International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Probit Regressed Feature Selection Based Linear Programming Boost Classification for Tumor Risk Factor Identification and Disease Diagnosis

P. S. Renjeni, B. Mukunthan

Abstract - Accurate diagnosis of survival rate in patients with tumor remains challenges due to the increasing complexity of treatment protocols, and different patient population samples. Due to the complexity, the risk factor of the patients gets increased. Therefore, a reliable and well-validated prediction needs to develop the automatic disease diagnosis for early detection of the tumor. The novel technique called Probit Regressed Feature Selection based Iterative Linear Programming Boost Classification (PRFS-ILPBC) is introduced for tumor risk factor identification and disease diagnosis of patient data with higher accuracy and lesser time consumption. In PRFS-ILPBC technique, Probit Regression model is a regression type to estimate the relationship between the features and the disease symptoms using bivariate correlation coefficient. Based on the correlation results, the features fall into any one of the two classes (i.e. relevant or irrelevant). With the help of relevant feature, Iterative Linear Programming (LP) Boost Classification model is applied to perform classification by combining the weak learner for tumor risk factor identification and disease diagnosis. LPBoost constructs the strong classifier through initiating with a set of weak classifiers. The training data (i.e. patient data) are taken as the input and added to the set of considered weak classifiers. The kernelized support vector machine act as weak learner compares the training features with the testing results to identify the risk factor and classify the patient data into normal or abnormal. The ensemble classifier improves disease diagnosis accuracy and reduces the false positive rate. Experimental evaluation of proposed PRFS-ILPBC technique and existing methods are carried out with different factors such as disease diagnosing accuracy, false positive rate, and computation time with respect to a number of patient data. The observed results reported that the proposed PRFS-ILPBC technique achieves higher disease diagnosing accuracy with minimum computation time as well as false positive rate than the conventional techniques.

Keywords: tumor disease diagnosis, feature selection, Probit Regression, bivariate correlation coefficient, Iterative Linear Programming Boost Classification, kernelized support vector machine

I. INTRODUCTION

In health monitoring, the human body comprises the many types of cells. Every cell performs a set of a specific function. These cells help to maintain the human body in the healthy and properly working. When a few cells lose their ability to control their growth, the additional cells formed a mass of tissue which is called a tumor.In the health care industry, a large volume of data presented and it is difficult to discover the necessary pattern with minimum complexity. In order to minimize the complexity in disease diagnosis, feature selection is performed to help in the important tasks of analyzing the risk factors.

Feature selection is the process of finding the subset of features using different approaches to minimize the prediction error as well as improve the accuracy. Several data mining and machine learning techniques are developed for tumor disease diagnosis.

A Radial basis function neural network (RBFNNs) with Artificial Bee Colony was introduced in [1] to categorize the EEG signal for detecting epileptic seizures. The designed classification technique was not minimized the epileptic seizures detection time. A complex-valued neural network (CVANN) was developed in [2] to analyze the EEG signals for epilepsy diagnostic purposes. But it failed to achieve higher accuracy while implementing the larger and more continuous data.

A novel ensemble classifier was developed in [3] to identify an epileptic seizure from compressed and noisy EEG signals with lesser complexity. The classification accuracy was not improved. A fuzzy classifier was introduced in [4] to detect the normal EEG signal from epileptic EEG signal. The classifier performance was not enhanced since it failed to select related features.

A k- nearest neighbor (kNN) classifier was introduced in [5] for identifying the seizure activity across the various regions of the brain automatically. The designed classifier failed to detect the seizure activity using big dataset.

A multilayer perceptron neural network (MLPNN) classifier was developed in [6] for accurately diagnosis the epileptic seizures with minimum time. But it failed to minimize the false positive rate. An ensemble classification techniques were developed in [7] for brain tumor disease diagnosis using feature selection. The techniques failed to identify the disease with minimum time.

A global maximal information coefficient (MIC) method was introduced in [8] to precisely distinguish the seizure activities using EEG signal. But the designed method failed to improve the accuracy of seizure activity detection. A Long Short-Term Memory (LSTM) networks using deep convolutional neural networks (CNN) was developed in [9] for epileptic seizure prediction using EEG signals.

Manuscript published on 30 September 2019
* Correspondence Author
P. S. Renjeni*, Research Scholar, Jairams Arts & Science College, Karur-3.
Dr. B. Mukunthan, Research Supervisor, Jairams Arts & Science College, Karur-3.

The method failed to perform the epileptic seizure prediction with more EEG data. An optimum allocation technique with logistic model trees (LMT) was introduced in [10] to identify the epileptic seizure from EEG signals. But the performance of the time complexity remained unsolved. The most significant issues in the tumor disease diagnosis are identified from the existing literature are lesser diagnosing accuracy, more time complexity, high false positive rate, and so on.

Such kinds of issues are overcome by introducing a novel technique called PRFS-ILPBC.

The contribution of PRFS-ILPBC technique is summarized as follows.

- To improve the disease diagnosing accuracy, PRFS-ILPBC technique uses the Iterative LPBoost classification. The PRFS-ILPBC technique initially constructs the Kernelized support vector classifier to categorize the patient data as a normal or abnormal using hyperplane. The weak classifier results are combined to make a strong for achieving the accurate classification results.
- To minimize the false positive rate, Iterative LPBoost Classification combines a set of weak learners and calculates the training error. The weight of the weak classifier is updated based on the training error value. The ensemble classifier finds the weak classifier with minimum error.
- To minimize the computation time, PRFS-ILPBC technique uses the probit regression function. The regression analyzes the features and disease symptoms using bivariate correlation. The correlation results are used to identify the relevant features and irrelevant features. The relevant features are selected for patient data classification.

