Neighbor Embedding Feature Selected Light Gradient Boosting Classification for Breast Cancer Detection with Gene Expression Data

S.Rajasekaran, S.Sathyabama

Abstract:- ***Breast cancer is one of the most frequently diagnosed cancers among women worldwide. Accurate detection of Breast cancer is essential for providing better treatment and risk minimization of the patients. Recently, the collection of biological data like gene expression, protein sequences, DNA sequences are used due to improvements of accessible data mining techniques to diagnosis the disease at an earlier stage. The current state-of-art methods reported to have certain limitations in their diagnostic capability. In order to improve the breast cancer classification, an efficient technique called Gaussian Kernelized Neighbor Embedding based Light Gradient Boost Classification (GKNE-LGBC) technique is introduced. The GKNE-LGBC technique considers the benchmark microarray dataset and performs two processes such as feature selection and classification for detecting breast cancer using gene expression data. The number of gene and the data are collected from the microarray dataset. After collecting, the Gaussian Kernelized stochastic neighbor embedding algorithm is applied to select the relevant features (i.e. genes) and remove the irrelevant features based on the distance similarity. Next, the classification of the gene expression data is done with the help of steepest descent light gradient boosting algorithm. The boosting algorithm initially constructs' number of weak learners i.e. bivariate regression tree to classify the input expression data into normal or cancerous with the selected features. Then the weak classifiers are combined into strong by minimizing the training error. This helps to improve breast cancer detection accuracy and minimizes the false positive rate. The experimental evaluation is carried out using gene microarray dataset with various parameters such as breast cancer detection accuracy, false positive rate and breast cancer detection time with a number of genes. The experimental results confirm that the proposed GKNE-LGBC technique accurately identifies breast cancer with higher accuracy, and minimal time complexity as well as false positive rate as compared to the state-of-art- methods.***

Keywords: ***benchmark microarray dataset, gene expression data, breast cancer detection, Gaussian Kernelized stochastic neighbor embedding, feature selection, steepest descent light gradient boosting algorithm, bivariate regression tree***

I. PREAMBLE

Cancer detection is the major research area in medical field to accurately predict the different type of tumor. Earlier cancer detection has always provided better treatment to the patient. Generally, the biomedical research is carried out with the different microarray dataset which consists of gene expression, protein sequences and so on. The microarray database is a repository which includes a microarray gene expression data. The main application of microarray database is to store the thousands of genes and accurately identifies the disease at an earlier stage with minimum time. The several methods have been conducted with the statistical and machine learning for breast cancer classification, but there are some major concerns that make it an error, inaccurate and more time-consuming. In order to perform accurate disease identification, the dimensionality reduction technique called feature selection is introduced since the large volume of genes reduces the performance of the classifier. Therefore, the feature selection is an important processing step before the classification.

A principal component analysis and autoencoder neural network with the AdaBoost algorithm (PCA-AE-Ada) was designed in [1] for efficiently detects the clinical outcomes of cancer patients with gene expression profile. The designed algorithm failed to identify which features are most relevant for cancer prediction. A deep learning-based multi-model ensemble method was developed in [2] for identifying the cancer with gene expression data. But the accurate cancer prediction was not obtained with minimum time complexity.

A fuzzy-based Logistic regression was developed in [3] for feature selection and cancer gene data classification. The designed method considered the missing data other attributes for prediction which minimized the cancer prediction accuracy. A Gene Interaction Regularized Elastic Net (GIREN) model was introduced in [4] for predicting cancer with gene expression. But the method failed to minimize the computational complexity of cancer prediction. A neighborhood entropy-based uncertainty measures were introduced in [5] for selecting the relevant feature to perform the cancer classification. The designed method failed to improve the classification performance of cancer detection.

An extensive assessment of machine learning systems was introduced in [6] for identifying the colon cancer using gene expression data. But the designed systems was not minimized the classification error.

Logistic regression with elastic net regularization model was introduced in [7] for breast cancer data classification using gene expression variables. The method failed to minimize the time complexity of data classification.

A Mutual Information method was introduced in [8] for finding the robust gene signature from microarray data to predict breast cancer. Accurate gene selection was not performed to improve breast cancer prediction accuracy.

Revised Manuscript Received on September 10, 2019.

S.Rajasekaran, Research Scholar, Bharathiar University, Coimbatore, Tamilnadu. India.
(E-mail: srs20may@gmail.com)
Dr.S.Sathyabama, Assistant Professor, Department of Computer Science, Thiruvalluvar Government Arts College, Rasipuram, Namakkal, Tamilnadu,.India.
(E-mail: sathyaaksrct@yahoo.com)

A Random Subspace-based SVM (RS-SVM) classifier was developed in [9] to predict cancer using gene expression profiles. The designed classifier consumed more time for cancer prediction. An intelligent decision support system (IDSS) was introduced in [10] for detecting breast cancer with gene expression. The system failed to enhance prediction accuracy. The multi-objective and Machine Learning techniques were introduced in [11] for gene selection to detect cancer. But the designed feature selection algorithm failed to handle large data with minimum time.

A deep neural network model based on Attention method was developed in [12] for prediction of breast cancer by extracting the useful information from the gene profile data. The designed method increased the survival time of cancer disease prediction. A hierarchical stacked autoencoder framework was introduced in [13] to combine the gene expression and transcriptome data for detecting the cancer subtypes. But the performance of the cancer detection accuracy was not improved. A weighted K-means support vector machine (wKM-SVM) was developed in [14] for cancer prediction. The designed classifier was not improved the classification accuracy.

A two-phase search strategy was developed in [15] for diagnosing the prostate cancer from the gene expression data. Though the strategy increases the classification accuracy, the time complexity was not minimized. Hidden Markov models (HMMs) was introduced in [16] for classifying cancer with gene expression profiles. Though the model minimizes the classification time, the training error was not minimized. A new ensemble approach was developed in [17] for classifying cancer-based on gene expression profiles. But the classification accuracy was not improved using an ensemble approach.

