Jeyasingh and Veluchamy, J Cancer Sci Ther
2018, 10:8
DOI: 10.4172/1948-5956.1000549

Journal of Cancer Science & Therapy

Review Article Open Access

Polytomous Logistic Regression Based Random Forest Classifier for Diagnosing Cancer Disease

Suganthi Jeyasingh[1*] **and Malathy Veluchamy**[2]

[1]*Department of Computer Science and Engineering (CSE), Raja College of Engineering and Technology, Madurai, Tamil Nadu, India*

[2]*Anna University Regional Centre, Madurai, Tamil Nadu, India*

[*]**Corresponding author:** Suganthi Jeyasingh, Department of Computer Science and Engineering (CSE), Raja College of Engineering and Technology, S V Raja Nagar, Veerapanjan, Madurai-625020, Tamil Nadu, India, Tel: +04522429280; E-mail: gksuganthi123@gmail.com

Rec date: June 15, 2018; **Acc date:** August 13, 2018; **Pub date:** August 16, 2018

Abstract

An early prediction of cancer disease is important for successful treatment. Recently, many research works have been designed for classification of cancer diseases and early detection. Performance of classifying different stages of cancer diseases was not efficient. In order to overcome such limitation, Multinomial Logistic Regression Based Random Forest Classifier (MLR-RFC) technique is proposed to improve the performance of prediction with higher classification accuracy. Initially, MLR-RFC technique performs pre-processing task to extract the relevant features for disease prediction and to reduce the classification time. Then, MLR-RFC technique applies Random Forest Classifier for predicting the cancer disease with higher disease prediction rate. Finally, MLR-RFC technique uses MLR based Random Forest Classifier for classifying different stages of cancer disease as normal, moderate, mild and severe with higher classification accuracy. The performance of MLR-RFC technique is measured in terms of disease prediction rate, classification accuracy and classification time by using two data sets namely Breast cancer dataset and Wisconsin Breast cancer dataset. The experimental result shows that the MLR-RFC technique is able to improve the disease prediction rate and classification accuracy when compared to state-of-the-art-works.

Keywords: Breast Cancer; Classification; MLR based Random Forest Classifier (MLR-RFC); Pre-processing

Introduction

Classification is one of the most important tasks in machine learning and data mining. In medical diagnosis, classification model attains great attention for cancer pattern identification. The early detection of cancer is highly useful in curing the disease. Recently, lot of research have been designed by using data mining and machine learning techniques to classify cancer disease. A Random forest classifier is designed by combining the classifier and feature selection technique to improve the prediction accuracy for breast cancer diagnosis and prognosis [1]. However, multi class classification is not considered. A Hybrid Adaptive Ensemble Learning (HAEL) framework [2] is developed for cancer disease classification with higher classification accuracy. But, the classification time is high.

A Cost-Sensitive Classifier with Gentle Boost Ensemble is introduced for classification [3]. This system reduced the misclassification costs and enhanced the overall classification performance. But, classification accuracy is not at required level. Adaptive Penalized Logistic Regression [4] is applied to accurately analyze high-dimensional DNA microarray data for cancer classification with lower misclassification rate and to improve the classification accuracy in which, high-dimensional multi class classification cancer data was not considered. Shuffled Frog Leaping with Levy Flight (SFLLF) [5] that employs k-Nearest Neighbor (k-NN) technique is developed to classify cancer diseases. However, SFLLF takes more time for disease classification. A novel method [6] is intended to solve the multiclass imbalance classification problem of cancer microarray data with the support of ensemble classifier and to provide more balanced and robust classification results though, classification performance is not efficient.

An artificial neural network ensemble [7] is designed with the histogram of oriented gradient genomic features for automated screening and early prediction of lung cancer with higher accuracy. A Genetic Bee Colony (GBC) algorithm [8] is presented for considering the gene selection problem in both binary and multi-class cancer classification. The GBC algorithm acquired highest classification accuracy with average number of selected genes. However, the time complexity remained unaddressed. A re-balancing sample method [9] is introduced to discover novel genes related to individual cancers by using a two-step logistic regression. However, time complexity remained unaddressed. A novel strategy based on multi-tasking learning methods [10] is intended to identify the significant genes from large gene expression datasets to predict cancers.

The Probabilistic Neural Network is applied for the efficient detection and classification of the breast cancer [11]. The classifier is trained with the features obtained from the Wisconsin cancer dataset and tested. A novel algorithm is developed by integrating Cuckoo Optimization Algorithm (COA) and Genetic Algorithm (GA) [12] to achieve high cancer classification performance. A better global minimum is obtained with only little iteration. A novel approach is proposed by integrating the feature selection approach and transductive Support Vector Machine based classification [13]. A forward greedy search algorithm is used to obtain the potential gene markers. The proposed approach has successfully used unlabeled gene expression data and achieved better empirical success. However, if the labeled and unlabeled data follow different distributions, integrating the unlabeled data may lead to poor performance. A prediction scheme that combines Fuzzy Preference Based Rough Set (FPRS)

method for feature selection with the semi-supervised Support Vector Machine (SVM) is developed [14].