The rest of the paper is arranged into five different sections. Section 2 describes the related works in the field of tumor disease diagnosis. In Section 3, the proposed PRFS-ILPBC technique is explained with the neat architecture diagram. In Section 4, experimental evaluation is performed with EEG dataset. The observed results are discussed with different parameters in Section 5. Section 6 provides the conclusion.

II. RELATED WORKS

A convolutional neural network (CNN) was introduced in [11] for classifying the EEG signals. The performance of CNN was not improved by increasing the number of samples. A supervised machine learning model was developed in [12] for classifying the seizure and nonseizure records. But the model failed to accurately minimize the training error in the classification.

A feature-based linear topological classifier was introduced in [13] for identifying the epileptic seizures. The designed classifier failed to perform the epileptic seizures detection with minimum time. A random forest classifier was developed in [14] to identify the epileptic seizure from the EEG signal. Though the technique selects the relevant feature, the time complexity was not minimized. A novel machine learning classification was developed in [15] using EEG–ECG features to detect the increased risk of epileptic seizures. But the more relevant feature selection was not performed. A least square support vector machine (LS_SVM) classifier was introduced in [16] for categorizing the EEG signals using feature selection. The designed classifier failed to achieve higher accuracy. A hybrid method uses the support vector machines for the classification of the medical data was developed in [17] using feature selection. But the false positive rate was not minimized in the medical data classification.

A deep learning approach was introduced in [18] for automatic identification of epileptic seizures with EEG signals. The approach improves the detection accuracy but it failed to minimize the time. A nonlinear sparse extreme learning machine (SELM) was developed in [19] to identify the epileptic seizure. Though the SELM minimizes computational complexity, accurate detection was not performed. A k-nearest neighbors (k-NN) algorithm was designed in [20] to classify epilepsy at various ages of the patients. But the feature selection was not performed to minimize the time complexity.

In this paper, PRFS-ILPBC technique is introduced for addressing the existing problems identified from the above-said literature. The brief description of the PRFS-ILPBC technique is presented in the next section.

III. Probit Regressed Feature Selection based Iterative Linear Programming Boost Classification for tumor disease diagnosis

A Probit Regressed Feature Selection based Iterative Linear Programming Boost Classification (PRFS-ILPBC) technique is introduced to improve the tumor disease diagnosis accuracy with minimum time complexity. The complexity of the disease diagnosis is minimized by selecting the less number of features from the dataset. A Feature selection is a process of selecting the attributes from the tumor dataset. In real-world applications, a large number of features present in the training data set may enhance the risk level for disease prediction. This also contains the high dimensional feature space. Therefore, feature selection is significant for dimensionality reduction. Dimensionality reduction is the process of removing the irrelevant or redundant features and selects the most significant features (i.e. risk factors). The risk factor influences the severity of tumor disease for early diagnosis. The main benefits of feature selection are to minimize the computation time and improve the performance of training sample sets. The architecture diagram of the PRFS-ILPBC technique is shown in figure 1.

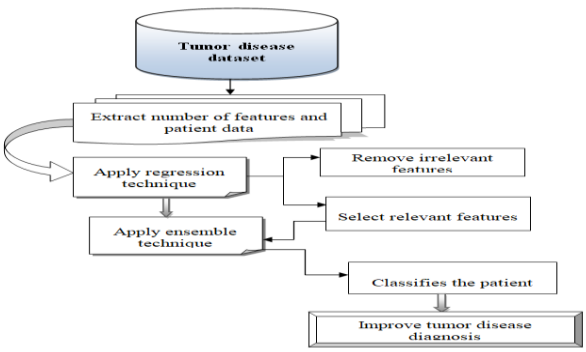

Figure 1 architecture of PRFS-ILPBC technique

Figure 1 depicts the architecture of the proposed PRFS-ILPBC technique to obtain better disease diagnosing results with minimum time complexity. Initially, patient data $d_1, d_2, d_3, \ldots d_n$ and features $f_1, f_2, f_3, \ldots f_n$ are extracted from the dataset. The PRFS-ILPBC technique performs two major processes namely feature selection and classification. The feature selection is the process of selecting the subset of the features (i.e. risk factors) with the help of the regression analysis. Followed by, the classification is performed with the selected relevant features using iterative linear program boost ensemble classifier. The classification results clearly shows that the PRFS-ILPBC technique effectively diagnosing the tumor disease with higher accuracy. These two process of the PRFS-ILPBC technique is clearly described in the below subsection.

1.1 Probit regression based feature selection

The first process in the PRFS-ILPBC technique is the feature selection is also called as attribute selection for disease diagnosis and risk factor analysis. Feature selection is very effective for reducing the dimensionality by removing the irrelevant features. The PRFS-ILPBC technique uses the Probit regression and the word is coming from a probability + unit (Probit). Probit regression analysis is a statistical method that helps to evaluate the relationship between two variables such as independent (i.e. features f) and dependent variable (i.e. outcomes Y) where the dependent variable takes two results (i.e., relevant or irrelevant). The objective of the Probit Regression model is to estimate the correlation of features for disease diagnosis where the probability results fall into any one of these two classes (1 or 0). The bivariate correlation is used to find the relationship between the objective (i.e. tumor disease symptoms) and features in the given dataset. The bivariate correlation is estimated as follows,

$$\rho = \frac{\sum f_i s_t - \frac{\sum f_i * \sum s_t}{n}}{\sqrt{\left(\sum {f_i}^2 - \frac{(\sum f_i)^2}{n}\right)} * \sqrt{\left(\sum {s_t}_2{}^2 - \frac{(\sum {s_t}_2)^2}{n}\right)}} \quad (1)$$