Machine learning models were designed in [18] for breast cancer prediction with anthropometric and clinical features. The designed models failed to perform an accurate classification with a minimum error rate. A novel hybrid wrapper technique was introduced in [19] to find the optimal gene subsets from gene expression data for identifying cancer. Though the technique improves the classification accuracy, the time was not minimized. A class imbalance-aware Relief algorithm was designed in [20] for classifying the tumors with high-dimensional imbalanced microarray gene expression data. But the designed algorithm failed to minimize the classification time.

The major issues are identified from the survey are overcome by introducing a novel technique called GKNE-LGBC technique. The proposal contribution is summarized in below subsection on the contrary to existing techniques.

1.1 Proposal Contribution

The contribution of GKNE-LGBC technique is summarized on the contrary to existing techniques. The contributions are given as below,

A GKNE-LGBC technique is introduced for effectively dealing with breast cancer detection scenario including microarray gene expression data. To achieve this contribution, feature selection and ensemble classification is performed.

To improve the breast cancer detection accuracy, steepest descent light gradient boost ensemble classifier is designed. The ensemble classifier initializes the empty set of weak learners as a bivariate regression tree with the number of training gene expression data. The tree measures the relationship between the extracted features and cancer disease features. If these two features are highly correlated, the data are classified into cancerous. Otherwise, the data is classified as normal. The weak classifier results are summed and obtained strong results with minimum training error. This helps to minimize the false positive rate.

To minimize the time complexity, relevant feature selection is performed using Gaussian Kernelized stochastic neighbor embedding algorithm. The algorithm measures the similarity between the feature and objective (i.e. breast cancer detection) to find the relevant or irrelevant. The relevant features are selected for breast cancer detection and remove irrelevant features.

1.2. Structure organization

The structure of the paper is organized into different sections as follows. Section 2 provides the problem statement involving the existing breast cancer classification techniques. In section 3, gives a detailed description of the feature selection and then the proposed classification. Section 4 illustrates the experimental evaluation and different parameters settings using microarray dataset. Followed by, the statistical results analysis of the different parameters and various techniques are presented in section 5. The conclusion of the paper is given in section 6.

II. PROBLEM STATEMENT

In this section, the breast cancer detection problem is described with the gene expression dataset. While processing a large number of genes, dimensionality reduction is significant for minimizing the time complexity of disease diagnosis and treatment of disease. The challenges involved dealing with a huge number of irrelevant attributes (i.e. features). The presence of these irrelevant features incurs extra computation time in both the training and testing phase of the classifier but also minimizes the performance of classification. This may increase the disease prediction time of the patient and also increases the risk level. In addition, disease identification accuracy is affected by the presence of thousands of attributes most of which are unnecessary for the classification. Thus, a major issue of microarray data classification is to select the minimum possible set of genes that helps to achieve better detection accuracy. Therefore, an efficient machine learning technique is required to solve these kinds of issues in breast cancer detection.

III. METHODOLOGY

A Gaussian Kernelized Neighbor Embedding based Light Gradient Boost Classification (GKNE-LGBC) technique is introduced to improve the breast cancer detection at an earlier stage with minimum time. The proposed GKNE-

LGBC technique uses the Gaussian kernelized stochastic neighbor embedding algorithm to select the relevant genes i.e. features from the microarray dataset. With the selected features, the steepest descent Light Gradient Boost classification technique is applied for identifying the gene expression data as normal or abnormal. The architecture diagram of the proposed GKNE-LGBC technique is shown in figure 1.

