Hybrid Feature Selection based on BTLBO and RNCA to Diagnose the Breast Cancer

Mohan Allam
School of Computer Science and Engineering
Lovely Professional University, India
allam.27289@lpu.co.in

Nandhini Malaiyappan
Department of Computer Science
Pondicherry University, India
mnandhini2005@yahoo.com

Abstract: *Feature selection is a feasible solution to improve the speed and performance of machine learning models. Optimization algorithms are doing a significant job in searching for optimal variables from feature space. Recent feature selection methods are purely depending on various meta heuristic algorithms for searching a good combination of features without considering the importance of individual features, which makes classification models to suffer from local optima or overfitting problems. In this paper, a novel hybrid feature subset selection technique is introduced based on Regularized Neighborhood Component Analysis (RNCA) and Binary Teaching Learning Based Optimization (BTLBO) algorithms to overcome the above problems. RNCA algorithm assigns weights to the attributes based on their contribution in building the learning models for classification. BTLBO algorithm computes the fitness of individuals with respect to the weights of features and selects the best ones. The results of similar feature selection methods are matched with the proposed hybrid model and proved better performance in terms of classification accuracy, recall and AUC measures over breast cancer datasets.*

Keywords: *Hybrid feature selection, binary teaching learning based optimization, neighborhood component analysis.*

Received August 23, 2021; accepted December 1, 2022
https://doi.org/10.34028/iajit/20/5/5

1. Introduction

Machine Learning is an emerging field of artificial intelligence which enables computers in building models from experience data to predict and automate the decision-making process. These models learn to improve their performance using various features available in the datasets. Datasets contain numerous features in which some are useful, and others are irrelevant or redundant. Correlated features need to be selected out of the total features in the dataset to boost the attainment of a classifier. Selecting optimal subset of features from a huge dataset for a particular application is a difficult job. Feature Selection methods can downsize the dimension of data to be used in constructing a machine learning model for predictive analysis. The main goal of feature selection is to make machine learning models simple for interpretation and to handle the curse of dimensionality by selecting an optimal subset of features. Thus, feature selection aids well in understanding the information of datasets to improve the performance of predictions and reduces the attribute search space as well as computing time of learning models. Feature selection methods are classified into 3 groups. The first one is the filter method, in which features are ranked based on their relevance and selects the best ones. The second is the Wrapper model, in which classification models will play a key role in selecting a good combination of features.

The last one is the embedded method, which can take advantage of both the above methods.

Most of the feature selection methods make use of searching algorithms to provide solutions for overfitting problems with generalized models. Different meta-heuristic algorithms [16] are used for searching attributes in the feature space to improve the efficiency of the classifiers. Genetic Algorithm (GA) is a widespread nature-inspired evolutionary method used for optimization and searching problems. But GA has few limitations in dealing with exploitation of search space. Guha *et al*. [8] resolved this problem with a local search algorithm (Great Deluge Algorithm) and achieved better performance with multiple classification models. Particle Swarm Optimization (PSO) is another popular optimization method for searching the solutions in the population. Too *et al*. [26] proposed a wrapper-based feature selection algorithm using binary competitive swarm optimizer and achieved high accuracies with the UCI machine learning repositories. Ahmad *et al*. [1] developed a wrapper method for text feature selection from customer reviews using Ant Colony Optimization (ACO) and k-Nearest Neighbor (k-NN) for sentiment analysis. Alfarraj *et al*. [2] proposed a method using firefly and gravitational ant colony algorithm for selecting optimized features from big data to solve the dimensionality problem in predictive analytics.

The performance of meta-heuristic algorithms mainly depends on the controlling parameters throughout the optimization process. Auto tuning of these parameters is important for improving the results of feature selection applications. This problem can be resolved with other advanced optimization algorithms which need only few or no parameters during the selection process. Teaching learning based optimization [18, 19] is a metaheuristic algorithm used in various research problems and does not use any external controlling variables during the optimization process. Allam and Nandhini [4, 5] proposed two wrapper selection methods based on TLBO and Binary TLBO algorithms for selecting optimal features from WDBC dataset and achieved better results in terms of accuracy, precision, recall and AUC measures. Neighborhood Component Analysis (NCA) [7] is a non-parametric learning model used for the classification and dimensionality reduction of datasets by learning a distance metric with a linear conversion of input data. NCA learns the feature weights with respect to the classification loss of training data for feature selection. But the authors acknowledged the overfitting problem of the NCA model with high dimensional data and upgraded the method with regularization concept [30]. A new parameter “λ” (regularization parameter) is added for the new model, called RNCA, to avoid the overfitting possibility of NCA method.

In this proposal, a new hybrid feature selection technique is presented that integrates RNCA in BTLBO algorithm to approximate the fitness of individuals with respect to feature weights. The feature selection process uses the BTLBO algorithm for searching a good combination of attributes by considering the importance (weight) of individual variables in the original dataset. The performance of the proposed system is measured using various measures such as classification accuracy, recall, F-Measure and Area Under the Curve (AUC) values. The key contribution of the proposed method is, designing a hybrid version of the feature selection method for optimal feature selection using the combination of BTLBO and RNCA algorithms.

The remaining sections will describe the proposed work. In the second section, earlier research on hybrid methods for attribute selection are discussed using various optimization algorithms. Section 3 deals with the proposed hybrid attribute selection model by means of BTLBO and RNCA algorithms. The discussion on the results of existing and current work is stated in the fourth section. The final section gives the outline of current work along with future scope.

2. Related Works

The performance of a machine learning model primarily hangs on the association of individual features towards a specific problem. The selected attributes with higher weights in a subset will decide the outcome of the learning model. Different hybrid methods are identified for selecting the best feature subsets using various optimization techniques. Liang *et al*. [13] used the power of both ACO and brainstorm optimization algorithms to develop a feature selection model and produced improved accuracy outcomes with 6 different datasets. Hsu *et al*. [9] developed a hybrid variable selection model by merging the filter and wrapper feature selection approaches to select useful genes from microarray cancer data. They also predicted the protein disordered region using the attributes selected from the proposed model and evaluated the performance with SVM classifier.

Most of the variable selection methods employed in disease diagnosis applications [6] are based on evolutionary algorithms. Rajalaxmi [17] developed a Hybrid Binary Cuckoo Search and Genetic Algorithm (HBCS-GA) to extract relevant features from Type-2 Diabetes datasets. The HBCS-GA model achieved better accuracies with various classifiers like SVM, k-NN and Decision Trees. Sun *et al*. [24] proposed a novel hybrid (filter-wrapper) gene selection technique using ReliefF and ACO algorithms for tumor classification on DNA microarray datasets. They used ReliefF algorithm for dividing the gene weights to reduce the dimensionality of datasets. Khourdifi and Bahaj [11] proposed a hybrid model optimized by PSO and ACO algorithms to predict heart disease using Fast Correlation-Based feature selection. Jain and Salau [10] used a different filter, wrapper, and embedded based feature selection methods for selecting important variables from images. The authors evaluated the proposed model using k-NN and SVM classification models with various approaches.