In (1), ρ denotes a correlation coefficient, n denotes a number of features, f_i denotes a features, s_t denotes a tumor disease symptoms, $\sum f_i s_t$ refers to the sum of the product of paired score, $\sum f_i$ is the sum of f_i score, $\sum s_t$ is the sum of s_t score, $\sum {f_i}^2$ is the sum of the squared score of f_i and $\sum {s_t}^2$ is the sum of the squared score of s_t. The probit regression analyses the correlation and returns the value for selecting the relevant features and removing the irrelevant features.

$$P(y|f_i) = \begin{cases} 1 & ; positive\ correlation \\ 0 & ; negative\ correlation \end{cases} \quad (2)$$

In (2),P denotes a probability of the classes, y denotes an output classes (i.e. relevant and irrelevant), f_i denotes a features. The regression function returns '1' when the coefficient provides the positive correlation and the probability is '0' when the coefficient provides the negative correlation. The positive correlation provides the higher relationship between the feature and tumor disease symptoms. Whereas the negative correlation indicates there is no relationship between a feature and the tumor disease symptoms. Therefore, positive correlation falls into the relevant class whereas the negative correlation falls into the irrelevant class. In this way, the proposed PRFS-ILPBC technique selects the relevant features and removes the irrelevant features from the dataset. The feature selection minimizes the time complexity in the tumor disease diagnosis. The algorithmic process of the feature selection is described as follows,

Input: tumor disease dataset D_t , features $f_1, f_2, f_3, \ldots . f_n$
Output: Select relevant features and remove irrelevant features
Begin
1. **Extract the** number of features $f_1, f_2, f_3, \ldots . f_n$
2. Measure correlation ρ
3. **If** $(\rho = +1)$ **then**
4. Positive correlation
5. $P(y|f_i)$ returns '1'
6. Select relevant features
7. **else**
8. Negative correlation
9. $P(y|f_i)$ returns '0'
10. Remove irrelevant features
11. **end if**

end

Algorithm 1 probit regression based feature selection

Algorithm 1 describes the step by step process of the feature selection for achieving the higher diagnosing accuracy with minimum time complexity. Initially, the numbers of features are extracted from the dataset. Then the probit regression function finds the correlation between the features and tumor disease symptoms.

The positive correlation is used for selecting the relevant features. Otherwise, the features are said to be an irrelevant feature. The relevant features are selected for tumor disease diagnosis and the irrelevant features are removed. This helps to minimize the disease diagnosing time.

1.2 Iterative Linear Programming Boost Classification for disease diagnosis

After selecting the relevant features, the classification is performed using Iterative Linear Programming Boost technique with patient data. Boosting is a machine learning ensemble algorithm that converts weak learners into strong ones for achieving higher classification accuracy. A weak learner is a classifier that failed to provide true classification. On the contrary, a strong learner is also a classifier that provides accurate classification results by combining the weak learner. The structure of the ensemble classification algorithm is shown in figure 2.

Figure 2 shows the structure of the ensemble classification algorithm. The ensemble classifier considers the training samples $\{x_i, y_i\}$ where x_i denotes a number of patient data (i.e. input) and y_i denotes an output results.

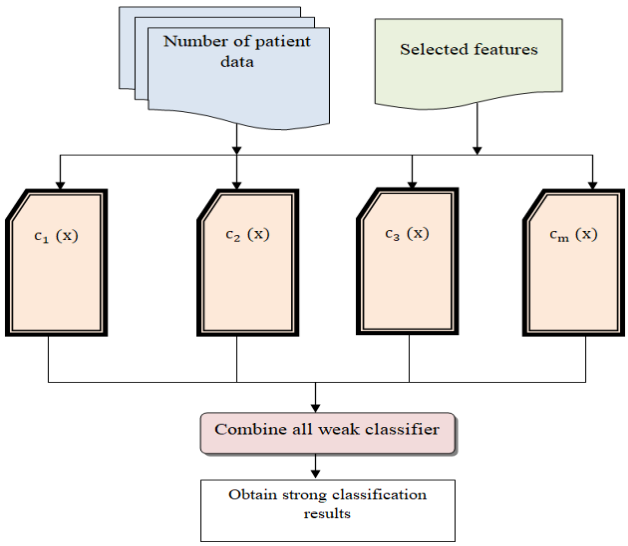

Figure 2 structure of the ensemble classification algorithm

Initially, patient data is given to the weak learners $c_1(x), c_2(x), c_3(x), \ldots . c_n(x)$. The ensemble classifier uses the kernelized support vector classifier for categorizing the patient data into different classes using optimal hyperplane. The hyperplanes is a decision boundary between the classes. The kernelized support vector categorizes the patient data on either side of the decision boundary by analyzing the risk factors with the testing results. A separating hyperplane (h) is defined as follows,

$$\vec{\alpha}.d_i + \vec{v} = 0 \quad (3)$$

In (3),$\vec{\alpha}$ represents the normal weight vector to the hyperplane, d_i denotes a training samples (i.e. Patient data), $\vec{v}$ denotes a bias. If the training samples are linearly separable, two marginal hyperplanes are selected to separate the patient data into two classes.

$$h_1 \rightarrow \vec{\alpha}.d_i + \vec{v} > 0 \quad (4)$$
$$h_2 \rightarrow \vec{\alpha}.d_i + \vec{v} < 0 \quad (5)$$

From (4) and (5), h_1, h_2 denotes lower and upper marginal hyperplanes to classify the patient data into above and below the boundary. The classification result of the kernelized support vector classification is shown in figure 3.