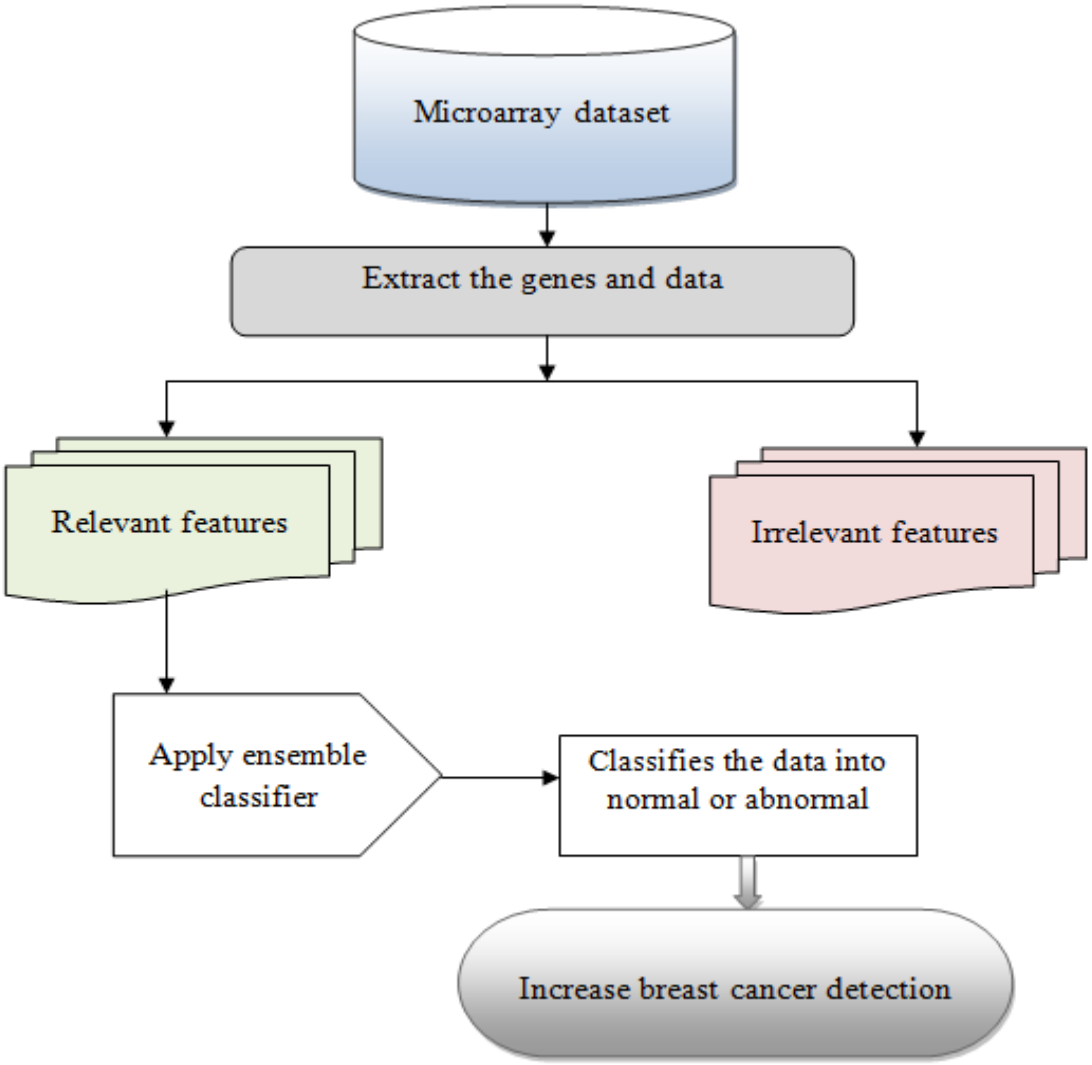

Figure 1 architecture of the GKNE-LGBC technique

Figure 1 illustrates the architecture of the proposed GKNE-LGBC technique to improve breast cancer detection accuracy with minimum time. Let us consider the gene microarray dataset D_M contains a number of genes $G_1, G_2, G_3, \ldots G_n$ and data $D_1, D_2, D_3 \ldots . D_n$. The relevant feature selection is initially performed to chosen the most relevant features (i.e. gene) from the microarray dataset to minimize the breast cancer detection time.

Secondly, ensemble classifier is applied to identify the breast cancer which helps to improve the performance of bioinformatics data analysis. The brief explanation of GKNE-LGBC technique is presented in the following section.

3.1 Gaussian kernelized stochastic neighbor embedding for feature selection

The first process in the design of the proposed GKNE-LGBC technique is to select the relevant features for minimizing the complexity involved in breast cancer detection. In general, the gene expression dataset consists a lot of irrelevant attributes (i.e. features) so one of the major steps is in the machine learning is to perform the feature selection before the classification process. Therefore, the proposed technique designed a Gaussian kernelized stochastic neighbor embedding (GKSNE) algorithm for selecting the features from the gene microarray dataset. The stochastic neighbor embedding is a machine learning algorithm as well as the dimensionality reduction technique by embedding the high-dimensional data into a low-dimensional space. In GKSNE, stochastic is a random probability distribution over the pairs of high-dimensional data (i.e. gene) in such a way that the similar features have a high probability of being picked while the unrelated features

have an extremely small probability of being picked.

Let us consider gene microarray dataset D_M with the genes $G_1, G_2, G_3, \dots G_n$. The stochastic neighbor embedding initially calculates the probability between the features and objective based on the Gaussian kernel which is mathematically calculated as follows.

$$P_{ij} = \frac{\exp\left(-\frac{1}{2}\frac{\left\|F_i - O_j\right\|^2}{\sigma^2}\right)}{\Sigma \exp\left(-\frac{1}{2}\frac{\left\|F_i - o_j\right\|^2}{\sigma^2}\right)} \qquad (1)$$

In (1),P_{ij} denotes a probability, F_i denotes a i^{th} feature, O_j denotes breast cancer detection i.e. objective, σ denotes a deviation. The Gaussian kernel measures the distance similarity between F_i and O_j using Euclidean distance $\left\|F_i - O_j\right\|$. The probability value obtains between zero and one. Based on the probability value, the threshold (i.e. 0.5) is set to identify the relevant features.

$$y = \begin{cases} P_{ij} > 0.5 \;\; ; relevant\ features \\ P_{ij} < 0.5 \;\; ; \; irrelevant\ features \end{cases} \qquad (2)$$

In (2),y denotes an output function. The probability results greater than the threshold is said to a relevant feature. Otherwise, the features are said to be an irrelevant features. Finally, the relevant features are selected for the classification and the irrelevant features are removed from the dataset. The Gaussian kernelized stochastic neighbor embedding algorithm is described as follows,

Input: Gene microarray dataset D_M , Number of genes $G_1, G_2, G_3, \dots G_n$

Output: Select relevant features from the dataset

Begin

1. For each feature $F_i$in dataset D_M

2. calculate probability P_{ij} between F_i and O_j

3. if ($P_{ij} > 0.5$**) then**

4. The feature is said to be a relevant

5. Select the features for classification

6. **else**

7. The feature is said to be an irrelevant

8. Remove the features

9. **End if**

10. end for

End

Algorithm 1 Gaussian kernelized stochastic neighbor embedding algorithm

Algorithm 1 describes the Gaussian kernelized stochastic neighbor embedding algorithm to find the relevant and irrelevant features from the dataset. Initially, the probability is calculated based on the Gaussian kernel function for measuring the distance similarity. Then the probability results are verified with the threshold value for identifying the relevant and irrelevant features. The relevant features are selected for classification and remove irrelevant features. Therefore, the feature selection process of proposed technique minimizes the time complexity in breast cancer detection.

3.2 Steepest decent light gradient boost classification for breast cancer detection

After the feature selection, the gene classification is performed to identify the cancerous samples from the dataset. The main objective of this method is to construct the boosting classifier that correctly distinguished the cancerous samples and normal samples from the gene expression profiles. The proposed GKNE-LGBC technique uses the Steepest Decent Light Gradient Boost Classification (SDLGBC) algorithm to identify breast cancer with higher accuracy. The SDLGBC algorithm is a machine learning ensemble technique that converts the weak classifier into a strong one. A weak learner is a base classifier that failed to provide an accurate classification. On the contrary, a strong learner is also a classifier that is arbitrarily well-correlated and provides true classification by combining the weak classifier results. Therefore, the ensemble learning algorithm provides accurate classification results than normal classifier.