A few researchers also used TLBO algorithm for selecting valuable attributes in various applications. Tuo *et al*. [27] developed a hybrid algorithm using harmony search for global exploration and TLBO for local exploitation of feature space to solve complex problems of high dimensionality. Sevinç and Dökeroğlu [22] presented a new hybrid algorithm with a combination of TLBO and extreme learning machines for classifying different class of problems on UCI datasets. A novel multiple objective feature selection technique was developed [12] using TLBO algorithm for binary classification problems. The proposed method achieved better results compared to other metaheuristic algorithms like GA, PSO, and Tabu search with UCI datasets. Satapathy *et al*. [20, 21] proposed a feature selection method with a new combination of TLBO and rough set theory and compared results with other hybrid selection models. Taghanaki *et al*. [25] developed an intrusion detection system which uses NCA for feature transformation and GA for feature selection. The system achieved better results than other techniques with KDD Cup99 dataset. In the paper, Zhao *et al*. [31] used NCA for selecting key attributes and SVM classifier to build a model for predicting the HBV Reactivation in Primary Liver cancer from the subset. Shang *et al*. [23] developed a hybrid model to predict the traffic incident duration using NCA for selecting optimal variables, Random Forest (RF) to build the model for classification and Bayesian optimization algorithm to tune the RF parameters, respectively. A confusion matrix was produced to measure the performance of the RF model constructed with NCA based subset of features.

In this work, a novel hybrid model is developed for selecting a subset of optimal attributes from feature space. The model is formed with the combination Binary TLBO for optimal feature search and NCA for evaluating the feature importance in terms of weights. The results are compared with various kinds of exiting algorithms and observed significant enhancement with WDBC & Coimbra breast cancer datasets. In the following section, the process of the proposed hybrid (BTLBO-NCA) model is discussed with a flowchart. Ali *et al*. [3] presented the benefits of LDA algorithm for the reduction of the dimension of big datasets and also enhanced the performance of several classification models.

In literature study, we observed that most hybrid feature selection methods have been designed with a combination of homogeneous meta heuristic algorithms. These feature selection methods are not giving any weightage to individual features which are very important for providing solution to the specific problem. The classification models built using these features will suffer from local optima or overfitting problems.

3. Methods and Materials

Feature Selection process searches the entire feature space and identifies the likely attributes suitable for the problem. The designated variables are taken from the initial dataset containing all features to construct optimal dataset with useful features. The entire procedure of selecting the relevant attributes from the given population will happen in several generations. In every generation, the fitness of individuals is computed as accuracy (or error) from classification models in wrapper methods or metrics like correlation coefficient and mutual information in filter methods. Instances having high fitness values will be carried to the next generation with a good number of related attributes. Succeeding generations will carry the best individuals with optimal features. A revolutionary method is required to explore and exploit the whole feature space. Here, a novel hybrid method proposed with BTLBO optimization algorithm to explore the dataset for a diverse combination of features in the population and RNCA for assigning weights to the features used in estimating the fitness of the instances.

3.1. BTLBO Algorithm

TLBO is one of the population-oriented optimization algorithms which does not depend on any internal arguments. The binary version of TLBO procedure is used to quest the feature set for choosing the ideal subset of attributes. The procedure operates in two steps, called teacher and learner stages, for exploring the complete dataset. The algorithm ends with a specified number of iterations by selecting an individual with an optimal subset of attributes. Each individual (student) has ‘t’ number of attributes (features(f) from 1 to t) and ‘s’ instances (individuals(i) from 1 to s). The variable ‘v’ limits the procedure in terms of iterations. The working steps of BTLBO is shown below:

- *Step* 1: specify the number of individuals (binary) and features $X_{f,i}$ in the population and termination criteria.
- *Step* 2: for each iteration v, calculate the mean for attributes individually as $M_{f,v}$.
- *Step* 3: evaluate the fitness of instances by Equation (1).

$$\text{Fitness}\left(X_{f,i,v}\right) = \text{Error}\left(X_{f,i,v}\right) \tag{1}$$

Teacher Phase:

- *Step* 4: adapt all individuals (Students) with respect to the best individual (Teacher)

a. Choose the top student (Best fitness) from the class as an instructor.
b. Compute the mean difference for all the features using the top student as shown in Equation (2).

$$\text{Dif\_Mean}_{f,i,v} = r_v\left(X_{f,itop,v} - T_F M_{f,v}\right) \tag{2}$$

Where, $X_{f,itop,v}$ is the top student in the subject *f*. T_F is the teaching factor (1 or 2) and r_v is the random number (0 to 1).

c. The top student performs as an instructor and educates the outstanding students. Update students in the class with Equation (3).

$$\begin{gathered} X'_{f,i,v} = 0 \quad \text{if } X_{f,i,v} + \text{Dif\_Mean}_{f,v} < 0.5 \\ X'_{f,i,v} = 1 \quad \text{if } X_{f,i,v} + \text{Dif\_Mean}_{f,v} \geq 0.5 \end{gathered} \tag{3}$$

Here, $X'_{f,i,v}$ is the altered value of $X_{f,i,v}$.

d. If $X'_{f,i,v}$ is finer than $X_{f,i,v}$,
use the updated value
else,
keep the old value.

Learner Phase:

- *Step* 5: upgrade every student using neighbour students in the class with Equations (4) and (5).

a. Elect two individuals G and H with the condition $X'_{total-G,k} \neq X'_{total-H,v}$ at random.
Where, $X'_{total-G,v}$, $X'_{total-H,v}$ are updated variables of $X_{total-G,v}$, $X_{total-H,v}$ with regard to G and H individuals respectively.
b. If $X'_{total-G,v}$ is finer than $X'_{total-H,v}$

$$X''_{f,G,v} = 0 \ \text{ if } X'_{f,G,v} + r_v\left(X'_{f,G,v} - X'_{f,H,v}\right) < \ 0.5$$

$$X''_{f,G,v} = 1 \text{ if } X'_{f,G,v} + r_v \left(X'_{f,G,v} - X'_{f,H,v}\right) \geq 0.5 \quad (4)$$

else,

$$X''_{f,G,v} = 0 \text{ if } X'_{f,G,v} + r_v \left(X'_{f,H,v} - X'_{f,G,v}\right) < 0.5$$

$$X''_{f,G,v} = 1 \text{ if } X'_{f,G,v} + r_v \left(X'_{f,H,v} - X'_{f,G,v}\right) \geq 0.5 \quad (5)$$

```
c. If $X''_{f,G,v}$ is finer than $X'_{f,G,v}$
        Use the updated value
   else,
        Keep the old value.
```

- Step 6: If the end criteria satisfied,

```
        display the outcome
    else,
        proceed to Step 2
```

The population is in binary format representing the existence and non-existence of a specific attribute in the instance. Each individual solution in the dataset contains binary bit patterns identical to the number of attributes. The bits '1' and '0' signify the existence and absence of specific attributes in the population. In the initial stage, the students will learn from the instructor by means of the difference between them. If the added variation in difference is bigger than a limit value (0.5), the student will follow the teacher solution in terms of feature availability as shown in Equation (3). In the later stage, the students will acquire and update the information in dual correspondence as shown in Equations (4) and (5).