The proposed prediction scheme is found to be biologically relevant for the cancer diagnosis and drug discovery. An Incremental Learning Machine is proposed for the classification of cancer [15]. A novel gene selection approach is introduced based on the substantial modification of Analytic Hierarchy Process (AHP) [16]. A semi-supervised projective non-negative matrix factorization method [17] is proposed for achieving higher cancer classification performance. A Multiplicative Update Rule (MUR) is developed to optimize the proposed method and proved the convergence property. A Binary Quantum-Behaved Particle Swarm Optimization (BQPSO) is proposed for the selection of cancer feature genes and combined with the SVM for the cancer classification [18].

The Computer-aided diagnostic (CAD) is proposed for the automatic segmentation and classification of each lung into normal or cancer [19]. The proposed CAD achieved 95% accuracy in case of the linear classifier. A Graphical User Interface is designed and Naïve Bayes classifier is used to detect the probability of occurrence of Breast cancer among women [20]. The existing classification methods do not achieve better prediction accuracy and are computationally expensive.

In order to overcome the existing issues, Multinomial Logistic Regression Based Random Forest Classifier (MLR-RFC) technique is designed. The main objective of the proposed MLR-RFC technique is to improve the classification and prediction accuracy of cancer disease diagnosis for early detection. The research objective of the proposed technique is formulated as follows:

1. To remove redundant, unwanted, noisy attributes in data set and to reduce the classification time.

2. To improve the prediction accuracy of cancer disease, by employing Random Forest Classifier.

3. To efficiently classify the different stages or levels of cancer disease with higher classification accuracy, by applying the MLR based Random Forest Classifier.

The rest of the sections in this paper are organized as follows.

- Section 2 presents Multinomial Logistic Regression Based Random Forest Classifier (MLR-RFC) technique for prediction and classification of cancer diseases.
- Section 3 and Section 4 presents the experimental section with details performance analysis.
- Section 5 explains the related works.
- Finally, Section 6 concludes this research finding.

Experimental Techniques

Multinomial Logistic Regression Based Random Forest Classifier (MLR - RFC)

The MLR-RFC technique is designed to improve the classification performance of cancer disease for early prediction. The MLR-RFC technique initially takes breast cancer data set as input and then performs pre-processing to extract useful features from input data. Next, Random Forest Classifier is applied for predicting the cancer disease with higher prediction rate. Then, MLR-RFC technique employs Multinomial Logistic Regression Based Random Forest Classifier (MLR based Random Forest Classifier) for classifying the cancer disease with different levels such as normal, moderate, mild and severe. This in turn helps for predicting the cancer disease in an early detection. The overall architecture diagram of MLR-RFC technique is shown in Figure 1.

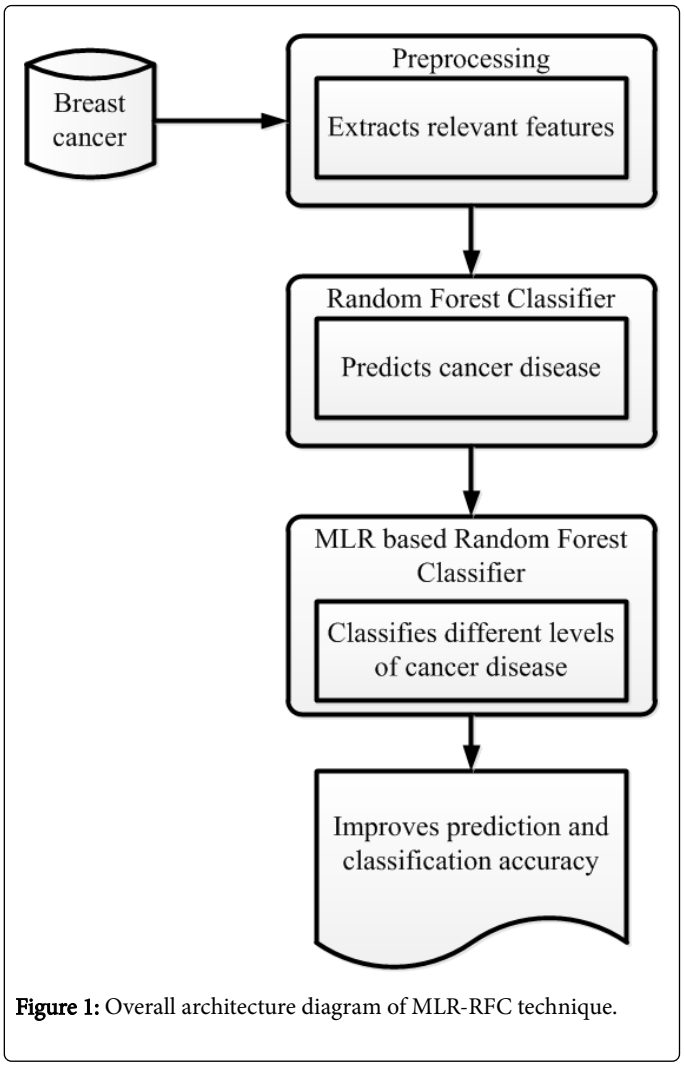

Figure 1: Overall architecture diagram of MLR-RFC technique.

