are evaluated for each subject and averaged over data of 18 training sessions of dataset 2. These parameters range between 0 and 1, with 1 indicating the highest classification performance. Table 4.3 lists the average values of confusion metrics obtained using different feature selection approaches. Obtained results suggest that the

Table 4.2 Comparison of classification accuracy (CA), kappa coefficient and the number of features selected (FS) between ReliefF, PCA, GA and RNCA for motor imagery data of 19 different training sessions of 2 different datasets from BCI competition II (dataset III) and IV (dataset 2b). Values in boldness indicate the largest value compared with all others.

	Subject ID	**All Features**			**ReliefF**		
		CA	Kappa	FS	CA	Kappa	FS
Dataset 1	BCI1	77.1	0.542	42	80.7	0.614	26
Dataset 2	B0101T	84.2	0.684	42	88.3	0.766	16
	B0102T	82.5	0.650	42	91.7	0.834	10
	B0201T	62.5	0.250	42	64.2	0.284	22
	B0202T	60.0	0.200	42	62.5	0.250	12
	B0301T	58.3	0.166	42	64.2	0.284	13
	B0302T	49.2	-0.016	42	51.7	0.234	33
	B0401T	87.5	0.75	42	80.0	0.600	10
	B0402T	92.1	0.842	42	91.4	0.828	38
	B0501T	75.8	0.516	42	90.0	0.800	11
	B0502T	**73.6**	**0.472**	42	72.1	0.442	13
	B0601T	79.2	0.584	42	80.8	0.616	30
	B0602T	76.7	0.534	42	85.0	0.700	12
	B0701T	67.5	0.350	42	65.8	0.316	19
	B0702T	61.7	0.234	42	**65.8**	**0.316**	23
	B0801T	67.5	0.350	42	73.1	0.462	13
	B0802T	75.8	0.516	42	65.8	0.316	7
	B0901T	67.5	0.35	42	85.0	0.700	8
	B0902T	73.3	0.466	42	81.7	0.634	25
Mean Values		72.2	0.444	42	75.7	0.526	18

Table 4.2 (continued)

	Subject ID	**PCA**			**GA**			**RNCA**		
		CA	Kappa	FS	CA	Kappa	FS	CA	Kappa	FS
Dataset 1	BCI1	69.3	0.586	11	**81.4**	**0.628**	7	80.7	0.614	11
Dataset 2	B0101T	72.2	0.444	15	91.7	0.834	19	**93.3**	**0.866**	6
	B0102T	65	0.300	12	91.7	0.834	8	**92.5**	**0.850**	7
	B0201T	57.5	0.150	12	72.5	0.450	8	**73.3**	**0.466**	16
	B0202T	56.7	0.134	15	70.8	0.416	8	**73.3**	**0.466**	7
	B0301T	54.2	0.084	11	66.7	0.334	4	**67.5**	**0.350**	7
	B0302T	50.8	0.016	10	**54.2**	**0.084**	4	49.2	-0.016	3
	B0401T	97.5	0.950	12	95.8	0.916	25	**99.2**	**0.984**	4
	B0402T	77.1	0.542	10	**94.3**	**0.886**	3	**94.3**	**0.886**	5
	B0501T	64.2	0.284	9	**91.7**	**0.834**	6	**91.7**	**0.834**	6
	B0502T	70.0	0.600	15	70.7	0.414	7	**73.6**	**0.472**	8
	B0601T	60.0	0.200	11	86.7	0.734	8	**94.2**	**0.884**	3
	B0602T	60.8	0.216	9	87.5	0.750	5	**88.3**	**0.766**	2
	B0701T	59.2	0.184	11	65.8	0.316	5	**70.8**	**0.416**	11
	B0702T	50.8	0.016	9	63.3	0.266	11	64.2	0.284	9
	B0801T	62.5	0.250	13	**78.8**	**0.576**	5	78.1	0.562	2
	B0802T	70.0	0.400	10	78.3	0.566	6	**80.8**	**0.616**	6
	B0901T	66.7	0.334	11	75.0	0.500	6	**88.3**	**0.766**	2
	B0902T	57.5	0.150	13	**83.3**	**0.666**	8	81.7	0.634	11
Mean Values		64.3	0.307	11.52	78.9	0.579	8	**80.7**	**0.615**	6.6

RNCA improved the classification performance in comparison to the other compared feature selection algorithms.