3.2. RNCA Algorithm

NCA algorithm measures a value, called Mahalanobis distance, in k-NN learning model for classification. This method will learn the distance to maximize the performance of a k-NN classifier. RNCA will do feature weighting for selecting a subset of optimal features. The feature weighting process measures a weighting vector by maximizing the target function of Leave-One-Out (LOO) accuracy of a classification model [29]. The RNCA algorithm has the following steps,

- *Step* 1: The training set T has N number of individuals represented by T = {(x_i, y_i), i = 1, 2, 3, ... N}, Here x_i is the attribute vector & y_i is the target vector with class labels.
- *Step* 2: Let D_w is the distance between two individuals (x_i, x_j). The D_w is computed with respect to the weights defined in Equation (6). Where, w_k is the weight assigned to the kth attribute.

$$D_w(x_i, y_i) = \sum_{k=0}^{q} w_k^2 \mid x_{i,k} - x_{j,k} \mid \quad (6)$$

- *Step* 3: The x_i will select the nearest neighbor x_j as a reference point with the probability P_{ij} (Equation (7)) to maximize its LOO classification accuracy or minimizes the classification error on the training vector T. The kernel function k gives big values with small distances.

$$P_{ij} = \frac{k(D_w(x_i, y_i))}{\sum_{j=1, j\neq i}^{n} k(D_w(x_i, y_i))} \quad (7)$$

The kernel function is indicated as, $k(m) = exp\left(-\frac{m}{\sigma}\right)$. Here, the variable σ gives the width of the kernel and influences the probability.

- *Step* 4: The probability of the instance x_i being correctly classified will be measured as a summation of the Equation (2) and the training target vector is shown in Equation (8).

$$P_i = \sum_{j=0,\ j\neq i}^{n} P_{ij}\ Y_{ij} \quad (8)$$

This probability will be treated as a fitness function which needs to be minimized for classification error.

- *Step* 5: Regularize the NCA using the parameter λ as shown in Equation (9).

$$\text{Fitness} = \sum_{i=1}^{n} P_i - \lambda \sum_{k=0}^{q} w_k^2 \quad (9)$$

The fitness function should be maximized or minimized for classification accuracy or error respectively. RNCA model will achieve the best performance for classification and feature reduction with the optimal value of λ.

3.3. Hybrid Feature Selection Using BTLBO and RNCA

A novel hybrid feature selection model is developed with a combination of BTLBO and RNCA algorithms as shown in Figure 1. The Binary TLBO mainly focuses on 'exploration and exploitation' of the feature space for selecting a good combination of subset of attributes by imitating the regular learning process of a student from teachers and classmates in school. The best relevant features in the individuals will train the other individuals to find a reduced optimal feature set. The RNCA model will serve the model as a fitness evaluator by learning and associating weights with a view to the attributes based on their importance in the dataset. In the end, the hybrid system will select the attributes having higher weights than the others in the population.

The proposed model of BTLBO-RNCA FS starts with the initialization of individuals in the population, fixing the number of iterations and teaching factor value. The number of ones in the result string indicates the attributes selected in the individual. The BTLBO algorithm is organized into two modules named teacher phase and learner phase. In teacher phase, fitness is evaluated for each instance in the solution space with the help of the RNCA learning model.

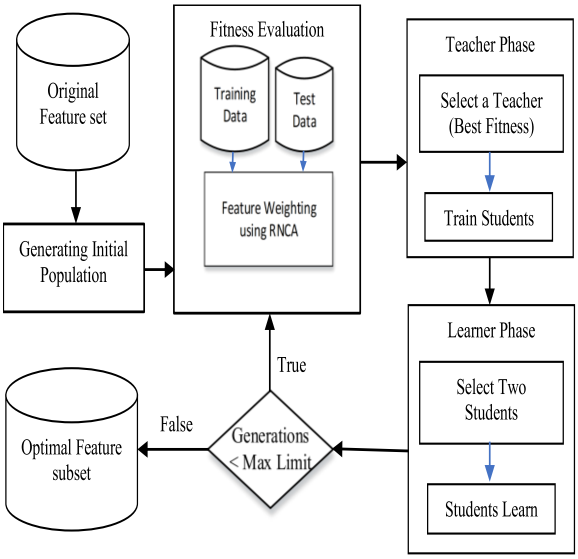


Figure 1. The workflow of BTLBO-RNCA hybrid model.

RNCA learns the weights with a view of the attributes for the training set and validates against the test set based on the correlation between input and output vectors in the population. Regularization of λ value will maximize the generalization capability of the NCA model to overcome the problem of overfitting while working towards classification problems. The candidate solutions having the highest weighting vector associated with the attributes will be chosen as an instructor (teacher). The instructor will calculate the difference mean (Equation (2)) for each attribute in the dataset and update the remaining individuals based on the computation shown in Equation (3). The updated candidate is compared with the old one to keep the best in the final population of the current generation.

In the learner phase, every candidate solution is updated with another randomly selected candidate using Equations (4) and (5) based on the preference given to fitness. The initial and updated fitness values of every individual are compared to elect individual solutions from the population for the next generation. Check the termination criteria (i.e., the maximum number of generations) at the end of every iteration and finalize the algorithm process by selecting the best individual with an optimal subset of features having the highest weights. The features having higher weights will have more efficiency in dealing with machine learning models. The key contributions of the proposed method are,

a) To improve the classification performance by incorporating the RNCA model with BTLBO for considering the importance of individual features.
b) Minimize the dimension of features to construct a classification model for predicting classes.
c) Avoid the overfitting problem of a classification model using the regularized NCA.

To showcase the above attainments, various experiments have been conducted using the medical dataset for classifying benign and malignant tumors. The results are compared with various existing related feature selection methods and presented in the form of graphs and tables in the next section.

4. Results and Discussion

The BTLBO-RNCA model is implemented for feature selection and assessed using different classification measures like accuracy (or Error rate), precision, recall, F-Measure, and AUC values. The feature weights of original attributes, RNCA selected attributes and BTLBO-RNCA based attributes are presented in graphs.