As shown in Figure 1, the MLR-RFC Technique takes Breast Cancer dataset as input and then performs preprocessing for extracting relevant features from data set. This preprocessing task removes the unwanted and redundant attributes (i.e., features) and efficiently mines relevant features for disease prediction which in turn assists for reducing classification time. Finally, MLR-RFC Technique applied MLR based Random Forest Classifier is designed for classifying the different diseased level of cancer, resulting in improved classification accuracy. The detailed explanation about the proposed MLR-RFC Technique is described in following sections.

Pre-processing

The MLR-RFC Technique performs preprocessing with the aid of Chi squared test to reduce the classification time and space complexity for disease diagnosis. The Chi squared test is the statistical sampling distribution test that is employed to discover the correlation relationship between two attributes for reducing the redundant features for disease prediction. The following figure illustrates

preprocessing using Chi squared test is shown in Figure 2. The chi squared test also called as χ^2 test is used for extracting the redundant attributes for the disease diagnosis. The sampling distribution of the analysis value is a chi-squared distribution while the null hypothesis is true. A chi-squared test is employed to remove the features that are irrelevant. Besides, the chi squared method is used in MLR-RFC technique for mining the relevant attribute and reducing the redundant attribute for performing disease diagnosis. The chi squared method selects features which is relevant to the class vectors and eliminates the rest of the features. While discovering the most relevant features, the chi-squared distribution in the MLR-RFC Technique reduces the redundancy among the selected attributes. Let us assume the discretization factor 'χ^2' using interval measured in the proposed MLR-RFC Technique formulates an accurate number of intervals, therefore changing the continuous data values into the discrete values which is represented as given below,

$$x^2 = \sum \frac{(observed - Expected)^2}{Expected} \quad (1)$$

$$x^2 = \sum_{i=1}^{c} \sum_{j=1}^{n} \frac{\left(Oi_j - E_{ij}\right)^2}{E_{ij}} \quad (2)$$

From equations (1) and (2), discretization factor χ^2 denotes the difference between Oij at corresponds to the two features that related with the cancer disease in 'ith' interval, 'jth' class and the expected count 'Eij'. The expected count is evaluated by using mathematical formula,

$$E_{ij} = \frac{\left(A = a_i\right) * count\,(B = b_i)}{N} \quad (3)$$

From equation (3), the expected count is obtained by means of the product of two features in the 'ith' interval and jth class with respect to the total number of patients 'N'. Larger value of χ^2 is more likely the attributes that correlated to diagnose the disease. Chi-square determines the variation between collected counts and expected counts. If the difference is large, it presents a significant change and allows removing the redundant attributes. If the scores of the two features are too similar, then it is concluded that they are relevant features.

Pre-processing Algorithm using Chi Squared Method
Input: Set of features in Breast Cancer dataset
Output: Obtain the relevant feature and reduce the redundant feature
Step 1: Begin
Step 2: For each attribute in objects
Step 3: Compute the discretization factor using (2) and (3)
Step 4: Obtain the relevant attributes
Step 5: Remove the redundant attributes
Step 6: End for
Step 7: End

Table 1: The algorithmic process of pre-processing using Chi-square method.

For each feature from breast cancer dataset, the chi squared pre-processing algorithm estimates the discretization factor and then modified factor depends on the equivalence measure with objective of removing the relevant attributes. This in turn helps for reducing classification time. The algorithmic process of pre-processing using chi squared method is shown below Table 1.

With the help of the above algorithmic process, MLR-RFC Technique obtains the relevant attributes in the dataset for disease diagnosis. For each attribute, the similarity between the two attributes is determined with the help of chi squared method. The higher value of χ^2 provides the more related attributes for disease prediction. This in turn helps MLR-RFC Technique to select the highlevel features to perform the cancer disease diagnosis.

Random forest classifier for disease prediction

After selecting the relevant features, Random Forest Classifier is used for predicting the occurrences of cancer disease. It is an ensemble classifier that consists of many decision trees where the final predicted class for a test example is obtained by combining the predictions of all individual trees. It combines the bootstrap aggregation and random feature selection to make a collection of decision trees exhibiting controlled variation. The training set for each individual tree in a random forest is constructed by sampling N examples at random with replacement from the N available examples in the dataset. This is called as bootstrap sampling and bagging explains the aggregation of predictions from the resulting collection of trees. As a result of the bootstrap sampling procedure, approximately one third of the N examples do not exist in the training set of each tree. This is known as the "out-of-bag" data of the tree for which internal test predictions is to be performed. By combining the out-of-bag data predictions of all trees, an internal estimate of the generalization error of the random forest is identified.