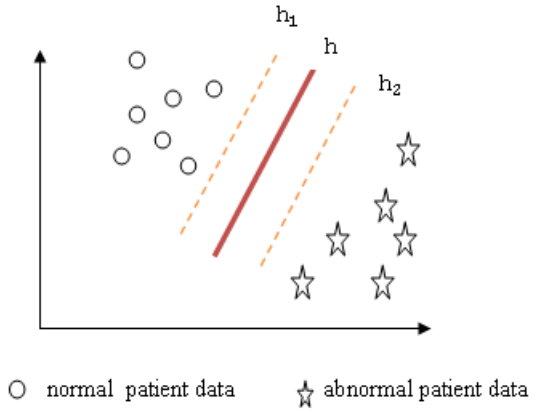

Figure 3 Kernelized support vector classifier

Figure 3 illustrates a kernelized support vector classifier to categorize the patient data into normal or abnormal. As shown in figure 3, the patient data are classified on either side of the hyperplane (h). The upper sides of the samples are called as normal patient data whereas the lower labeled samples are abnormal patient data. The output of the kernelized support vector classifier is obtained as follows,

$$c\,(x) = sign\ \sum \alpha_i y_i\ k\,(x_i, x_{i'}) \quad (6)$$

In (6), $c\,(x)$ denotes a predicted classification results of the weak learner, α_i denotes a weights of the training data, y_i represents the dependent variable (i.e. output) whose value is determined by observation. k denotes a kernel function that measures the similarity between any pair of inputs. '$sign$' determines whether the predicted classification output either positive or negative. The positive output results provide the higher similarity whereas the negative similarity provides less similarity.

The kernelized support vector classifier has some training errors which reduce the accurate disease diagnosis and risk prediction. This problem is overcome by combining all the weak learners' results. The output of the strong classifier is expressed as follows,

$$H\,(x) = \sum_{i=1}^{n} c_i\,(x) \quad (7)$$

From (7),$H\,(x)$ represents the output of strong classifier, $c_i\,(x)$ denotes a weak classifiers results. Iteratively, the linear program boost classifier assigns the weights to the current set of weak classifiers as it is given below.

$$H\,(x) = \sum_{i=1}^{n} \delta * c_i\,(x) \quad (8)$$

From (8)δ denotes a weights of the weak learner. The weight is the random integer number. After assigning the weight, the ensemble classifier calculates the training error of the each weak learner to achieve the higher diagnosing accuracy. The training error is defined as the squared difference between the actual and the predicted results of the kernelized support vector classifier. Therefore, training error is calculated using the following mathematical formula.

$$error = \left(c_o\,(x) - c_i\,(x)\right)^2 \quad (9)$$

From (9),$error$ represents the training error, $c_o\,(x)$ represents the actual output of the weak classifier, $c_i\,(x)$ represents observed results of the weak learner. Based on the error value, the initial weights are adjusted to find the accurate classification results. This is reiterated until no weak classifiers to add remain. Hence the name is called as iterative linear program boost classifier. When the weak learner correctly categorizes the patient data into a normal or abnormal, the initial weight gets increased. Otherwise, the initial weight gets decreased. The adjusted weight of the kernelized support vector classifier with the strong classification is given below,

$$H\,(x) = \hat{\delta} * c_i\,(x) \quad (10)$$

In (10),$H\,(x)$ represents a strong classification results, $\hat{\delta}$ denotes an adjusted weight of the weak classifier based on the error.

The ensemble classifier finds the weak learner with minimum error. In order to improve the classification performance and also minimizes the incorrect patient data classification, the linear program boosting ensemble classifier increases the margin of the different classes. The strong classifier outputs are subject into the margin,

$$\sum_{i=1}^{n} \hat{\delta} * c_i\,(x) + \vartheta_n \geq \mathrm{m_r} \quad \text{Where } \vartheta_n \geq 0 \qquad (11)$$

In (11),ϑ_n represents a non-negative vector of the slack variable, $\mathrm{m_r}$ denotes a margin of the classes. If the ensemble classification results are larger than the margin, all the patient data are correctly classified into particular class resulting in minimizes the incorrect classification i.e. false positive rate. As a result, the iterated linear program boost classifier attains the true class labels resulting in increases the tumor disease diagnosing accuracy. The algorithmic process of ensemble classification is described as follows,

Input: patient Data $d_1, d_2, d_3, \ldots d_n$, selected features
Output: Improve disease diagnosing accuracy
Begin

1. **for** each d_i
2. **construct 'n' weak learners**
3. Construct optimal hyperplane h
4. Find two marginal hyperplane h_1, h_2
5. The output of the weak classifier is $c\,(x) = sign\ \sum \alpha_i y_i\ k\,(x_i, x_i')$
6. **If** $(c\,(x) > 0)$ **then**
7. Patient Data is classified as 'normal'
8. **else**
9. Patient Data is classified as 'abnormal'
10. **end if**
7 **C**ombine all weak classifier results into strong $H\,(x) = \sum_{i=1}^{n} c_i\,(x)$
8 Assign weights to the weak classifier $\sum_{i=1}^{n} \delta * c_i\,(x)$
9 **For** each $c_i\,(x)$
10 Calculate the training error
11 **End for**
12 Adjust weight $\hat{\delta}$ to weak learners
13 Find best weak classifier with minimum training error
14 **If** $(\sum_{i=1}^{n} \hat{\delta} * c_i\,(x) + \vartheta_n \geq \mathrm{m_r})$ **then**
15 Classify all patient data
16 **End if**
17 Obtain strong classification results
18 **End for**

End

Algorithm 2 iterated linear program boosting classification

Algorithm 2 describes the step by step process of ensemble classification to improve the disease diagnosing accuracy with minimum false positive rate. Initially, the ensemble classifier constructs 'n' weak classifier for categorizing the patient data. The weak classifier constructs the hyperplane which acts as a boundary for verifying the training patient data with the testing results. Then the weak classifier categorizes the patient data into either side of the hyperplane. After classifying the data, the set of weak learner results are combined into one classifier. The similar weight is assigned to a set of the weak classifier. For each set of the classifier, the training error is computed. Followed by, the initial weight is adjusted. The ensemble classifier finds the best classifier with minimum error. Finally, the results are subjected to the margin for categorizing all the patient data. As a result, the ensemble classification results increases disease diagnosing accuracy and minimizes the false positive rate.