4.1. Experimental Setup and Dataset

Breast cancer dataset WDBC of the UCI repository [28] is used to validate the proposed FS method. The dataset has 569 instances in which 212 are having a malignant tumor and 357 are having benign tumor patient records. The features are extracted from digital images of breast mass and each one has 30 real-valued multivariate attributes. The authors [15] compared different classification models k-NN and feature reduction techniques PCA to diagnosis breast cancer using WDBC dataset. They evaluated the performance with average accuracy which is computed in multiple numbers of folds. The RNCA learning model has been used as a classifier as well as a feature selection method to evaluate the performance of the proposed FS model. Cross-validation is performed with 80:20 partition ratio for training and testing the proposed BTLBO-RNCA FS model on the dataset.

Coimbra breast cancer dataset [14] also has been used to test the performance of the hybrid model. The dataset has 116 instances with 64 breast cancer patient records and 52 normal patient records. Instances are collected from the regular blood samples and each record has 10 attributes which predict cancer disease. This dataset also has been divided into train and test subsets with 93 and 23 instances, respectively.

4.2. Performance Evaluation of the BTLBO-RNCA Model

The hybrid method is evaluated by computing various performance measures of a learning model that is trained with selected features. The same metrics have been measured with overall features for performance comparison. Table 1 shows the important classifier performance measures. The accuracy of a learning model can be measured as the ratio of the number of correct predictions to the total predictions made by a classifier for a test dataset. The precision is the ratio of actual correct versus total positive recognitions and recall is correctly identified versus actual positives in the test set. F-Measure can be computed using recall and precision as a weighted harmonic mean. The Receiver

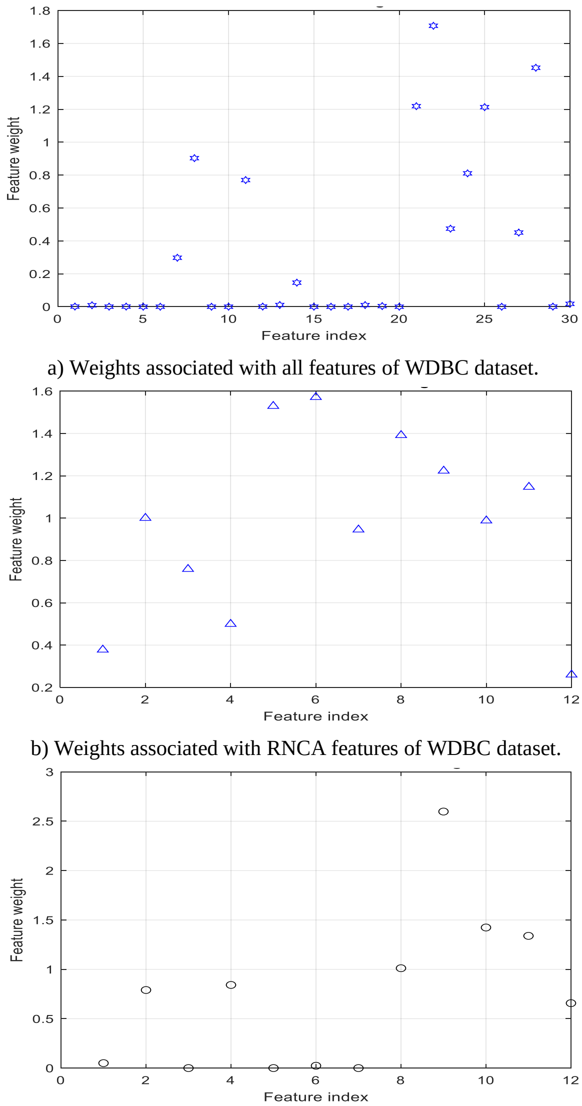
Operating Characteristic (ROC) curve measures the performance of a classification model with respect to two parameters, called the true positive rate and the false positive rate. AUC measures the range below the ROC curve. The AUC value ranges from 0 (indicating 100% wrong predictions) to 1 (indicating 100% correct predictions).

Table 1. The Performance measures of BTLBO-RNCA model with WDBC dataset.

Classification Measures	**Overall Features**	**BTLBO-RNCA Selected Features**
Accuracy	96.46	99.12
Error Rate	3.53	0.88
Precision	0.98	0.99
Recall	0.904	0.976
F_Measure	0.95	0.988
AUC	0.996	0.997

The classifier achieved an accuracy of 96.46 with complete features set and 99.12 with the features selected by the hybrid model. Error rates are measured to validate the accuracy metric of the learning model. The precision value of the classifier is greater than 0.9 for both the total and optimal subset of features. The remaining measures are above the value of 0.9 (Maximum value is 1) for both the cases. The higher accuracy (or the low error rate) of the learning model proved the performance improvement of the proposed hybrid technique for optimal feature selection. Better values are achieved by Recall and F_Measure metrics and comparable results are produced with Precision and AUC values. The classifier produced better performance in terms of accuracy with the optimal subset of features. But accuracy itself is not enough to decide the performance of a learning model and need to consider the remaining measures to finalize the results. The hybrid FS model also achieved comparable results for recall and F-Measure with the classifier.

The contribution of features in building a learning model for classification is characterized by feature weights using the NCA model. Figure 2-a) represents a graph plotted between the overall features and the weights associated with the attributes by the NCA model. In the graph, X-axis maintains the feature index and Y-axis indicates the weights associated with the attributes. Here, the features associated with weights greater than zero will influence the efficiency of a learning model. The remaining features with weight values closer to zero will not contribute to the results of a learning model and increase the complexity and training time. Figures 2-b) and 2-c) give the association of weights distributed to the features selected by RNCA and the proposed BTLBO-RNCA FS methods. Both RNCA and hybrid models selected 12 out of 30 features having weights greater than zero.



a) Weights associated with all features of WDBC dataset.

b) Weights associated with RNCA features of WDBC dataset.

c) Weights associated with BTLBO-RNCA features of WDBC dataset.