For every node in a tree, d≪D features are randomly chosen from the dataset and the node is divided with the aid of the best possible binary split. A parent node np is split into child nodes nr. nr using Gini index that measures the likelihood of incorrect labeling of an instance, if it is randomly classified based on the distribution of labels within the node. For a binary split, the Gini index of a node 'n' is expressed as

$$I_G(n) = 1 - \sum_{c=1}^{2} p_c^2 \quad (4)$$

From equation (4), pc is the relative proportion of examples that belong to the class c, present in node n. The best possible binary split is the one that increases the improvement in the Gini index which is formulated as follows:

$$\Delta I(n_p) = I_G(n_p) - p_l I_G(n_l) - p_r I_G(n_r) \quad (5)$$

From equation (5), pl and pr denote the proportions of examples in the node np that are assigned to the child nodes nl and nr respectively. The Gini index is used to obtain the relative importance of features for classification. A measure of the importance of an individual feature is calculated by summing the decrease in the Gini index occurring at all nodes in the forest that are divided based on that feature. The structure of Random forest classifier for disease prediction is shown in Figure 2. A random forest classifier consists of N decision trees trained by a completely random approach. For each decision tree Tn the features are selected randomly from the pool of training sample. It is a subset of

all the training samples. After N trees are created, the final decision combines all the outputs of tree by means of considering the average of all N outputs. In decision tree, each node measures Gini index that split the features to be classified. The training step aims to discover an estimate depends on training data of the posterior distribution over the classes in each leaf. From Figure 3, each node is divided based on a single feature and each branch is finished in a terminal node. Terminal nodes afford a prediction for the class of a test example based on the path taken through the tree. The color of a terminal node indicates its class prediction. The final predicted class for a test example is obtained by combining the predictions of all individual trees.

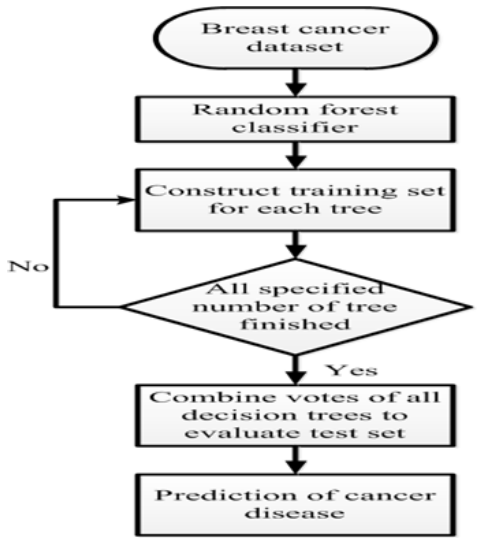

Figure 2: Flow diagram of random forest classifier for prediction of cancer disease.

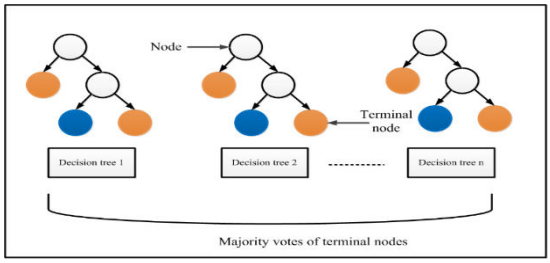

Figure 3: Structure of random forest classifier.

In the training case, a random forest comprises multiple binary trees. Each binary tree yields a different partition of the attribute space. Each node chooses the best point pair as the best weak classifier and the random forest combines the results of each weak classifier to a strong classifier which is mathematically formulated as below,

$$F(d) = \arg\max d_c(d) \quad (6)$$

$$F(d) = arg\ max\left(\frac{1}{N}\right)\sum_{n=1,2\ldots n} d_n,\ d(f(d)=c) \quad (7)$$

From (6) and (7), N is the total number of binary trees and n represents single binary tree, d is the data attribute (i.e., feature) to be classified and C is the label of class. Here, dn, d is the probability classified by the nth binary tree that the data attribute belong to the disease. The class value of '0' and '1' is efficiently used to differentiate between the presence and absence of cancer disease. In MLR-RFC Technique c=1 denotes the data attribute of patient influenced by cancer disease whereas c=0 signifies the absence disease. Figure 4 shows the flow diagram of the Random forest classifier process for the prediction of cancer disease.

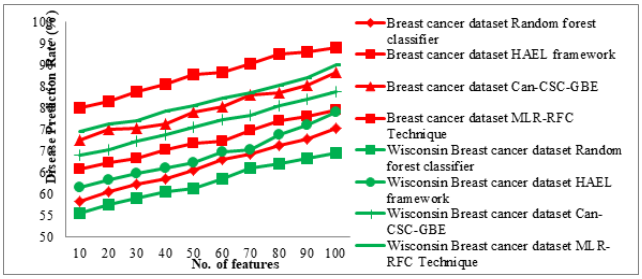

Figure 4: Measure of disease prediction rate.

As shown in Figure 4, random forest classifier initially takes breast cancer data set as input and constructs multiple binary trees for available attributes in data set. After N trees are constructed, Random forest combines the results of each weak classifier to a strong classifier with the aid of equation (4) for predicting the cancer disease. Therefore, MLR-RFC Technique improves the prediction accuracy cancer diseases in an efficient manner. After predicting the occurrence of cancer disease, Multinomial Logistic Regression based Random Forest Classifier is used for classifying different level of cancer diseases such as normal, moderate, mild and severe with higher classification accuracy.

Multinomial Logistic Regression based Random Forest Classifier (MLR-RFC) for Disease Classification

Multinomial Logistic Regression (MLR) based Random Forest Classifier is a classification technique which generalizes a binary logistic regression model to a multiclass problem. It is used to predict the probabilities of possible outcomes of a categorical response variable for a set of independent variables. The MLR based Random Forest Classifier is an attractive model for analysis since it does not assume normality, linearity, or homoscedasticity. The MLR based Random Forest Classifier does have assumptions. Besides, assumes non-perfect separation. If the collections of response variable are efficiently separated via the predictor(s), then unrealistic coefficients is evaluated and the effect of size will be significantly enlarged.