IV. Experimental Settings

Experimental evaluation of proposed PRFS-ILPBC technique and existing methods are RBFNN [1] and CVANN [2] are implemented in Java language using Epileptic Seizure Recognition Dataset. The Epileptic Seizure Recognition Dataset https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition is taken from the UCI Machine Learning Repository. The dataset comprises the recording of the brain activity data. The dataset comprises the 11500 instances and 179 attributes. The 178 attributes are explanatory variables represented as X1, X2…, and X178. The last attribute is a class label used for identifying the tumor in the brain area. The attributes characteristics are an integer and real. The dataset characteristics are multivariate and time series data. The associated tasks performed by the dataset are classification and clustering. For the experimental consideration, the numbers of patient data (i.e. instances) are taken as input from 50 to 500. The performances are evaluated with different metrics such as disease diagnosing accuracy, false positive rate and computation time.

V. RESULT AND DISCUSSIONS

The experimental results and discussion of the PRFS-ILPBC technique and existing RBFNN [1], CVANN [2] are described in this section with different performance parameters as disease diagnosing accuracy, false positive rate and computation time. The results are evaluated with the help of tables and graphical representation. The mathematical calculations of different metrics are provided for each subsection.

5.1 Impact of disease diagnosis accuracy

Tumor disease diagnosis accuracy is defined as the number of patient data (i.e. patient) correctly classified as normal or abnormal from the number of patient data taken as input. The mathematical formula for calculating the tumor diagnosis accuracy is given below,

$$DDA = \frac{Number\ of\ patient\ data\ correctly\ classified}{Number\ of\ patient\ data} * 100 \qquad (12)$$

In(12), DDA represents the disease diagnosis accuracy which is measured in terms of percentages (%).

Sample Calculation for disease diagnosis Accuracy

- **Proposed PRFS-ILPBC**: Number of patient data correctly classified is 44 and the total number of patients is 50. Then disease diagnosis accuracy is calculated as,

$$DDA = \frac{44}{50} * 100 = 88\%$$

- **Existing RBFNN:** Number of patient data correctly classified is 42 and the total number of patients is 50. Then disease diagnosis accuracy is calculated as,

$$DDA = \frac{42}{50} * 100 = 84\%$$

- **Existing CVANN**: Number of patient data correctly classified is 40 and the total number of patients is 50. Then disease diagnosis accuracy is calculated as,

$$DDA = \frac{40}{50} * 100 = 80\%$$

Let us consider the 50 patient data, proposed PRFS-ILPBC technique classifies 44 patient data. Therefore, the disease diagnosing accuracy is 88% whereas the disease diagnosing accuracy of RBFNN [1] and CVANN [2] are 84% and 80%. Similarly, the nine various results are calculated with a different number of patient data. The ten various results of disease diagnosis accuracy of three different techniques is reported in table 1.

Table 1 Disease diagnosis accuracy versus number of patient data

Number of patient data	Disease diagnosis accuracy (%)		
	PRFS-ILPBC	**RBFNN**	**CVANN**
50	88	84	80
100	91	85	74
150	93	87	79
200	92	88	83
250	95	89	85
300	93	88	84
350	92	87	83
400	94	88	81
450	92	86	79
500	91	85	79

Table 1 reports the experimental results of disease diagnosis accuracy versus a number of patient data collected from the Epileptic Seizure Recognition Dataset. The reported results show that the proposed PRFS-ILPBC technique improves disease diagnosis accuracy when compared to existing RBFNN [1] and CVANN [2].

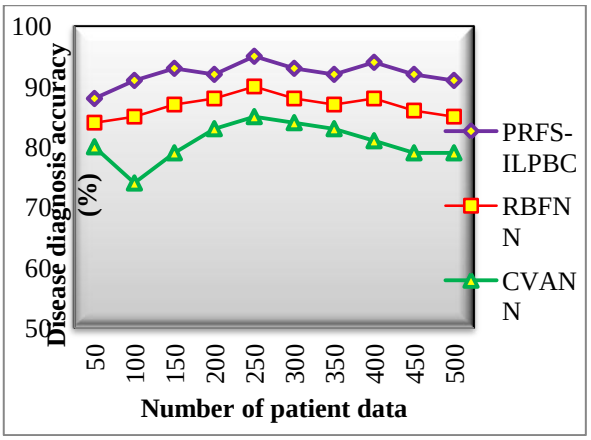

Figure 4 Performance results of disease diagnosis accuracy

Figure 4 illustrates the performance results of the disease diagnosis accuracy with respect to a number of patient data. As shown in figure 4, the numbers of patient data are taken as input in the 'x' axis and the results of the disease diagnosis accuracy are obtained at 'y' direction. The above graphical result clearly shows that the proposed PRFS-ILPBC technique increases the tumor disease diagnosis accuracy than the conventional classification methods. This significant development of the proposed PRFS-ILPBC technique is achieved by applying the ensemble classification technique. The ensemble classification initially constructs the number of weak learners. The kernelized support vector classifier is used as a weak learner categorizes two classes i.e. normal or abnormal by constructing the hyperplane. The weak classifier analyzes the risk level of the patient to identify the disease. Then the set of weak classifiers results are combined into strong. The ensemble classifier computes the training error based on the actual and predicted classification results of the weak learner. Then the weights of the weak classifier are adjusted based on the training error and find the weak learner with minimum error. As a result, the ensemble classifier accurately classifies the patient data into a normal or abnormal with the selected features. The obtained results of disease diagnosis accuracy of proposed PRFS-ILPBC technique are compared to the accuracy of conventional techniques. The average of different results shows that the proposed PRFS-ILPBC technique improves the disease diagnosis accuracy by 6% and 14% when compared to existing RBFNN [1] and CVANN [2]]respectively.