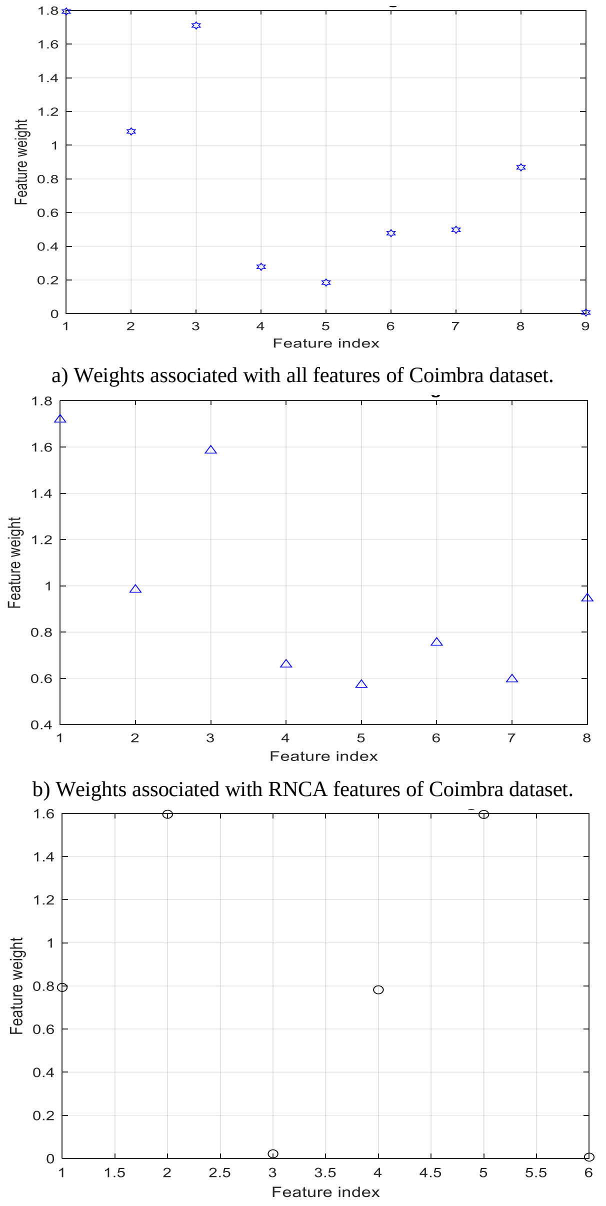
Figure 2. Comparison of weights related to the features of WDBC dataset.

Table 2 shows the performance measures of the proposed model with the Coimbra breast cancer dataset. The classifier accomplished an accuracy of 73.91 with complete features set and 99.3 with the features selected by the proposed hybrid model. Precision, recall and F-Measure values of the classifier are 0.71, 0.83 and 0.77 with overall attributes and 0.91 for optimal subset of features. The large difference in the accuracy between the models trained with all features and optimal subset of features proved the importance of new hybrid models with the best combination algorithms for feature selection. A small improvement from 0.9 to 0.98 is noticed in the AUC value of the classifier.

Table 2. The Performance measures of BTLBO-RNCA model with coimbra breast cancer dataset.

Classification Measures	**Overall Features**	**BTLBO-RNCA Selected Features**
Accuracy	73.91	91.3
Error Rate	26.09	8.67
Precision	0.71	0.91
Recall	0.83	0.91
F_Measure	0.77	0.91
AUC	0.90	0.98

The classifier achieved good results using the proposed hybrid FS method with the Coimbra dataset having only a few attributes for deciding the type of instance. Figure 3 presents the weights given to the attributes in the dataset by the classifier with the help of various FS methods. The classifier assigned almost similar weights for the complete attributes and RNCA selected attributes. Major variations in weights are given by the classifier using the proposed method assigned. RNCA FS method assigned and eliminated the least weight to the last attribute. But BTLBO-RNCA selected the last attribute by assigning some weight along with the other attributes.



a) Weights associated with all features of Coimbra dataset.

b) Weights associated with RNCA features of Coimbra dataset.

c) Weights associated with BTLBO-RNCA features of coimbra dataset.

Figure 3. Comparison of weights related to the features of WDBC dataset.

4.3. Comparison of Performance Measures of Feature Selection Algorithms

The results achieved by the individual features selection methods (RNCA and BLTBO) are compared with the results of the proposed model. The performance of FS methods is evaluated for WDBC dataset using different measures as shown in Table 3. The hybrid BTLBO-RNCA feature selection method achieved best accuracy (99.12) or least error rate and better recall (0.97) values when compared with the remaining FS methods. The proposed model also achieved comparative results of F-measure and AUC with the classifier. The precision value is greater than 0.9 for all the feature selection methods.

Table 3. Classification performance measures of WDBC dataset.

Classification Measures	**Selected Features**		
	RNCA	**BTLBO**	**BTLBO-RNCA**
Accuracy	97.34	98.23	99.12
Error Rate	2.66	1.77	0.88
Precision	0.98	0.99	0.99
Recall	0.928	0.952	0.976
F_Measure	0.962	0.975	0.988
AUC	0.997	0.976	0.997

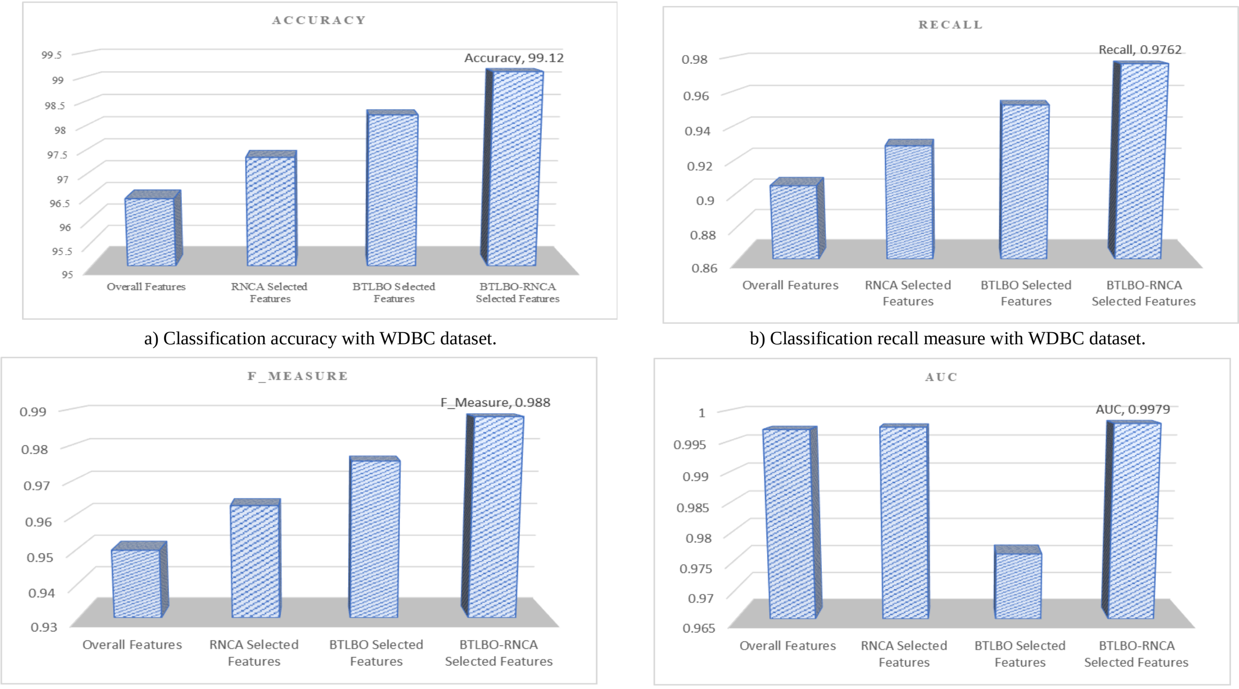
The attributes count is presented in Table 4 for the specified FS methods. The table also includes the indexes of selected features in the records of the dataset. The proposed method selected 12 features out of 30 to accomplish the best performance with the classification model.