There are diverse parameter estimation methods based on the inferential objectives of MLR based Random Forest Classifier analysis. The generalized linear modeling technique of MLR based Random Forest Classifier is used to model unordered categorical response variables. It is a simple extension of logistic regression. This method permits comparison of each category of unordered response variable to a random reference category providing a number of logistic regression models. This procedure yields a number of logistic regression models that formulate specific comparisons of the response categories. The MLR based Random Forest Classifier contains j–1 logic equations, while there are j categories of the response variable.

The MLR based Random Forest Classifier classifies d-dimensional real-valued input vectors x ∈ Rd into one of the k outcomes c ∈ {0,...,k-1} using k–1 parameter vector $\beta_0,...,\beta(k-2) \in Rd$. Let the response variable Y ∈ {1,...,k}

have k possible values (categories). A common representation of the MLR based Random Forest Classifier model is mathematically formulated as,

$$p(Y = r \mid x) = \frac{\exp(x^T \beta_r)}{\sum_{s=1}^{k} \exp(x^T \beta_r)} = \frac{\exp(n_x)}{\sum_{s=1}^{k} \exp(n_x)} \quad (8)$$

Where $\beta_r^T = (\beta_{r0}, \ldots \beta_{rp})$

It is obvious that one has to specify some additional constraints since the parameters $\beta_{1....}^{t} \ldots \beta_k^T$ are not identifiable. An often active side constraint is based on selecting a reference category (RSC). When category k is chosen, one set $\beta_k^T = (0, \ldots, 0)$ yields ηk=0. Fitting a model using a reference category, the corresponding model is formulated as, $p(Y = r \mid x) = \frac{\exp(x^T \beta_r)}{1 + \sum_{s=1}^{q} \exp(x^T \beta_s)} for\ r = 1, \ldots \ldots q$ (9)

Multinomial Logistic Regression based Random Forest Classifier Algorithm **Input**: **Breast Cancer Data set**
Output: Improved Classification and Prediction Accuracy
Step 1: Begin
Step 2: For attributes in data set
Step 3: Constructs decision tree
Step 4: For decision tree
Step 5: Features are selected randomly from the training sample pool
Step 6: Parent node is partitioned into child nodes according to the Gini index using (4) and (5)
Step 7: If (N number of trees is constructed) then
Step 8: Random forest classifiers combines the results of all weak classifier to a strong classifier for disease prediction using (6) and (7)
Step 9: MLR based Random Forest Classifier modelis applied for classifying multiple levels of cancer disease as normal, moderate, mild and severe using (8) and (9)
Step 10: End if
Step 11: End for
Step 12: End for
Step 13: End

Table 2: The algorithmic process of multinomial logistic regression based random forest classifier for classification of cancer disease.

By using equation (8) and (9), Multinomial Logistic Regression based Random Forest Classifier efficiently classifies the diverse level of cancer disease as normal, moderate, mild and severe for early detection. This in turn helps the MLR-RFC Technique to improve the classification accuracy with minimum time.

The algorithmic process of Multinomial Logistic Regression based Random Forest Classifier for Classification of Cancer Disease is shown below Table 2.

With the help of above algorithmic process, MLR-RFC Technique efficiently predicts and classifies the diverse level of cancer disease as normal, moderate, mild and severe. As a result, MLR-RFC Technique improves the prediction and classification accuracy of cancer diseases with minimum time.

Experimental Setting

The proposed Multinomial Logistic Regression Based Random Forest Classifier (MLR-RFC) technique is implemented using JAVA language with Breast cancer dataset [21] and Wisconsin Breast cancer dataset [22] obtained from UCI Machine learning repository. The Breast cancer dataset is extracted from the UCI repository, which contains 201 instances of one class and 85 instances of other class. The instances are characterized through 9 attributes, some of which are linear and some are nominal. The Wisconsin Breast cancer dataset is extracted from UCI repository, which has 699 instances and 10 number of attributes. The effectiveness of MLR-RFC Technique is compared against with existing Random forest classifier [1], Hybrid Adaptive Ensemble Learning (HAEL) framework [2] and Cost-Sensitive Classifier with Gentle Boost Ensemble (Can-CSC-GBE) [3].

Results and Discussion

In this section, the result analysis of MLR-RFC Technique is evaluated. The performance of MLR-RFC Technique is compared against with existing Random forest classifier, Hybrid Adaptive Ensemble Learning (HAEL) framework and Cost-Sensitive Classifier with Gentle Boost Ensemble (Can-CSC-GBE). The performance of MLR-RFC Technique is evaluated along with the metrics such as disease prediction rate, classification accuracy and classification time.

Measurement of disease prediction rate

In MLR-RFC Technique, disease prediction rate is defined as the ratio of number of features correctly predicted as disease to the total number of features. The disease prediction rate is measured in terms of percentage (%) and mathematically formulated as,

$$Disease\ Prediction\ Rate = \frac{number\ of\ features\ correctly\ predicted}{total\ number\ of\ features} * 100 \quad (10)$$

From the equation (10), disease prediction rate of cancer is obtained. While disease prediction rate is higher, the method is said to be more efficient.