In general, several ensemble learning techniques have been presented. Recently, the size of the microarray dataset is increased and it becomes complex using conventional boosting algorithms to provide faster classification results. On the contrary, the proposed GKNE-LGBC technique uses the SDLGBC algorithm as it provides the fast classification results, and significantly provides high accuracy while handling the number of gene expression data. Here the 'Light' indicates the high speed of the classification. The SDLGBC algorithm performs only the tree-based classification which constructs the decision tree classifier in the form of vertical manner.Therefore, the proposed classifier is also known as a leaf-wise decision tree algorithm which minimizes the more training loss when compared to another decision tree algorithm. The structure of the SDLGBC algorithm is shown in figure 2.

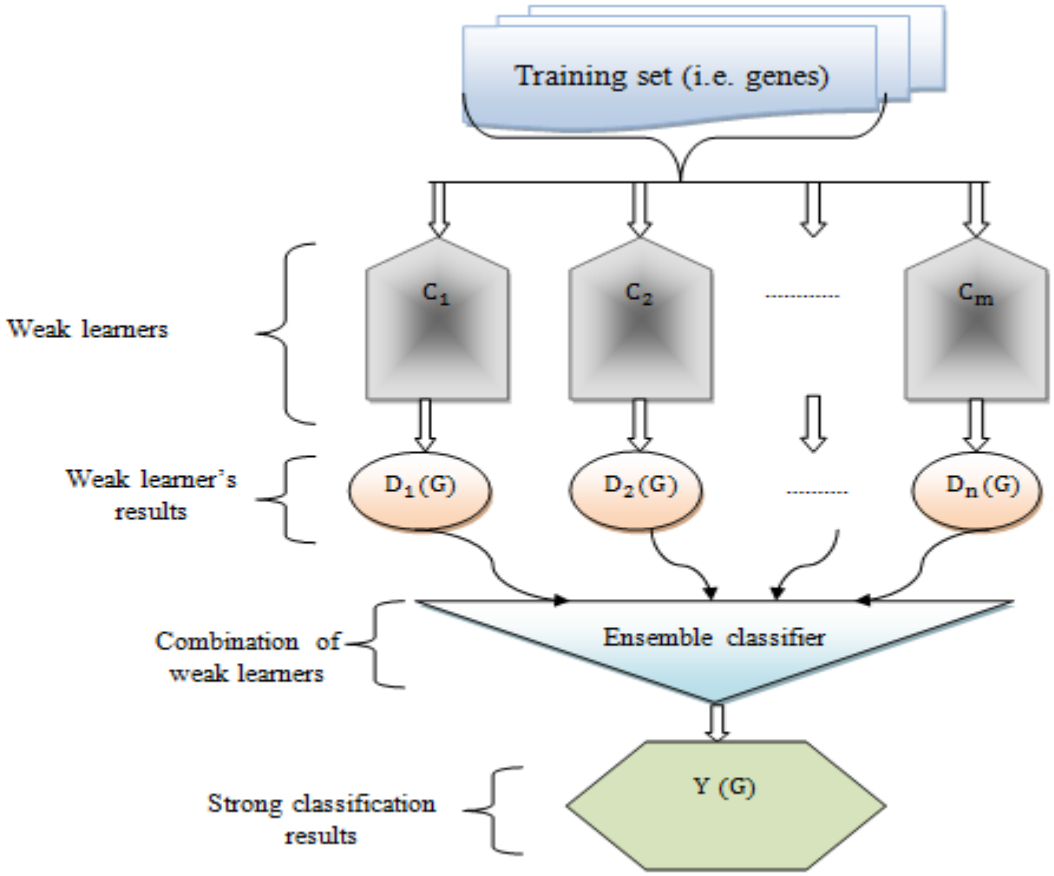

Figure 2 flow process of SDLGBC algorithm

Figure 2 illustrates the flow process of the SDLGBC algorithm. Let us consider the training sets $\{X_i, Y_i\}$ where X_i denotes an input i.e. gene expression data $D_1, D_2, D_3, \dots D_n$ and Y_i represents the final strong classification results. The bootstrap classifier initially constructs an empty set of 'm' weak learners $C_1, C_2, C_3, \dots C_m$ with a number of gene expression data and the selected feature subset. The ensemble classifier uses the bivariate regression tree as a weak classifier. A bivariate regression tree is a decision based classifier used to find the relationship between the training data and testing data (i.e. cancerous data). The regression tree is constructed with a root node, branch node and leaf node. The root node performs the "test" on a gene data based on their relationship and branch node represents the outcome of the test, and finally the leaf node provides the output class labels. The path from the root node to leaf node denotes a certain classification rules.

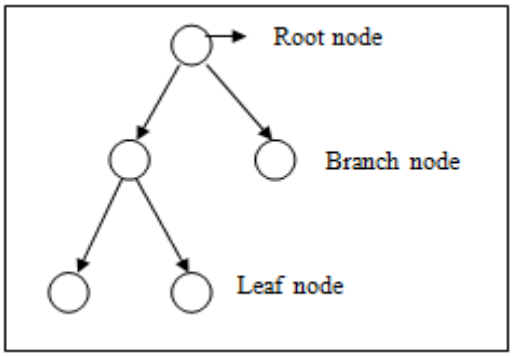

Figure 3 Bivariate Regression tree

Figure 3 illustrates a bivariate regression tree to classify the input gene expressed data as normal or cancerous based on the certain rule (i.e. relationship).

The root node represents the training gene data.