Table 4. Feature count and corresponding indexes of WDBC dataset.

Feature Selection Methods	**Features Count**	**Feature Indexes**
RNCA	12	7 8 11 14 21 22 23 24 25 27 28 30
BTLBO	11	5 6 7 10 11 12 17 21 25 27 30
BTLBO-RNCA	12	1 2 9 11 12 13 18 22 24 25 27 29

Improvements in the performance measures of a classifier with the proposed model are represented using graphs as shown in Figure 4. Considerable increment is observed in the accuracy, recall and F_Measure values of various FS methods from the graph. The AUC value of the BTLBO FS is just behind the remaining FS methods.

The curves are plotted between false positive rates and true positive rates (ROC curves) for the features selected by RNCA, BTLBO-RNCA algorithms, and overall features. All curves are closer to corner indicating proportional performance with the AUC value of 0.99 for WDBC dataset as shown in Figure 5.



a) Classification accuracy with WDBC dataset.

b) Classification recall measure with WDBC dataset.

c) Classification f_Measure measure with WDBC dataset.

d) Classification AUC measure with WDBC dataset.

Figure 4. The Graphical comparison of classification measures of WDBC dataset for various feature selection methods.

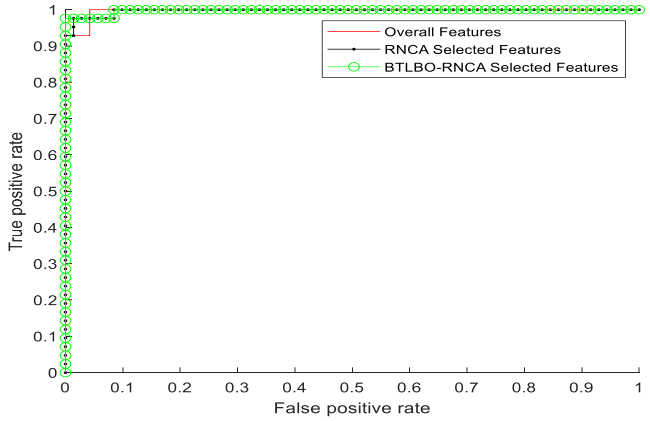


Figure 5. The ROC curves for Overall, RNCA, and BTLBO-RNCA selected features of WDBC dataset.

Different metrics of the proposed FS model are compared with the other FS methods using Coimbra BC dataset. The performance measures of the learning model constructed with Coimbra dataset using RNCA and BTLBO FS methods are shown in Table 5. The classifier attained an accuracy of 73.91 and 78.26 with total features and RNCA selected features. BTLBO improved the accuracy to 86.96 with its teaching and learning based search strategy. The current hybrid model achieved an accuracy of 91.3 with the combined performance of RNCA and BTLBO algorithms. The attributes selected by the RNCA FS method minimally improved the efficiency of the classifier in terms of other measures. BTLBO FS succeeds in improving the efficiency of the learning model except for AUC value (0.87). All measures of BTLBO-RNCA FS, including AUC (0.98) justified the hybrid proposal for feature selection.

The count of attributes selected by various FS methods with the best classification value is represented in Table 6. The table also includes the indexes of selected features from the dataset. The proposed method selected 6 features to accomplish the best performance with the classification model. RNCA FS eliminated only one attribute to form a subset for training the classifier. But BTLBO performed well by selecting minimal attributes and achieved comparable results with the classifier.

Graphs in Figure 6 give a comparative analysis of various performance measures with the specified (Table 5) FS methods. The bars in the first graph show the accuracy percentages in the y-axis. There is a significant increment in the accuracy values (73.91, 78.26, 86.96, and 91.3) by the classifier with the FS methods. The classifier performed well in terms of recall and F_Measure values with the hybrid and BTLBO FS methods. The value (0.98) of AUC placed the proposed FS method at the top position for all the measures of the classifier. ROC curves are drawn for every mentioned FS method in Figure 7. Significant performance differences of the classifier can be observed for various FS methods with ROC curves presented in the graph.

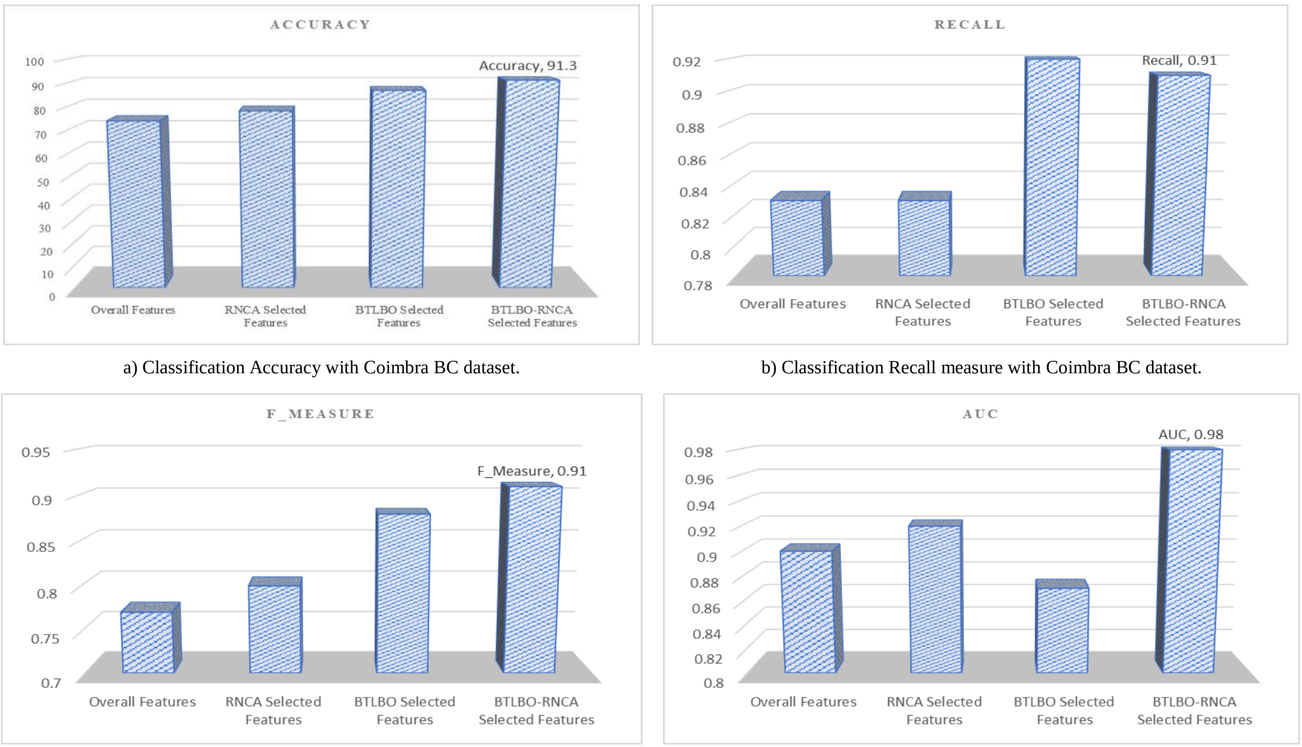
Table 5. Classification performance measures of coimbra dataset.