Table 3 depicts the comparative result analysis of disease prediction rate of cancer diseases with respect to two different dataset namely Breast cancer dataset and Wisconsin Breast cancer dataset. The numbers of features are considered as input for cancer disease diagnosis and it varies from 10 to 100. From the table value, it is illustrated, that the disease prediction rate of proposed MLR-RFC Technique using Breast cancer dataset and Wisconsin Breast cancer dataset is higher when compared to other data set and existing methods.

No. of features	Disease Prediction Rate (%)							
	Breast cancer dataset				Wisconsin Breast cancer dataset			
	Random forest classifier	HAEL framework	Can-CSC-GBE	MLR-RFC Technique	Random forest classifier	HAEL framework	Can-CSC-GBE	MLR-RFC Technique
10	58.14	65.87	72.54	80.12	55.54	61.46	68.95	74.48
20	60.55	67.14	74.89	81.45	57.46	63.17	70.14	76.19
30	62.18	68.33	75.13	83.65	58.96	64.78	72.36	77.12
40	63.42	70.15	76.14	85.47	60.48	65.95	73.65	79.23
50	65.48	71.64	78.93	87.69	61.26	67.23	75.61	80.45
60	67.89	72.17	80.17	88.13	63.58	69.75	77.18	82.36
70	69.15	74.63	82.98	90.14	65.93	70.15	78.16	83.47
80	71.35	76.91	83.41	92.47	67.12	73.77	80.47	85.35
90	72.68	78.12	85.12	93.11	68.36	76.12	81.96	86.96
100	75.13	79.38	88.23	94.05	69.48	78.98	83.65	89.91

Table 3: Disease prediction rate.

Figure 5 portrays the impact of disease prediction rate for cancer disease diagnosis with respect to different number of features in the range of 10 to 100. Two different datasets are used to perform the experimental results based on four methods. In the figure, the red color line indicates the disease prediction rate for Breast cancer dataset, whereas green color line indicates Wisconsin Breast cancer dataset. Among the two datasets, the proposed MLR-RFC Technique using Breast cancer dataset provides better disease prediction rate for cancer disease diagnosis when compared to other data set and existing methods.

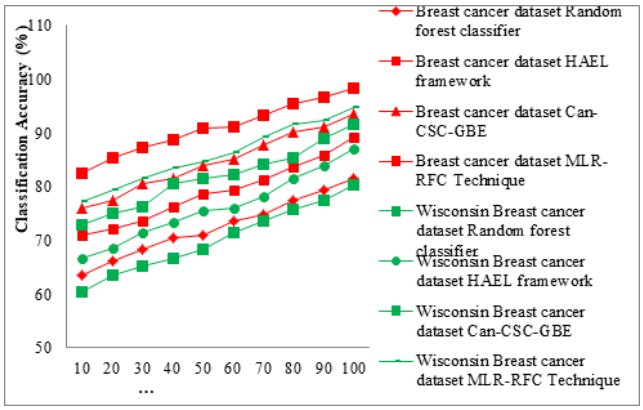

Figure 5: Measure of classification accuracy.

Besides, while increasing the number of features, the disease prediction rate also gets increased using all the four methods.

But comparatively, the disease prediction rate of proposed MLR-RFC Technique using Breast cancer dataset is higher. This is due to the application of Random Forest Classifier in MLR-RFC Technique where it combines the results of all weak classifier to strong classifier for efficiently predicting the occurrences of cancer disease. This in turn helps for improving the disease prediction rate in a significant manner. As a result, MLR-RFC Technique improves the disease prediction rate by 32%, 21%, and 9% with Breast cancer dataset when compared to existing Random forest classifier [1], HAEL framework [2] and Can-CSC-GBE [3] respectively. Similarly for Wisconsin Breast cancer dataset, MLR-RFC Technique improves the disease prediction rate by 30%, 18%, and 7% as compared to the existing Random forest classifier, HAEL framework and Can-CSC-GBE respectively.

Measurement of classification accuracy

In MLR-RFC Technique, classification accuracy is defined as the ratio of the difference between the correctly classified feature and incorrectly classified feature to the total number of features. Classification accuracy is also called precision, measured in terms of percentage (%). Classification accuracy of cancer disease is obtained using

$$Classification\ accuracy = \frac{Correctly\ classifed\ features\ as\ disease}{Total\ number\ of\ feature} \times 100 \quad (11)$$

The comparative result analysis of classification accuracy for cancer diseases diagnosis with respect to two dissimilar datasets namely Breast cancer dataset and Wisconsin Breast cancer dataset is presented in Table 4. From the table value, it clear that the classification accuracy of proposed MLR-RFC Technique using Breast cancer dataset is higher as compared to other data set and existing methods.