The branch node finds the condition based on the correlation between the gene expression data. Therefore, the bivariate correlation between the selected features with the testing features are measured as follows,

$$\beta = \frac{(\sum F_{Ts}F_{Tr}) - (\sum F_{Ts})(\sum F_{Tr})}{\sqrt{\{\sum {F_{Tr}}^2 - (\sum F_{Tr})^2\}\{\sum {F_{Ts}}^2 - (\sum F_{Ts})^2\}}} \qquad (3)$$

From (3), β represents the bivariate correlation coefficient, F_{Ts} is the testing feature set (i.e. cancerous data) , F_{Tr} represents the training feature set, $\sum F_{Ts} F_{Tr}$ is a sum of the product of paired score, $\sum F_{Ts}$ denotes the sum of F_{Ts} score, $\sum F_{Tr}$ denotes a sum of F_{Tr} score, $\sum {F_{Tr}}^2$ denotes a sum of the squared score of $\sum F_{Tr}$ and $\sum {F_{Ts}}^2$ represents the sum of the squared score of $\sum F_{Ts}$.

The bivariate correlation coefficient provides the two results −1 and +1. The coefficient provides '+1' indicates the positive correlation between the two features and then classifies the gene expression data into the cancerous. The coefficient provides '-1'represents the negative correlation between the features. If the correlation coefficient provides a negative correlation, then the leaf node classifies the gene expression data as normal. In this way, the entire gene

expression data are trained with the weak classifier to distinguish the normal and cancer from the microarray dataset. But the weak learner has some training error in the classification process. In order to obtain strong classifier, the outputs of weak learner results are summed into one which is expressed as follows.

$$Y = \sum_{i=1}^{m} D_i\ (G) \quad (4)$$

In (4),Y represents the output of the strong classifier, $D_i\ (G)$ represents the output of the weak classifier results. Then the similar weight is assigned to each classification results.

$$Y = \sum_{i=1}^{m} \alpha * D_i\ (G) \quad (5)$$

In (5),α indicates the weight of the classifier, the main aim of ensemble classifier is to minimize the objective function or criterion. Here, the objective function is also called as a cost loss function or error function. The mean square error is calculated for finding the accurate classification results of the weak learner. The mean square error is mathematically calculated based on the squared difference between the actual and predicted classification results of the weak learner.

$$E_{MS} = (D_o(G) - D_i\ (G))^2 \quad (6)$$

In (6), E_{MS} represents the mean square error, $D_o(G)$ denotes actual results of the weak learner, $D_i\ (G)$ represents the predicted results. Based on the squared error, the initial weights of the weak learners are updated. If the weak learner correctly classified the data, then their weight is decreased. Otherwise, the weight gets increased. Then the proposed ensemble classifier finds the weak learner results with minimum square error using steepest descent function. The steepest descent function is used in classification to find minimum of a function.

$$G(x) = \arg\min\ [E_{MS}] \quad (7)$$

In (7),$G(x)$ represents the steepest descent function, $\arg min$ denotes an argument of the minimum function to find the minimum error of the weak classification results, E_{MS} indicates the mean square error. Therefore, the output of the ensemble classifier with minimum error and the updated weight as given below,

$$Y = \sum_{i=1}^{m} \alpha^t * D_i\ (G) \quad (8)$$

In (8),Y denotes a output of the strong classifier, α^t denotes a updated weight of weak classifier $D_i\ (G)$. The output of the strong classifier provides the accurate classification results with minimum error rate. The ensemble classification result clearly shows that the input gene expression data are correctly distinguished as a normal and cancerous with less false positive rate. In this way, the proposed technique accurately predicts the breast cancer from the gene expression data at an earlier stage. The algorithmic process of Steepest Decent Light Gradient Boost Classification is described as follows,

Input: Number of gene expression data

$D_1, D_2, D_3, \ldots . D_n$, selected genes

Output: Improve breast cancer detection accuracy

Begin

1. **For each gene**
2. Construct' empty set of weak learners $\{C_1, C_2, C_3, \ldots . C_m\}$
3. Measure the bivariate correlation
4. **If** $(\beta = +1)$ **then**
5. Positive correlation between $F_{Ts}\ and\ F_{Tr}$
6. Gene expression data is classified into cancerous
7. **else**
8. Negative correlation between $F_{Ts}\ and\ F_{Tr}$
9. Gene expression data are classified into normal
10. **end if**
11. Combine a set of weak learners $Y = \sum_{i=1}^{m} D_i\ (G)$
12. **For** each $D_i\ (G)$
13. Assign the similar weight $Y = \sum_{i=1}^{m} \alpha\ D_i\ (G)$
14. Calculate E_{MS}
15. Find the weak learner with minimum E_{MS} i.e. $G(x) = \arg\min[E_{MS}]$
16. Obtain strong classification results $Y = \sum_{i=1}^{m} \alpha^t * D_i\ (G)$
17. **end for**
18. **end for**

End

Algorithm 2 steepest descent light gradient boost ensemble classification

The algorithm 2 clearly describes the step by step process of the steepest descent light gradient boost ensemble classification to accurately detect the breast cancer from the given gene expression dataset. Initially, the ensemble classifier constructs an empty set of weak learner with the training gene expression data. The bivariate regression tree is used as a weak classifier to measure the correlation between the training feature and cancerous feature set. If

both the sets are highly correlated, then the leaf node classifies the input samples as cancerous. Otherwise, it classified as normal samples.

Then the classification results of weak learners are combined to create strong classification results by finding the accurate classification results with minimum error.

The steepest decent function is used in the proposed ensemble classifier for finding the best classifier with minimum training error. As a result, the ensemble classifier improves breast cancer detection accuracy and minimizes the false positive rate.