5.2 Impact of false positive rate

The false positive rate is defined as the number of patient data incorrectly classified as normal or abnormal from the number of patient data taken as input. The false positive rate is mathematically calculated as follows,

$$false\ positive\ rate = \frac{Number\ of\ patient\ data\ incorrectly\ classified}{Number\ of\ patient\ data} * 100 \quad (13)$$

The false positive rate is measured in terms of percentage (%).

Sample Calculation for false positive rate:

- **Proposed PRFS-ILPBC**: Number of patient data incorrectly classified is 6 and the total number of patients is 50. The false positive rate is computed as follows,

$$false\ postive\ rate = \frac{6}{50} * 100 = 12\%$$

- **Existing RBFNN**: Number of patient data incorrectly classified is 8 and the total number of patients is 50. The false positive rate is computed as follows,

$$false\ postive\ rate = \frac{8}{50} * 100 = 16\%$$

- **Existing CVANN**: Number of patient data incorrectly classified is 10 and the total number of patients is 50. The false positive rate is computed as follows

$$false\ postive\ rate = \frac{10}{50} * 100 = 20\%$$

Table 2 false positive rate versus number of patient data

Number of patient data	False positive rate (%)		
	PRFS-ILPBC	**RBFNN**	**CVANN**
50	12	16	20
100	9	15	26
150	7	13	21
200	8	10	17
250	5	11	15
300	7	12	16
350	8	13	17
400	6	12	19
450	8	14	21
500	9	15	21

Table 2 describes the experimental results of false positive rate with respect to a number of patient data. For the experimental consideration, numbers of patient data are taken from 50 to 500. The various performance results of false positive rate confirm that the proposed PRFS-ILPBC technique minimizes the false positive rate in the patient data classification. Let us consider the number of patient data is 50, false positive rate of proposed PRFS-ILPBC technique is 12% whereas the false positive rate of other two existing RBFNN [1] and CVANN [2] are 16% and 20%. Similarly, the various results are illustrated in figure 5.

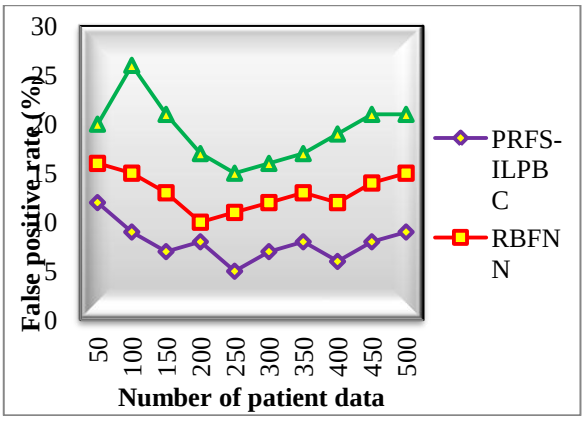

Figure 5 Performance results of false positive rate

Figure 5 illustrates the experimental results of false positive rate with respect to a number of patient data taken from the elliptic seizer dataset using three different classification techniques PRFS-ILPBC, RBFNN [1] and CVANN [2]. The above graphical result clearly shows that the PRFS-ILPBC technique minimizes the false positive rate in the tumor disease diagnosis. This is because of the PRFS-ILPBC technique outperforms well in the disease classification by minimizing the training error. The PRFS-ILPBC technique calculates the training error for all the weak classifiers. The ensemble classifier finds the weak learner with minimum error. In addition, the classification results are subjected to the margin with the selected variables for categorizing all the patient data into normal or disease-affected classes. This helps to improve accurate disease diagnosis and minimize the false positive rate.

The average of ten results shows that the false positive rate of PRFS-ILPBC technique is minimized by 40% when compared to RBFNN [1]. In addition, the false positive rate of PRFS-ILPBC technique is significantly minimized by 59% when compared to CVANN [2].

5.2 Impact of computation time

The computation time is defined as the amount of time taken to diagnosis the tumor disease with the number of patient data. The computation time is calculated using mathematical formula,

$$CT = number\ of\ patient\ data * time\ (diagonasis\ one\ patient\ data) \quad (14)$$

In (14) CT represents the computation time and it is measured in terms of milliseconds (ms).