No. of features	Classification Accuracy (%)							
	Breast cancer dataset				Wisconsin Breast cancer dataset			
	Random forest classifier	HAEL framework	Can-CSC-GBE	MLR-RFC Technique	Random forest classifier	HAEL framework	Can-CSC-GBE	MLR-RFC Technique
10	63.56	70.87	76.12	82.45	60.32	66.56	72.95	77.25
20	66.12	72.24	77.46	85.47	63.65	68.67	75.14	79.34
30	68.36	73.63	80.57	87.36	65.27	71.45	76.36	81.55
40	70.46	76.37	81.63	88.78	66.78	73.39	80.65	83.46
50	71.1	78.56	83.85	90.96	68.26	75.5	81.61	84.71
60	73.69	79.41	85.23	91.25	71.58	76.02	82.18	86.23
70	74.75	81.32	87.89	93.42	73.65	78.3	84.16	89.12
80	77.36	83.81	90.15	95.36	75.72	81.44	85.47	91.68
90	79.46	85.93	91.23	96.74	77.39	83.94	88.96	92.41
100	81.52	89.3	93.56	98.25	80.38	87.14	91.65	94.68

Table 4: Comparison of classification accuracy.

Figure 6 represents the impact of classification accuracy for cancer disease diagnosis with respect to the diverse number of features in the range of 10 to 100. Two different datasets are employed to perform the experimental results based on four methods. In the figure, the red color line point outs the classification accuracy for Breast cancer dataset, whereas green color line designates Wisconsin Breast cancer dataset. Among the two datasets, the proposed MLR-RFC Technique using Breast cancer dataset provides better classification accuracy for cancer disease diagnosis when compared to other data sets and existing methods.

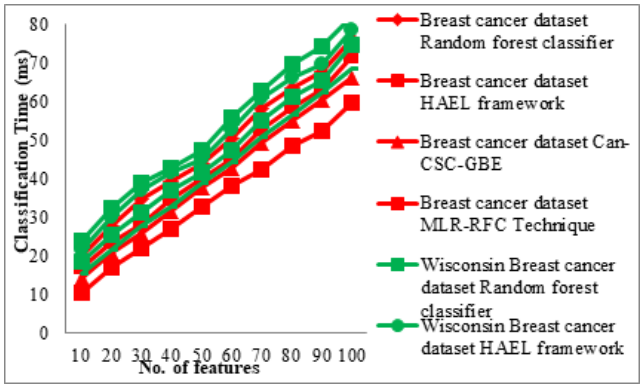

Figure 6: Measure of classification time.

In addition, while increasing the number of features, the classification accuracy also gets increased using all the four methods. But comparatively, the classification accuracy of proposed MLR-RFC Technique using Breast cancer dataset is higher. This is owing to the application of MLR based Random Forest Classifier in MLR-RFC Technique where it predicts the probabilities of different possible outcomes of a categorical response variable for a given data set.

This in turn assists for improving the classification accuracy in an efficient manner. Therefore, MLR-RFC Technique improves the classification accuracy by 26%, 15%, and 7% with Breast cancer dataset when compared to existing Random forest classifier, HAEL framework and Can-CSC-GBE respectively. Similarly for Wisconsin Breast cancer dataset, MLR-RFC Technique improves the classification accuracy by 23%, 13%, and 6% as compared to existing Random forest classifier, HAEL framework and Can-CSC-GBE respectively.

Measurement of classification time

In MLR-RFC Technique, classification time measures the amount of time taken for classifying the diseases effectively, by using MLR based Random Forest Classifier. The classification time is measured in terms of milliseconds (ms) and formulated as,

Classification Time=Time (Classify the feature as disease) (12)

The result analysis of classification time taken for cancer diseases diagnosis using two dataset namely Breast cancer dataset and Wisconsin Breast cancer dataset is demonstrated in Table 5. Different number of features is considered as input for classifying the multiple levels of cancer diseases and it varies from 10 to 100.

From the table value, it is expressive that the classification time of proposed MLR-RFC Technique using Breast cancer dataset is lower as compared to other data set and existing methods.'

Figure 6 presents the impact of classification time for cancer disease diagnosis based on dissimilar number of features in the range of 10 to 100. The two different datasets are utilized to perform the experimental results based on four methods. Among the two datasets, the proposed MLR-RFC Technique using Breast cancer dataset provides better classification time for cancer disease diagnosis as compared to other data set and existing methods.

No. of features	Classification Time (ms)							
	Breast cancer dataset				Wisconsin Breast cancer dataset			
	Random forest classifier	HAEL framework	Can-CSC-GBE	MLR-RFC Technique	Random forest classifier	HAEL framework	Can-CSC-GBE	MLR-RFC Technique
10	19.9	16.8	14.2	10.5	23.7	21.7	18.5	15.3
20	27.7	23.5	20.6	16.8	32.2	30.1	25.4	21.7
30	34.5	28.1	25.7	21.9	38.9	36.8	31.3	27.6
40	39.3	34.7	31.6	27.1	42.8	41.6	36.8	32.8
50	43.7	39.9	37.9	32.7	47.5	45.2	41.7	38.7
60	50.6	45.2	42.8	38.2	55.7	52.7	47.5	44.2
70	58.1	52.7	49.3	42.4	62.6	60.9	55.1	50.9
80	63.6	58.4	55.2	48.6	69.8	66.2	61.3	56.6
90	67.8	63.3	60.4	52.4	74.3	69.7	65.4	62.3
100	76.3	71.8	66.1	59.7	82.1	78.5	74.6	68.4

Table 5: Comparison of classification time.