IV. EXPERIMENTAL EVALUATION AND PARAMETER SETTINGS

Experimental evaluations of proposed GKNE-LGBC and existing methods PCA-AE-Ada [1] and deep learning-based multi-model ensemble method [2] are performed using Java language with breast cancer microarray dataset taken from the http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. This dataset comprises the 244881 attributes and 97 instances. The attributes from 1 to 244880 represent the genes. The last attributes is a class attributes which provides the two class labels such as relapse and non-relapse.

The relapse indicates the gene expression data in abnormal conditions (i.e. cancerous) and non-relapse denotes a no abnormality (i.e. normal). In this dataset, the attribute is denoted in the form of ordered sequence of @attribute statements. The attribute characteristics are numeric i.e. real or integer numbers. Among the number of genes, the more relevant genes are selected from the dataset. After that, the classification of cancerous or normal is performed with the selected genes. For the experimental evaluation, the numbers of gene expression data are taken as input for identifying breast cancer.

The performances of GKNE-LGBC technique and existing methods PCA-AE-Ada [1] and deep learning-based multi-model ensemble method [2] are evaluated with the different parameters such as breast cancer detection accuracy, false positive rate, and time complexity. The experimental result of proposed GKNE-LGBC technique is compared with two existing techniques. The comparative analyses of the different parameters are presented in the following sections

V. COMPARATIVE ANALYSES UNDER DIFFERENT PARAMETERS & RESULTS

The experimental results of GKNE-LGBC technique and existing methods PCA-AE-Ada [1] and deep learning-based multi-model ensemble method [2] are compared in this section with different parameters such as breast cancer detection accuracy, false positive rate, and time complexity. The obtained results are discussed with the help of either table or graphical representation. For each section, the statistical calculation is provided to show the performance of the proposed technique against conventional techniques.

5.1 Performance analysis of breast cancer detection accuracy

Breast cancer detection accuracy is measured as the ratio of the number of gene expression data are correctly classified as normal or abnormal to the total number of data taken as input. The detection accuracy is mathematically calculated as follows,

$$BCDA = \left[\frac{Number\ of\ data\ correctly\ classified\ as\ normal\ or\ abnomral}{n}\right] * 100 \quad (9)$$

In (9),$BCDA$ indicates the breast cancer detection accuracy, n denotes a number of gene expression data. The Breast cancer detection accuracy is measured in the unit of percentage (%). The statistical evaluation is given below.

Calculation:

Proposed GKNE-LGBC**:** Number of gene expression data correctly classified is 8 and the total number gene expression data is 9. Then the breast cancer detection accuracy is calculated as follows,

$$BCDA = \frac{8}{9} * 100 = 88.8\% \cong 89\%$$

Existing PCA-AE-Ada: Number of gene expression data correctly classified is 7 and the total number gene expression data is 9. Then the breast cancer detection accuracy is calculated as follows,

$$BCDA = \frac{7}{9} * 100 = 77.7\% \cong 78\%$$

Existing deep learning-based multi-model ensemble method: Number of gene expression data correctly classified is 6 and the total number gene expression data is 8. Then the breast cancer detection accuracy is calculated as follows,

$$BCDA = \frac{6}{9} * 100 = 66.6\% \cong 67\%$$

The statistical result shows that the GKNE-LGBC obtains 89% of breast cancer detection accuracy by correctly classified 8 data from the 9 data. The existing PCA-AE-Ada, deep learning-based multi-model ensemble method achieves 78% and 67% of breast cancer detection accuracy by correctly classified the gene expression data are 7 and 6 respectively. The different results of the Breast cancer detection accuracy with various gene expression data are shown in table 1.

Table 1 Breast cancer detection accuracy under different gene expression data

Number of gene expression data	Breast cancer detection accuracy (%)		
	GKNE-LGBC	PCA-AE-Ada	Deep learning-based multi-model ensemble method
9	89	78	67
18	94	89	83
27	93	85	81
36	94	89	83
45	96	87	80
54	93	85	81
63	97	86	79
72	94	89	83
81	95	91	88
90	96	90	86

Table 1 describes the performance results of breast cancer detection accuracy versus a number of gene expression data taken from the microarray dataset. For the experimental consideration 90 instances (i.e. gene expression data) are taken from the 97 instances of the dataset.

Totally ten different runs are performed with the input range from 9 to 90. By taking the different input (i.e. gene expression data), the various accuracy results are obtained as shown in table 1. The table values show that the breast cancer detection accuracy of three different techniques namely GKNE-LGBC technique and existing methods PCA-AE-Ada [1] and deep learning-based multi-model ensemble method [2]. Among them, the proposed GKNE-LGBC technique improves breast cancer detection accuracy using gene expression data. The GKNE-LGBC technique uses the steepest descent light gradient boost classification algorithm to minimize the training error of the normal classifier. The proposed classification algorithm has a set of bivariate regression tree as a normal classifier i.e. base classifier to distinguish the gene expression data as relapse or nonrelapse with selected relevant genes based on the correlation between the training data with cancerous data. The coefficient provides a positive correlation, and then the data is classified as a relapse. Otherwise, the regression tree classified the data as nonrelapse. Then the base classifier results are summed and to make a strong one by minimizing the error in the classification process. As a result, the GKNE-LGBC technique effectively finds the cancerous data which improving the breast cancer detection accuracy.

The results of GKNE-LGBC technique is compared to the existing techniques. The comparison results prove that the GKNE-LGBC technique improved the breast cancer detection accuracy by 8% as compared to PCA-AE-Ada [1] and 16% as compared to deep learning-based multi-model ensemble method [2] respectively.

5.2 Performance analysis of the false positive rate

The false positive rate is calculated as the ratio of the number of gene expression data are incorrectly classified as normal or abnormal to the total number of data taken as input. The mathematical formula for calculating the false positive rate is given below,

$$FPR = \left[\frac{Number\ of\ data\ incorrectly\ classified\ as\ normal\ or\ abnomral}{n}\right] i100 \quad (10)$$

In (10),FPR represents the false positive rate, n denotes a number of gene expression data. The false positive rate is measured in percentage (%).