Further, we have performed the Friedman test and a Wilcoxon signed ranks test as the post hoc statistical test to statistically validate the results obtained in Table 4.2. The null hypothesis assumes that the performance of all the algorithms is identical. Preceded by ranking all the algorithms for each subject ID separately, the Friedman test, then averages the ranks over all the subject Ids and calculate the p-value. A low p-value indicates that the null hypothesis is rejected and there is a statistical difference. To find out which algorithm has the source of difference¸ the Wilcoxon signed ranks test as post hoc statistical test is conducted. The algorithm with the best rank is selected as the control algorithm, then a pairwise comparison of all the other algorithm and the control algorithm is performed. We applied the Friedman test on the classification accuracies obtained in Table 4.2 of dataset 2 and found that the null hypothesis was rejected ($p=2.55e-9$). The RNCA achieved the best rank, 1.33 and selected as the control algorithm to perform the Wilcoxon signed ranks as post hoc test. Table 4.4 shows the mean rank obtained across the 18 data of dataset 2. Whereas, Table 4.5 presents the p values obtained by pairwise comparison of the algorithms. Statistical test results indicate that the classification performance of SVM with RNCA as feature selection is better than that of without feature selection ($p<0.05$), feature selection with ReliefF ($p<0.05$), PCA ($p<0.05$) and GA ($p<0.05$).

4.4 Discussion

The present study analyzed the effectiveness of the proposed feature selection RNCA approach on MI EEG classification. Although comparison with some baseline feature selection methods has proven the superiority of RNCA, further discussion on the mechanism of the algorithm is provided in this section declaring the potential advantages and limitations of the proposed method.

4.4.1 Selection of optimal features by RNCA

For MI EEG pattern recognition, sensory motor rhythms (SMR) are analyzed to discriminate between two classes of motor imagery [14]. In various studies, it is noted that the selection of an inappropriate frequency band leads to suboptimal classification performance; hence care has to be taken while selecting a frequency band [28,124]. In

Table 4.3 Mean rank of feature selection methods for the Friedman test.

Feature Selection Method	All Features	ReliefF	PCA	GA	RNCA
Mean Rank	3.83	2.97	4.52	2.33	1.33

Table 4.4 Statistical analysis of classification accuracy differences between the compared methods on BCI competition IV Dataset IIb: Results of Wilcoxon signed ranks test as post hoc with alpha=0.05.

Hypothesis	RNCA vs. All Features	RNCA vs. PCA	RNCA vs. ReliefF	RNCA vs. GA
P-value (alpha=0.05)	0.000438	0.001169	0.000232	0.017135

Table 4.5 Averaged classification performance metrics of SVM classifier on BCI competition IV dataset 2b present a comparison between compared feature selection methods. Values in boldness indicate the largest value compared with all others.

	All Features	**ReliefF**	**PCA**	**GA**	**RNCA**
Precision	0.7103	0.7518	0.6454	0.8277	**0.8541**
Recall or Sensitivity	0.6836	0.7408	0.6287	0.7408	**0.7895**
Specificity	0.7305	0.7483	0.6785	0.8224	**0.8341**
F1 Score	0.6951	0.7419	0.6355	0.7625	**0.8089**

our approach, filtering of EEG into three sub-bands of frequency ranges 4-8Hz, 8-15Hz and 15-30Hz, respectively is performed by DTCWT and accordingly, time, frequency, and phase features are extracted. Afterward, features with greater importance have been selected using RNCA. Results presented in Table 4.2, 4.3, 4.4 and 4.5 has shown that RNCA selects less number of features with increased classification performance compared to GA, ReliefF, and PCA. Another finding of the results is that the frequency domain feature, i.e., the power spectral density (PSD) of the channels C3 and C4 in the frequency range (8-15Hz) has high importance and been selected for 8 out of 10 subjects (see Figure 4.6). These results are backed by numerous studies [14,211,221] that have suggested that the PSD of the mu rhythm (8-13Hz) carries vital information about the motor imagery tasks. Furthermore, as we have used multiple feature extraction methods, our proposed algorithm RNCA has selected significant features which are different for different subjects, which shows that RNCA is a robust algorithm to be used for the design of a subject-specific BCI system.