Furthermore, while increasing the number of features, the classification time also gets increased using all the four methods. But comparatively, the classification time of proposed MLR-RFC technique using Breast cancer dataset is lower. This is because of preprocessing and MLR based Random Forest Classifier is applied in MLR-RFC Technique. The preprocessing task efficiently extracts the relevant attribute and removes the redundant attribute for performing disease prediction. Further, MLR based Random Forest Classifier predicts the probabilities of dissimilar possible outcomes of a categorical response variable for given a data set with minimum time. This in turn assists for reducing the classification time in an effective manner. Thus, MLR-RFC Technique reduces the classification time by 30%, 21%, and 15% with Breast cancer dataset when compared to existing Random forest classifier, HAEL framework and Can-CSC-GBE respectively. Similarly for Wisconsin Breast cancer dataset, MLR-RFC Technique reduces the classification time by 23%, 19%, and 10% as compared to existing Random forest classifier, HAEL framework and Can-CSC-GBE respectively.

Conclusion

An effective Multinomial Logistic Regression Based Random Forest Classifier (MLR-RFC) technique is developed for improving the performance of cancer diseases classification for early detection. At first, MLR-RFC technique accomplishes preprocessing task to remove the irrelevant features and to extract relevant features for disease prediction which in turn helps for reducing the classification time. Then, Random Forest Classifier is used to combine the results of all weak classifiers, to make a strong classifier for disease prediction resulting in improved disease prediction rate. Finally, MLR-RFC technique applies MLR based Random Forest Classifier for classifying different stages of cancer disease as normal, moderate, mild and severe resulting in enhanced classification accuracy. This in turn supports MLR-RFC technique for early detection of cancer disease.

The performance of MLR-RFC technique is tested with the metrics such as disease prediction rate, classification accuracy and classification time. With the experiments conducted for MLR-RFC technique, it is observed that the disease prediction rate of cancer disease provides more accurate results as compared to state-of-the-art works. The experimental results demonstrate that MLR-RFC technique provides better performance with an improvement of disease prediction rate and classification accuracy, when compared to the state-of-the-art works.

References

1. Nguyen C, Wang Y, Nguyen HN (2013) Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomedical Science and Engineering 6: 551-560.
2. Yu Z, Li L, Liu J, Han G (2015) Hybrid adaptive classifier ensemble. IEEE transactions on cybernetics 45: 177-190.
3. Ali S, Majid A, Javed SG, Sattar M (2016) Can-CSC-GBE: Developing cost-sensitive classifier with gentleboost ensemble for breast cancer classification using protein amino acids and imbalanced data. Comput Biol Med 73: 38-46.
4. Algamal ZY, Lee MH (2015) Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Syst Appl 42: 9326-9332.
5. Gunavathi C, Premalatha K (2014) A comparative analysis of swarm intelligence techniques for feature selection in cancer classification. ScientificWorldJournal Article ID 693831.
6. Yu H, Hong S, Yang X, Ni J, Dan Y, et al. (2013) Recognition of multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. Biomed Res Int Article ID 239628.
7. Adetiba E, Olugbara OO (2015) Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. ScientificWorldJournal Article ID 786013.
8. Alshamlan HM, Badr GH, Alohali YA (2015) Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Comput Biol Chem 56: 49-60.

9. Chen B, Shang X, Li M, Wang J, Wu FX (2016) Identifying individual-cancer-related genes by rebalancing the training samples. IEEE transactions on nanobioscience 15: 309-315.
10. Gao S, Xu S, Fang Y, Fang J (2013) Prediction of core cancer genes using multi-task classification framework. J Theor Biol 317: 62-70.
11. Azar AT, El-Said SA (2013) Probabilistic neural network for breast cancer classification. Neural Comput & Applic 23: 1737-1751.
12. Elyasigomari V, Mirjafari MS, Screen HR, Shaheed MH (2015) Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization. Appl Soft Comput 35: 43-51.
13. Maulik U, Mukhopadhyay A, Chakraborty D (2013) Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM. IEEE transactions on biomedical engineering 60: 1111-1117.
14. Maulik U, Chakraborty D (2014) Fuzzy preference-based feature selection and semisupervised SVM for cancer classification. IEEE transactions on nanobioscience 13: 152-160.
15. Nayyeri M, Noghabi HS (2016) Cancer classification by correntropy-based sparse compact incremental learning machine. Gene Reports 3: 31-38.
16. Nguyen T, Khosravi A, Creighton D, Nahavandi S (2015) Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification. PloS one 10: e0120364.
17. Zhang X, Guan N, Jia Z, Qiu X, Luo Z (2015) Semi-supervised projective non-negative matrix factorization for cancer classification. PloS one 10: e0138814.
18. Xi M, Sun J, Liu L, Fan F, Wu X (2016) Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine. Comput Math Methods Med Article ID 3572705.
19. Magdy E, Zayed N, Fakhr M (2015) Automatic classification of normal and cancer lung CT images using multiscale AM-FM features. Int J Biomed Imaging Article ID 230830.
20. Kharya S, Agrawal S, Soni S (2014) Naive bayes classifiers: A probabilistic detection model for breast cancer. Int Comp Appl 92: 10.
21. Zwitter M, Soklic M (1988) Breast Cancer Data Set.
22. Wolberg WH (1992) Breast cancer. Wisconsin (Original) Data Set.