Classification Measures	Selected Features		
	RNCA	BTLBO	BTLBO-RNCA
Accuracy	78.26	86.96	91.3
Error Rate	21.74	13.04	8.7
Precision	0.77	0.85	0.91
Recall	0.83	0.92	0.91
F_Measure	0.8	0.88	0.91
AUC	0.92	0.87	0.98

Table 6. Feature count and related indexes of Coimbra dataset.

Feature Selection Methods	Features Count	Feature Indexes
RNCA	8	1 2 3 4 5 6 7 8
BTLBO	4	1 2 3 8
BTLBO-RNCA	6	2 3 5 6 8 9

The curve generated by the proposed BTLBO-RNCA method occupied the top position and covered the complete area with a value AUC of 0.98. The curve for BTLBO has deviated initially with the curves of other FS methods.



a) Classification Accuracy with Coimbra BC dataset.

b) Classification Recall measure with Coimbra BC dataset.

c) Classification F_Measure measure with Coimbra BC dataset.

d) Classification AUC measure with Coimbra BC dataset.

Figure 6. Graphical comparison of classification measures for Coimbra BC dataset using various feature selection methods.

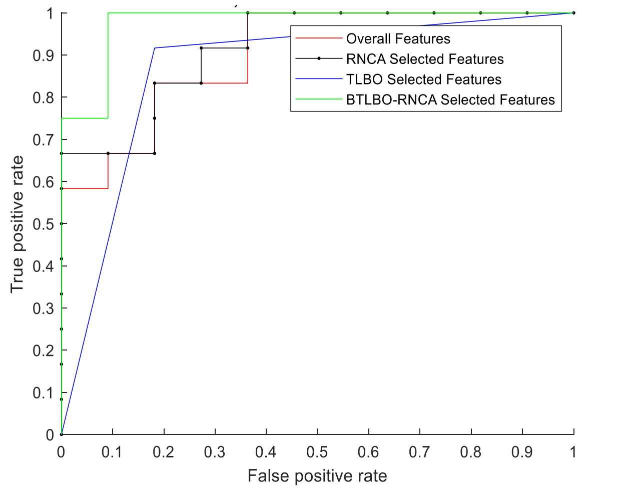


Figure 7. The ROC curves for Overall, RNCA and BTLBO-RNCA selected features of coimbra dataset.

From the above statistics we have discovered that the hybrid feature selection method (BTLBO-RNCA) improved the classification accuracy measures with the combination of best features. These results will be beneficial for medical experts to diagnose breast cancer with more accuracy. This feature selection method could be applied to the medical datasets related to other diseases for better classification results. This research work confirms that the efficiency of feature selection process is enhanced by the power evolutionary search method with respect to the individual feature weights.

5. Conclusions

In this paper, a new hybrid feature selection model was proposed with the help of binary TLBO and RNCA algorithms. In this hybrid model, the BTLBO optimization algorithm played a significant role in searching for best possible combination of features from the dataset with the emphasis on the relevance of the features to overcome the overfitting problems. The RNCA estimated the significance of features in the form of weights to evaluate the fitness of individuals during the feature selection process. The BTLBO-RNCA model improved the efficiency of the learning model by means of reduced optimal features with the WDBC and Coimbra datasets for diagnosing breast cancer. The classification measures of the new hybrid FS method are compared with the other related FS methods to validate the performance and proved significant improvements in the results with both the datasets. We would like to extend our work with the meta optimization approach to tune the internal parameters of the feature selection algorithm in the future to improve performance further.

References

[1] Ahmad S., Bakar A., and Yaakub M., "Ant Colony Optimization for Text Feature Selection in Sentiment Analysis," *Intelligent Data Analysis*, vol. 23, no. 1, pp. 133-158, 2019. DOI: 10.3233/IDA-173740

[2] AlFarraj O., AlZubi A., and Tolba A., "Optimized Feature Selection Algorithm Based on Fireflies with Gravitational Ant Colony Algorithm for Big Data Predictive Analytics," *Neural Computing and Applications*, vol. 31, pp. 1391-1403, 2019. DOI:10.1007/s00521-018-3612-0

[3] Ali A., Hussain Z., and Abd S., "Big Data Classification Efficiency Based on Linear Discriminant Analysis," *Iraqi Journal for Computer Science and Mathematics*, vol. 1, no. 1, pp. 7-12, 2020. DOI: https://doi.org/10.52866/ijcsm.2019.01.01.001

[4] Allam M. and Nandhini M., "Feature Optimization Using Teaching Learning Based Optimization for Breast Disease Diagnosis," *International Journal of Recent Technology and Engineering*, vol. 7, no. 4, pp. 78-85, 2018.

[5] Allam M. and Nandhini M., "Optimal Feature Selection Using Binary Teaching Learning Based Optimization Algorithm," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 2, pp. 329-341, 2018. https://doi.org/10.1016/j.jksuci.2018.12.001

[6] Allam M. and Nandhini M., "A Study on Optimization Techniques in Feature Selection for Medical Image Analysis," *International Journal on Computer Science and Engineering*, vol. 9, no. 3, pp. 75-82, 2017.

[7] Goldberger J., Hinton G., Roweis S., and Salakhutdinov R., "Neighbourhood Components Analysis," *Advances in Neural Information Processing Systems*, vol. 17, pp. 513-520, 2005.

[8] Guha R., Ghosh M., Kapri. S., Shaw S., Mutsuddi S., Bhateja V., and Sarkar R., "Deluge based Genetic Algorithm for Feature Selection," *Evolutionary Intelligence*, vol. 14, pp. 357-367, 2021.

[9] Hsu H., Hsieh C., and Lu M., "Hybrid Feature Selection by Combining Filters and Wrappers," *Expert Systems with Applications*, vol. 38, pp. 8144-8150, 2011.