However, RNCA has shown potential in selecting optimal features from appropriate frequency bands, a few studies [28,169] have demonstrated the importance of choosing a potential time window, since neural response time to different MI task is subject specific. Therefore, use of a fixed time window may degrade the classification performance of MI BCI system, this worth our consideration. Keeping this in view our work can be further extended by using shifted and varying time windows but it will also increase the computational cost.

4.4.2 Configurable parameter

By far, numerous wrapper structure-based feature selection approaches have been proposed to improve the classification performance of MI EEG [27,28,120,222]. Experimental results have found that the behavior of most of these algorithms is highly variant because of the diversity in the selection of the configurable parameters. The parameter controlling the nature of our approach is the length, L of the uniformly distributed array λ_{val}. We have investigated the classification accuracy against varying values of L. In particular, we varied the value of L from 40 to 160 in steps of 20 and evaluated the classification accuracy for five subjects (see Figure 4.7). It should be noted that the stability of our proposed algorithm is affected by varying values of L to some extent. For most of the subjects, CA is much stable over varying values of L especially in

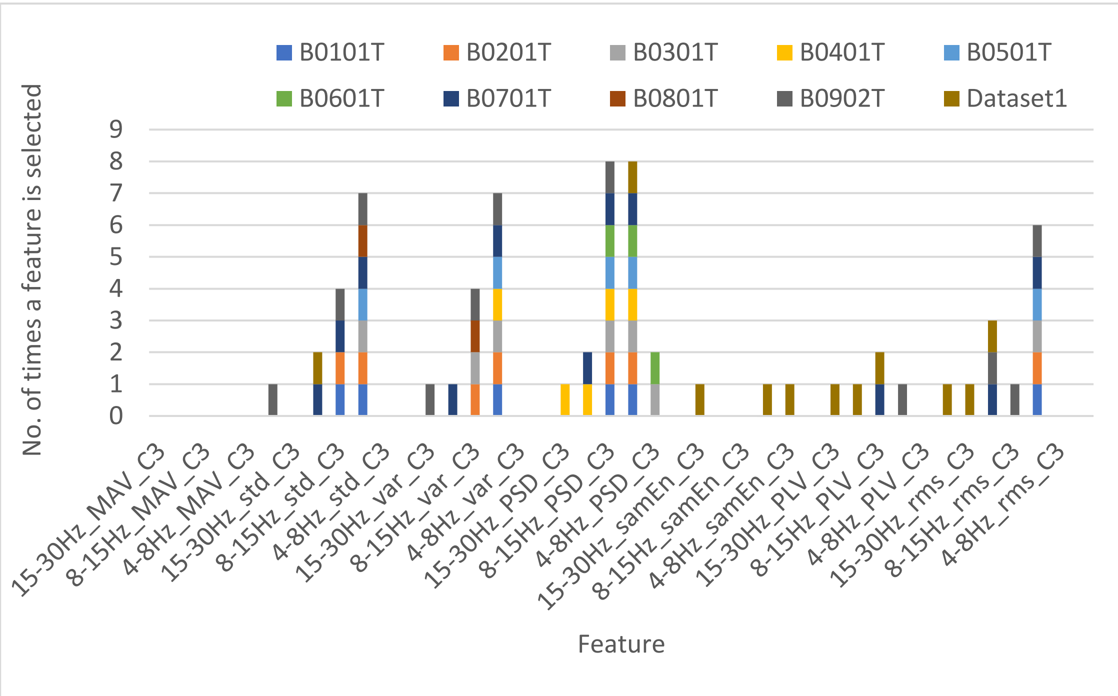


Figure 4.6 The number of times a feature is selected by RNCA for ten subjects. Results indicate that the RNCA feature selection method is capable of selecting optimal features that occur at different frequency bands for different subjects.