Sample Calculation for computation time:

- **Proposed PRFS-ILPBC:** Number of patient data is 50, and the time taken to diagnosis the one patient data is 0.26ms. The computation time is computed as follows,

$$CT = 50 * 0.26\ ms = 13ms$$

- **Existing RBFNN**: Number of patient data is 50, and the time taken to diagnosis the one patient data is 0.3ms. The computation time is computed as follows,

$$CT = 50 * 0.3ms = 15ms$$

- **Existing CVANN**: Number of patient data is 50, and the time taken to diagnosis the one patient data is 0.33 ms. The computation time is computed as follows,

$$CT = 50 * 0.33\ ms \approx 17ms$$

Table 3 Computation time versus number of patient data

Number of patient data	Computation time (ms)		
	PRFS-ILPBC	**RBFNN**	**CVANN**
50	13	15	17
100	22	25	28
150	30	36	41
200	36	44	50
250	45	53	58
300	52	57	63
350	57	70	74
400	64	76	84
450	68	86	95
500	75	90	100

Table 3 describes the computation time for diagnosing the disease based on the number of patient data. The numbers of patient data are taken as input in the ranges 50 to 500 for calculating the computation time. Initially, the patient information's are collected from the Epileptic Seizure Recognition dataset and identifies the tumor disease with the selected features through the classification. The PRFS-ILPBC technique minimizes the computation time in the disease diagnosis when compared to other classification techniques. The ten different results are plotted in the following graph.

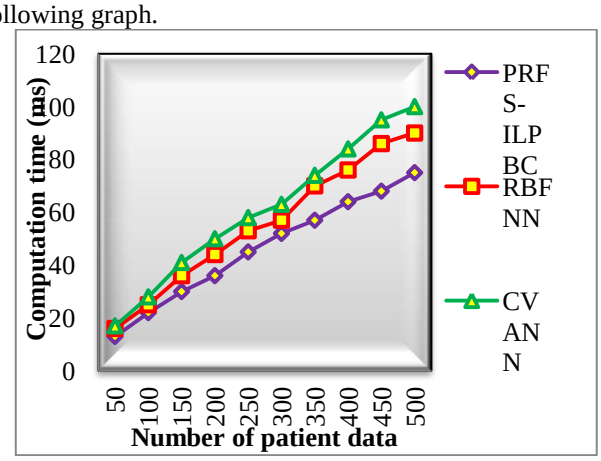

Figure 6 Performance results of computation time

Figure 6 depicts the computation time based on the number of patient data using three different classification techniques. The two-dimensional graphical results clearly show that the proposed PRFS-ILPBC technique minimizes the computation time when compared to existing classifier by selecting the relevant features from the dataset. The Epileptic Seizure Recognition dataset comprises the 178 attributes and 1 class labels. The classification is performed with the 178 attributes takes more time complexity. Therefore, most relevant features are selected before the classification for minimizing these time complexity in the disease diagnosis. The PRFS-ILPBC technique uses the bivariate correlation based probit regression for selecting the more similar features to identify the tumor disease. The regression finds the correlation between the features and tumor disease symptoms. If the features are more correlated to the tumor disease symptoms, then the feature is selected as a relevant for disease diagnosis. Otherwise, the features are removed. Then the PRFS-ILPBC technique performs the patient data classification using ensemble classifier only with the selected features. This helps to minimize the computation time for diagnosing the disease.

Let us consider the 50 patient data to calculate the computation time. The PRFS-ILPBC technique takes 13ms for diagnosing the patient as normal or abnormal. The computation time of other classification techniques RBFNN [1] and CVANN [2] are 15ms and 17ms. This discussion shows that the proposed PRFS-ILPBC technique minimizes the disease diagnosing time. The obtained results of PRFS-ILPBC technique is compared to other classification techniques. The comparisons result shows that the PRFS-ILPBC technique minimizes the computation time by 16% and 24% than the existing RBFNN [1] and CVANN [2] respectively. The above results discussions and statistical analysis prove that the proposed PRFS-ILPBC technique effectively improves the disease diagnosing accuracy with minimum time as well as false positive rate.

VI. CONCLUSION

An efficient technique called PRFS-ILPBC is developed for achieving higher disease diagnosing accuracy with minimal computation time. The PRFS-ILPBC technique collects the patient information from the dataset. Then the relevant feature selection and irrelevant feature removal are performed by applying the probit regression. The probit regression finds the relationship between the features and disease symptoms where the results are falls into one of the two classes i.e. relevant or irrelevant. The feature selection of PRFS-ILPBC technique is to minimize the computation time in the disease diagnosis. Finally, the classification is performed using iterative linear programming boost ensemble classification technique to categorize the patient as normal or disease-affected with the help of the selected features. The ensemble classification technique increases the tumor disease diagnosing accuracy and minimizes the false positive rate. The experimental evaluation is performed using brain tumor dataset with different metrics such as disease diagnosing accuracy, false positive rate and computation time.

The observed result shows that the PRFS-ILPBC technique improves the disease diagnosing accuracy and minimizes the computation time as well as false positive rate when compared to existing classification techniques. Hence it is concluded that PRFS-ILPBC is an efficient technique for accurate tumor disease diagnosis in the healthcare domain.