Calculation:

Proposed GKNE-LGBC: Number of gene expression data incorrectly classified is 1 and the total number gene expression data is 9. The false positive rate is calculated as follows,

$$FPR = \frac{1}{9} * 100 = 11\%$$

Existing PCA-AE-Ada: Number of gene expression data incorrectly classified is 2 and the total number gene expression data is 9. The false positive rate is calculated as follows,

$$FPR = \frac{2}{9} * 100 = 22\%$$

Existing deep learning-based multi-model ensemble method: Number of gene expression data incorrectly classified is 3 and the total number gene expression data is 9. Then the false positive rate is calculated as follows,

$$BCDA = \frac{3}{9} * 100 = 33\%$$

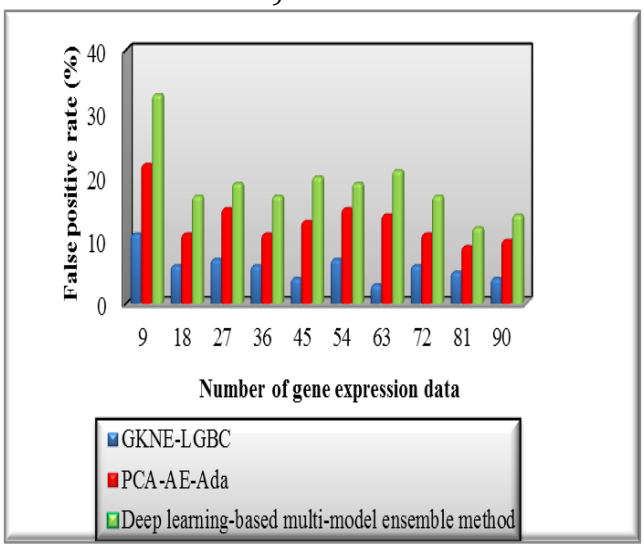

Figure 4 Performance results of false positive rate

Figure 4 depicts the experimental results of false positive rate with respect to a number of gene expression data. As shown in the graph, the number of data is given to the 'x' axis and the results are obtained at the 'y' axis. The graphical results illustrate that the false positive rate of GKNE-LGBC technique is minimized than the state-of-the-art methods. These significant results are obtained by minimizing the training error of the weak classifier. By applying an ensemble learning algorithm, the mean square error of the weak classifier is calculated based on the actual and predicted results. The proposed boosting technique uses the steepest descent optimization technique to find the minimum of an objective function i.e. error rate. This helps to minimize the incorrect classification of the gene expression data as compared to the existing techniques. The ten various simulation results of GKNE-LGBC technique and existing methods are compared. Then the average results confirm that the proposed technique minimizes the false positive rate by 55% and 68% as compared to the state-of-the-art methods.

5.3 Performance analysis of time complexity

Time complexity is defined as the amount of time required to identify the breast cancer through the classification of gene expression data. The mathematical formula for calculating the time complexity is expressed as follows,

$$TC = n * T(classifying\ one\ data) \quad (11)$$

From equation (11), TC represents the time complexity, n denotes a number of gene expression data , T denotes a time for classifying one data. The time complexity is measured in

milliseconds (ms). The sample mathematical calculations for the time complexity of three techniques are given below.

Calculation:

Proposed GKNE-LGBC: Number of gene expression data is 9 and the time taken for classifying the single gene expression data is 1.4ms.

Then the overall time complexity is calculated as follows,

$$TC = 9 * 1.4ms = 12.6ms \cong 13ms$$

Existing PCA-AE-Ada: Number of gene expression data is 9 and the time taken for classifying the single gene expression data is 1.6ms. Then the overall time complexity is calculated as follows,

$$TC = 9 * 1.6ms = 14ms$$

Existing Deep learning-based multi-model ensemble method: Number of gene expression data is 9 and the time taken for classifying the single gene expression data is 1.9ms. Then the overall time complexity is calculated as follows,

$$TC = 9 * 1.9ms = 17ms$$

Table 2 time complexity under different gene expression data

Number of gene expression data	Time complexity (ms)		
	GKNE-LGBC	**PCA-AE-Ada**	**Deep learning-based multi-model ensemble method**
9	13	14	17
18	14	18	23
27	18	22	27
36	21	24	29
45	25	29	34
54	28	33	38
63	30	35	39
72	33	37	42
81	37	41	45
90	45	51	58

As shown in table 2, the performance results of time complexity are reported with the number of gene expression data. The reported result shows that the GKNE-LGBC technique minimizes breast cancer detection time than the other two methods. This is due to the application of the relevant gene selection before the classification process. The microarray datasets contain a number of genes. While performing the classification with these genes, cancer detection time gets increased. This helps to improve the risk level for identifying disease patients. This problem is overcome by the GKNE-LGBC technique by selecting the relevant genes from the breast cancer microarray dataset. The gene selection is performed with the help of Gaussian kernelized stochastic neighbor embedding technique. The feature selection technique measures the distance similarity to identify the relevant genes for cancer detection among the number of genes. Then the threshold is set for identifying the relevant and irrelevant features. Finally, the classification is performed with the selected genes and identifies breast cancer with less time complexity. The performance results of time complexity are minimized by 14% and 26% using GKNE-LGBC technique as compared to the PCA-AE-Ada [1] and deep learning-based multi-model ensemble method [2].

The above parameter discussion clearly shows that the GKNE-LGBC technique accurately identifies breast cancer with minimum time by using gene expression data.