[10] Jain S. and Salau A., "An Image Feature Selection Approach for Dimensionality Reduction Based on kNN and SVM for AkT Proteins," *Cogent Engineering*, vol. 6, no. 1, 2019. https://doi.org/10.1080/23311916.2019.1599537

[11] Khourdifi Y. and Bahaj M., "The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart Disease," ICCWCS, 2019. http://dx.doi.org/10.4108/eai.24-4-2019.2284088

[12] Kiziloz H., Deniz A., Dokeroglu T., and Cosar A., "Novel Multiobjective TLBO Algorithms for the Feature Subset Selection Problem," *Neurocomputing*, vol. 306, pp. 94-107, 2018. https://doi.org/10.1016/j.neucom.2018.04.020

[13] Liang H., Wang Z., and Liu Y., "A New Hybrid Ant Colony Optimization Based on Brain Storm Optimization for Feature Selection," *The Institute of Electronics, Information and Communication Engineers*, vol. 102, no. 7, pp. 1396-1399, 2019. DOI:10.1587/transinf.2019EDL8001

[14] Patricio M., Pereira J., Crisostomo J., Matafome P., Gomes M., Seica R., Caramelo F., "Using Resistin Glucose Age and BMI to Predict the Presence of Breast Cancer," *BMC Cancer*, vol. 18, no. 1, 2018. doi: 10.1186/s12885-017-3877-1.

[15] Qiu Y., Zhou G., Zhao Q., and Cichocki A., "Comparative Study on the Classification Methods for Breast Cancer Diagnosis," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 66, no. 6, pp. 841-848, 2018. DOI: 10.24425/bpas.2018.125931

[16] Ramasamy R. Rani S., "Modified Binary Bat Algorithm for Feature Selection in Unsupervised Learning," *The International Arab Journal of Information Technology*, vol. 15, no. 6, pp. 1060-1067, 2018.

[17] Rajalaxmi R., "A Hybrid Binary Cuckoo Search and Genetic Algorithm for Feature Selection in Type-2 Diabetes," *Current Bioinformatics*, vol. 11, no. 4, pp. 490-499, 2016. DOI: 10.2174/1574893611666151228190309

[18] Rao R., "Review of Applications of TLBO Algorithm and a Tutorial for Beginners to Solve the Unconstrained and Constrained Optimization Problems," *Decision Science Letters*, vol. 5, pp. 1-30, 2016. DOI:10.5267/j.dsl.2015.9.003

[19] Rao R., Savsani V., and Vakharia D., "Teaching-Learning-based Optimization: A Novel Method for Constrained Mechanical Design Optimization Problems," *Computer-Aided Design*, vol. 43, no. 3, pp. 303-315, 2011. https://doi.org/10.1016/j.cad.2010.12.015

[20] Satapathy S., Naik A., and Parvathi K., "Rough set and Teaching Learning Based Optimization Technique for Optimal Features Selection," *Central European Journal of Computer Science*, vol. 3, no. 1, pp. 27-42, 2013. DOI: 10.2478/s13537-013-0102-4

[21] Satapathy S., Naik A., and Parvathi K., "Unsupervised Feature Selection Using Rough Set and Teaching Learning-Based Optimisation," *International Journal of Artificial Intelligence and Soft Computing*, vol. 3, no. 3, pp. 244-256, 2013. DOI:10.1504/IJAISC.2013.053401

[22] Sevinç E. and Dökeroğlu T., "A Novel Hybrid Teaching-Learning-Based Optimization Algorithm for the Classification of Data by Using

Extreme Learning Machines," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, pp. 1523-1533, 2019. 10.3906/elk-1802-40

[23] Shang Q., Tan D., Gao S., and Feng L., "A Hybrid Method for Traffic Incident Duration Prediction Using BOA-Optimized Random Forest Combined with Neighborhood Components Analysis," *Jounal of Advanced Transportation*, 2019. https://doi.org/10.1155/2019/4202735.

[24] Sun L., Kong X., Xu J., Xue Z., Zhai R., and Zhang S., "A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tongumor Classification," *Scientific Reports*, vol. 9, no. 1, 2019. doi: 10.1038/s41598-019-45223-x.

[25] Taghanaki S., Ansari M., Dehkordi B., and Mousavi S., "Nonlinear Feature Transformation and Genetic Feature Selection: Improving System Security and Decreasing Computational Cost," *ETRI Journal*, vol. 34, no. 6, pp. 847-857, 2012. https://doi.org/10.4218/etrij.12.1812.0032

[26] Too J., Abdullah A., and Saad N., "Binary Competitive Swarm Optimizer Approaches for Feature Selection," *Computation*, vol. 7, no. 2, 2019.

[27] Tuo S., Yong L., Deng F., Li Y., Lin Y., and Lu Q., "HSTLBO: A Hybrid Algorithm based on Harmony Search and Teaching-Learning Based Optimization for Complex High Dimensional Optimization Problems," *PLoS ONE12*, vol. 12, no. 4, 2017. https://doi.org/10.1371/journal.pone.0175114

[28] Wolberg W., Mangasarian O., Street N., and Street W., UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

[29] Yang W., Wang K., and Zuo W., "Neighborhood Component Feature Selection for High Dimensional Data," *Journal of Computers*, vol. 7, pp. 161-168, 2012. doi: 10.4304/jcp.7.1.161-168

[30] Yang Z. and Laaksonen J., "Regularized Neighborhood Component Analysis, Image Analysis," *Lecture Notes in Computer Science*, vol. 4522, 2007.

[31] Zhao Y., Liu Y., and Huang W., "Prediction Model of HBV Reactivation in Primary Liver Cancer-Based on NCA Feature Selection and SVM Classifier with Bayesian and Grid Optimization," *in Proceedings of the IEEE 3rd International Conference on Cloud Computing and Big Data Analysis*, Chengdu, pp. 547-551, 2018. 10.1109/ICCCBDA.2018.8386576

Mohan Allam holds a Ph.D. in Computer Science & Engineering from Pondicherry University and has over 15 years of experience in teaching in higher education institutions. Currently, works as an Assistant Professor at Lovely Professional University. He has mentored students on various research projects and provided guidance in selecting the best approach for their projects. The author's area of research includes Soft Computing, Image Processing, and IoT. The author has published several peer-reviewed journal articles and presented papers at various national and international conferences. He has also served as a reviewer for several international journals and conferences in the field of Computer Science and Engineering.

Nandhini Malaiyappan is an Associate Professor in the Department of Computer Science at Pondicherry University in Puducherry, India. Her research interests include Artificial Intelligence, Software Engineering, Evolutionary Algorithms, and Combinatorial Problem Optimization. She has published several research papers in various national and international conferences and journals. She has supervised several research projects and guided several Ph.D. and M.Tech. students in their research work. She has received grants from various funding agencies to support her research work. She has also served as a reviewer for several international journals and conferences in the field of Computer Science and Engineering.