REFERENCES

1. Sandeep Kumar Satapathy, Satchidananda Dehuri, Alok Kumar Jagadev, "ABC optimized RBF network for classification of EEG signal for epileptic seizure identification", Egyptian Informatics Journal, Elsevier, Volume 18, Issue 1, 2017, Pages 55-66,
2. Musa Peker,Baha Sen, Dursun Delen, "A Novel Method for Automated Diagnosis of Epilepsy Using Complex-Valued Classifiers", IEEE Journal of Biomedical and Health Informatics, Volume 20, Issue 1, 2016, Pages 108 – 118
3. Khalid Abualsaud, Massudi Mahmuddin, Mohammad Saleh, and Amr Mohamed, "Ensemble Classifier for Epileptic Seizure Detection for Imperfect EEG Data", the Scientific World Journal, Hindawi Publishing Corporation, Volume 2015, December 2014, Pages 1-15
4. VarshaHarpale and Vinayak Bairagi, "An adaptive method for feature selection and extraction for classification of epileptic EEG signal in significant states", Journal of King Saud University - Computer and Information Sciences, Elsevier, 2018, Pages 1-9
5. P. Fergus , A. Hussain , David Hignett , D. Al-Jumeily ,Khaled Abdel-Aziz , Hani Hamdan, "A machine learning system for automated whole-brain seizure detection", Applied Computing and Informatics, Elsevier, Volume 12, Issue 1, 2016, Pages 70-89
6. N. Sriraam , S. Raghu, Kadeeja Tamanna, Leena Narayan1, Mehraj Khanum1, A. S. Hegde and Anjani Bhushan Kumar, "Brain Informatics, Springer, Volume 5, Issue 10, 2018, Pages 1-10
7. Shamsul Huda, John Yearwood, Herbert F. Jelinek, Mohammad Mehedi Hassan3, Giancarlo Fortino, and Michael Buckland, "A Hybrid Feature Selection with Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis", IEEE Access, Volume 4, 2016, Pages 9145 – 9154
8. Hengjin Ke , Dan Chen , Xiaoli Li , Yunbo Tang , Tejal Shah , Rajiv Ranjan, "Towards Brain Big Data Classification: Epileptic EEG Identification With a Lightweight VGGNet on Global MIC", IEEE Access , Volume 6, Pages 14722 – 14733

9. Kostas M. Tsiouris , Vasileios C. Pezoulas , Michalis Zervakis, Spiros Konitsiotis, Dimitrios D. Koutsouris, Dimitrios I. Fotiadis, "A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals", Computers in Biology and Medicine, Elsevier ,Volume 99, 2018, Pages 24-37
10. Enamul Kabir, Siuly, Yanchun Zhang, "Epileptic seizure detection from EEG signals using logistic model trees", Brain Informatics, Springer, Volume 3, Issue 2, 2016, Pages 93–100
11. U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals", Computers in Biology and Medicine, Elsevier, Volume 100, 2018, Pages 270-278
12. Paul Fergus, David Hignett, Abir Hussain, Dhiya Al-Jumeily, and Khaled Abdel-Aziz, "Automatic Epileptic Seizure Detection Using Scalp EEG and Advanced Artificial Intelligence Techniques", BioMed Research International, Hindawi Publishing Corporation, Volume 2015, December 2014, Pages 1-17
13. Marco Piangerelli, Matteo Rucco, Luca Tesei, Emanuela Merelli, "Topological classifier for detecting the emergence of epileptic seizures", BMC Research Notes, Volume 11, Issue 392, 2018, Pages 1-7
14. Md Mursalin, Yuan Zhang, Yuehui Chen, Nitesh V Chawla, "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier", Neurocomputing, Elsevier, Volume 241, 2017, Pages 204-214
15. M. Valderrama, C. Alvarado, S. Nikolopoulos, J. Martinerie, C. Adam, V. Navarro, M. Le Van Quyen, "Identifying an increased risk of epileptic seizures using a multi-feature EEG–ECG classification", Biomedical Signal Processing and Control, Elsevier, Volume 7, 2012, Pages 237– 244
16. Hadi Ratham Al Ghayab , Yan Li , Shahab Abdulla , Mohammed Diykh ,Xiangkui Wan, "Classification of epileptic EEG signals based on simple random sampling and sequential feature selection", Brain Informatics, Springer, Volume 3, Issue 2, 2016, Pages 85–91
17. Mustafa Serter Uzer, Nihat Yilmaz, and Onur Inan, "Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification", The Scientific World Journal, Hindawi Publishing Corporation, Volume 2013, July 2013, Pages 1-10
18. Ramy Hussein, Hamid Palangi, Rabab Ward, Z. Jane Wang, "Epileptic Seizure Detection: A Deep Learning Approach", Electrical Engineering and Systems Science, Signal Processing, Pages 1-12, 2018
19. Yuanfa Wang, Zunchao Li, Lichen Feng, Chuang Zheng, and Wenhao Zhang, "Automatic Detection of Epilepsy and Seizure Using Multiclass Sparse Extreme Learning Machine Classification", Computational and Mathematical Methods in Medicine, Hindawi, Volume 2017, June 2017, Pages 1-10
20. Md. Kamrul Hasan, Md. Asif Ahamed, Mohiuddin Ahmad, and M. A. Rashid, "Prediction of Epileptic Seizure by Analysing Time Series EEG Signal Using k-NN Classifier", Applied Bionics and Biomechanics, Hindawi, Volume 2017, August 2017, Pages 1-12.

AUTHORS PROFILE

P.S. Renjeni received B.Sc Chemistry from Sree Devi Kumari College for women, Manonmaniam Sundaranar University, India and obtained MCA from Noorul Islam College of Engineering, Manonmaniam Sundaranar University, India. Presently, she is doing Ph.D in Bharathidasan University, Trichy, India. She has 15 years of teaching experience and working as Assistant Professor at the Department of Computer Science in V.T.M. College of Arts and Science, Arumanai, Tamil Nadu, India. Her research interest is in Data Mining.

B. Mukunthan pursued Bachelor of Science (Computer Science) from Bharathiar University, India in 2004 and Master of Computer Applications from Bharathiar University in year 2007 and Ph.D from Anna University - Chennai in 2013. He is currently working as Research Advisor in Department of Computer Science, Bharathidasan University, Tiruchirapalli since 2016. He is a member of IEEE & IEEE computer society since 2009, a life member of the MISTE since 2010. He has published more than 10 research papers in reputed international journals. He is also Microsoft Certified Solution Developer. His main research work focuses on Algorithms, Bioinformatics, Big Data Analytics, Data Mining, IOT and Neural Networks. He also invented a Novel and Efficient online Bioinformatics Tool and filed for patent. He has 12 years of teaching experience and 10 years of Research Experience.