VI. CONCLUSION

The accurate detection of breast cancer facilitates the precision cancer diagnosis at an earlier stage. The machine learning methods have been broadly used in cancer prediction than the other methods. In this paper, an efficient machine learning technique called GKNE-LGBC is developed for cancer detection. Specifically, the gene expression data obtained from the microarray dataset. Reducing the redundant or irrelevant genes from the gene expression datasets effectively minimize the time complexity of classification. In GKNE-LGBC technique, the Gaussian kernelized stochastic neighbor embedding algorithm is applied for relevant genes selection to increase the performance of classification. The results show that relevant gene selection minimizes the dimensionality of data. Then the ensemble learning algorithm is applied to categorize the gene expression data into the normal or cancerous samples with the selected relevant genes resulting in improving the breast cancer detection accuracy with minimum training error. The experimental evaluation is performed using GKNE-LGBC technique and existing techniques with benchmark microarray dataset. The observed results show that the GKNE-LGBC technique achieves better results in terms of breast cancer detection accuracy and minimizes the time complexity as well as false positive rate as compared to the state-of-the-art- methods.

VII. REFERENCES

1) Dejun Zhang, Lu Zou, Xionghui Zhou, Fazhi He, “Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer”, IEEE Access, Volume 6, 2018, Pages 28936 – 28944
2) Yawen Xiao, Jun Wu, Zongli Lin, Xiaodong Zhao, “A deep learning-based multi-model ensemble method for cancer prediction”, Computer Methods and Programs in Biomedicine, Elsevier, Volume 153, 2018, Pages 1-9
3) V.Nandagopal, S.Geeitha, K. Vinoth Kumar, J.Anbarasi, “Feasible analysis of gene expression –a computational-based classification for breast cancer”, Measurement, Elsevier, Volume 140, 2019, Pages 120-125
4) Lin Zhang, Hui Liu, Yufei Huang, Xuesong Wang, Yidong Chen, Jia Meng, “Cancer Progression Prediction Using Gene Interaction Regularized Elastic Net”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 14, Issue 1, 2017, Pages 145 – 154
5) Lin Sun, Xiaoyu Zhang, Yuhua Qian, Jiucheng Xu, Shiguang Zhang, “Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification”, Information Sciences, Volume 502, 2019, Pages 18-41

6) Md. Maniruzzaman, Md. Jahanur Rahman, Benojir Ahammed, Md. Menhazul Abedi, Harman S. Suri, Mainak Biswas, Ayman El-Baz, Petros Bangeas, Georgios Tsoulfas, Jasjit S Suri, "Statistical Characterization and ClassiÞcation
7) of Colon Microarray Gene Expression Data using Multiple Machine Learning Paradigms", Computer Methods and Programs in Biomedicine, Elsevier, Volume 176, 2019, Pages 173-193
8) Marta B. Lopes, André Veríssimo, Eunice Carrasquinha, Sandra Casimiro, Niko Beerenwinkel, and Susana Vinga, "Ensemble outlier detection and gene selection in triple-negative breast cancer data", BMC Bioinformatics, Volume 19, Issue 168, 2018, Pages 1-15
9) Mohammadreza Sehhati, Alireza Mehridehnavi, Hossein Rabbani, Meraj Pourhossein, "Stable Gene Signature Selection for Prediction of Breast Cancer Recurrence Using Joint Mutual Information", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 12, Issue 6, 2015, Pages 1440 – 1448
10) Liying Yang, Zhimin Liu, Xiguo Yuan, Jianhua Wei, and Junying Zhang, "Random Subspace Aggregation for Cancer Prediction with Gene Expression Profiles", BioMed Research International, Hindawi Publishing Corporation, Volume 2016, October 2016, Pages 1-10
11) Hanaa Salem, Gamal Attiya, Nawal El-Fishawy, "Early diagnosis of breast cancer by gene expression profiles", Pattern Analysis and Applications, Springer, Volume 20, Issue 2, 2017, Pages 567–578
12) Aman Sharma and Rinkle Rani, "C-HMOSHSSA: Gene Selection for Cancer ClassiÞcation using multi-objective Meta-heuristic and Machine Learning methods", Computer Methods and Programs in Biomedicine, Elsevier, Volume 178, 2019, Pages 219-235
13) Hongling Chen, Mingyan Gao, Ying Zhang, Wenbin Liang, and Xianchun Zou, "Attention-Based Multi-NMF Deep Neural Network with Multimodality Data for Breast Cancer Prognosis Model", BioMed Research International, Hindawi, Volume 2019, May 2019, Pages 1-11
14) Yang Guo, Xuequn Shang, Zhanhuai Li, "IdentiÞcation of Cancer Subtypes by Integrating Multiple Types of Transcriptomics Data with Deep Learning in Breast Cancer", Neurocomputing, Elsevier, Volume 324, 2019, Pages 20-30
15) SungHwan Kim, "Weighted K‑means support vector machine for cancer prediction" Springer Plus, Volume 5, 2016, Pages 1-11
16) Saleh Shahbei, Akbar Rahideh, Mohammad Sadegh Helfroush, Kamran Kazemi, "Gene expression feature selection for prostate cancer diagnosis using a two-phase heuristic–deterministic search strategy", IET Systems Biology, Volume 12 , Issue 4 , 2018, Pages 162 - 169
17) Thanh Nguyen, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi, "Hidden Markov models for cancer classification using gene expression profiles", Information Sciences, Elsevier, Volume 316, 2015, Pages 293-307
18) Sara Tarek, Reda Abd Elwaha, Mahmoud Shoman, "Gene expression based cancer classification", Egyptian Informatics Journal, Elsevier, Volume 18, Issue 3, 2017, Pages 151-159
19) Bikesh Kumar Singh, Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm", Biocybernetics and Biomedical Engineering, Elsevier, Volume 39, Issue 2, 2019, Pages 393-409
20) Alok Kumar Shukla, Pradeep Singh, Manu Vardhan, "A New Hybrid Wrapper TLBO and SA with SVM Approach for Gene Expression Data", Information Sciences, Elsevier, 2019, Pages 1-26
21) Yuanyu He Junhai Zhou , Yaping Lin, Tuanfei Zhu, "A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data", Computational Biology and Chemistry, Elsevier, Volume 80, June 2019, Pages 121-127