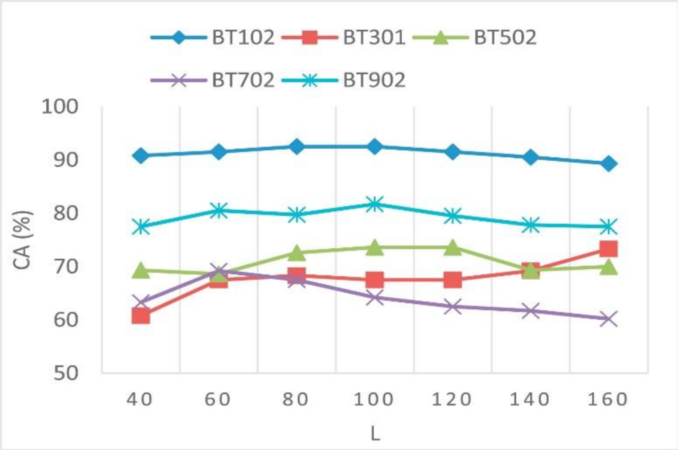


Figure 4.7 Change in classification Accuracy (CA) with respect to parameter L of RNCA for five subjects BT102, BT301, BT502, BT702, and BT902, respectively.

the range from 80 to 120. Although this range of the parameter L has given a potentially improved classification accuracy, this method of selecting the optimal range of parameter L decreases the processing speed of the algorithm and also requires additional dataset. This limits the practicality of the proposed feature selection method to some level. Hyperparameter optimization techniques in BCI have been introduced in some of the recent studies [223,224] and have shown potential in automatic selection of the optimal hyperparameter. As an extension to the proposed feature selection approach, some parameter optimization methods can be used to improve the BCI performance, which worth our future considerations.

4.4.3 Computational complexity

The complexity of the feature selection algorithm increases the total processing time. Table 4.6 lists the processing time, training time, and testing time of the different algorithms used in our work. From the results, it can be observed that the feature selection methods such as RNCA and GA that use a learning algorithm to optimize the classification performance have higher computation time than PCA and ReliefF. On the whole, the processing time of PCA is the fastest, with an average time of 0.3984s, while the processing time of RNCA and GA is 22.16s and 38.94s, respectively. Although RNCA feature selection time is higher due to the inner loop for estimation of classification error, it is a one-time procedure and doesn't affect the testing time. Also, due to the exclusion of a higher number of irrelevant features, RNCA with SVM classifier built a prediction model that attains a faster classification speed compared with prediction models made using GA, PCA, and ReliefF as feature selection methods and SVM classifier.

For the design of a practical BCI system, it is essential that the system must respond to different subjects with the same efficiency. Since neural response to MI task is subject and frequency band specific, with a fixed set of features high system performance can't be achieved. Hence, there lies the need for a practical feature selection approach capable of selecting an optimal subset of features for different subjects. This experimental study has demonstrated that the proposed RNCA feature selection method can efficiently be used for different subjects. Although the present study is conducted in an offline scenario, this work will be further extended to implement a real-time MI BCI system to control a prosthetic arm.

Table 4.6 Comparison of the average processing time of four feature selection methods: ReliefF, PCA, GA, and RNCA. SVM. Values in boldness indicate the largest value compared with all others.

	ReliefF	PCA	GA	RNCA
Algorithm Processing Time (s)	1.257415	0.3984	38.9404	22.1616
Training Time (s)	0.3478	0.3598	0.3758	0.3137
Testing Time (s)	0.0392	0.0250	0.0183	0.0124

4.5 Conclusion

This study presents the effectiveness of RNCA as a feature selection method to enhance the classification performance of the motor imagery tasks on datasets provided by BCI competition II and IV. For comparative study, we used two rank-based feature selection methods such as PCA and ReliefF and one learning-based method: GA. The SVM classifier has been used to evaluate the classification performance of each algorithm. We have investigated that the subset of features estimated by the proposed RNCA algorithm has successfully eliminated the irrelevant features and improved the overall classification performance of the SVM classifier for two class motor imagery problem. Results declare that the RNCA performed better feature selection task when compared with PCA, ReliefF, and GA. An important issue is the processing time of RNCA. Due to the diversity in the selection of configurable parameters that control the nature of RNCA and GA, the computational cost of these methods is higher than the rank-based feature selection methods. However, the execution time of the RNCA is faster than the GA. It is also concluded that due to the reduction in the feature space, the training and testing time of the prediction model was faster for RNCA than ReliefF, PCA, and GA. Further, this work can be extended by varying the feature space dimensions and optimizing the configurable parameters that control the nature of RNCA to attain a reasonable processing speed. This study concludes that to discriminate between different motor related mental tasks where high classification accuracy with a reasonable processing speed is required, feature selection using RNCA is the best choice.