International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 15, 2019 (Special Issue)

A Novel Approach for Gene Selection based on Random Forest-Variable Importance

Mrs. K.Uma Maheswari [1], Dr.A.Valarmathi [2]

[1]Assistant Professor, Department of Information Technology, Anna University, BIT Campus

Tiruchirappalli-24. umaravi03@gmail.com

[2] Assistant Professor, Department of Computer Applications, Anna University, BIT Campus

Tiruchirappalli-24. valar1030@yahoo.com

Abstract – Mining microarray gene expression is an imperative subject in bioinformatics in diagnosis of disease. Chronic degenerative diseases such as cardiovascular disease have emerged as the major causes of death. In present era, many diseases are caused by gene transformation. Asian Indians have a higher predisposition to Coronary Artery Disease as compared to any other global population. The aim of this article is to identify the relevant genes which are responsible for cardiovascular disease. The important genes i.e the marker genes which change their expression level in correlation with the risk or progression of the disease. In this article the gene dataset for GSE9820, GSE12288, GSE20681 and the Asian Indians (GSE42148) has been analyzed and the topmost genes which causes the CAD have been identified. Diagnostic tests and classification of patients can be done by using the marker genes, which will reduce laboratory cost and increase the accuracy.

Index Terms – Microarray, Coronary Artery Disease, Marker Genes, Risk, Expression Level, Gene Transformation.

I . INTRODUCTION

Heart disease is the top deadly disease throughout the world. Cardiovascular disease is one of the leading causes of death in human life, and is influenced by both environmental and genetic factors. With the recent advances in microarray tools and technologies there is potential to predict and diagnose heart disease using microarray DNA data from analysis of blood cells. It has also become evident that blood cells can also provide useful genomic information for cardiovascular conditions. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to the Heart diseases. In the age group of 25-69 years, there is 25% of deaths due to heart diseases. Genomics is way to study many genes (thousands of genes or even every gene) in an organism all at once. A microarray is a huge collection of spots that contain massive amounts of compressed data Each spot (one gene) of a microarray contains a unique DNA sequence The DNA Microarray generates gene expression data. A microarray database is a repository which contains microarray gene expression data. Microarrays are a recent technology to determine the expression levels of thousands of genes simultaneously in tissues. Significant information can be extracted from these genes by using machine learning techniques. Clinical microarray data can be analyzed from different viewpoints. The three main perspectives are: (1) making clinical predictions (classification), (2) discovering diagnostic classes (clustering experiments), and (3) selecting relevant genes (or groups of genes) or dimensionality reduction (feature extraction). A microarray gene expression data set can be represented in a tabular form, in which each row represents to one particular gene, each column to a sample, and each entry of the matrix is the expression level of a particular gene in a sample . By analysing the gene expression data, the genes which are responsible for causing diseases is identified.

The objective of this work is to identify the disease causing genes for Coronary Artery Disease. The paper is organized as follows: section II describes related work, section III provides an overview of dataset and methods used. Section IV explains the proposed methodology, Section V represents experimental results and analysis and the conclusion is given in section VI.

II. RELATED WORK

A number of approaches have been used for identifying important genes which are responsible for cardiovascular disease. Some of them is listed below.

Ali Anaissi et al(2013) developed a Balanced Iterative Random Forest algorithm to select the most relevant genes for a disease from gene expression microarray data. The biomarkers were validated by repeated training experiments and showed that the BIRF outperforms the state-of-the- art methods.

Hemant et al(2007) discussed the variable importance and pairwise variable associations in binary regression trees.

Ben Ishak et al(2016) compared SVR(Support Vector Regression) and RF(Random Forest) for the purpose of variable importance assessment. He concluded that the SVR score is recommended for variable ranking in linear situations and RF score is preferable in non-linear cases.

Hapfelmeier et al(2013) developed a new approach which can also be used to control the test-wise and family-wise error rate. He introduced a method that rates the relation between a variable and the outcome and other informative variables in the framework.

Arpita Nagpal et al (2018) developed a new feature algorithm which selects the appropriate selection feature set and overcomes the challenges with microarray data.

Vijendra singh et al(2018) proposed a feature selection algorithm for high dimensional data sets. He compared the proposed algorithm with Fast-Correlation Based Filter(FCBF) , a Fast Clustering-based Feature selection algorithm(FAST) and Random Forest on nine real-world cancer datasets and proved that the proposed algorithm works better in terms of classification accuracy.

Hong Han et al(2016) proposed a new method which is based on Random forest to select the variables using Mean Decrease Accuracy(MDA) and Mean Decrease Gini(MDG) and he demonstrated that the proposed method is powerful in both accuracy and CPU time.

Elnaz Pashaei et al(2016) used RFR(Random Forest Ranking) and BBHA-Binary Black Hole Algorithm for gene selection and showed that by selecting the least number of informative genes increases the prediction accuracy.

Chan Hee Park et al(2014) proposed a new feature selection method which is based on newly designed nearest-neighbor ensemble classifiers and shown that proposed method performs better especially when the number of features exceeds the number of observations.

Newton et al(2001) considered the problem of inferring fold changes in gene expression from cDNA microarray data. The methods are tested via simulation.

Peter et al(2008) performed a review through the workflow of a typical microarray experiment and showed that decisions made at each step from choice of platform through statistical analysis methods are all sources of this variability.

Dengju Yao et al(2015) developed a novel random forest based feature selection method which combines generalised sequence backward searching and generalised sequence forward searching strategies and showed that the proposed method improved the classification accuracy and also reduces the computation time of the feature selection.

Yunsong Qi et al(2011) proposed a method for ranking genes and estimated a threshold above which genes are differentially expressed. The proposed method is more sensitive to data sets with small differentially expressed values.

Zhang et al(2009) provided theoretical insight on the drawback of the double filtering procedure and developed the theoretically most powerful likelihood ratio test statistic.

III METHODS

A. Data Set

The gene expression dataset was taken from the Gene Expression Omnibus. GEO is a public functional genomics data repository. The gene expression dataset was taken from the Gene Expression Omnibus. Gene expression data are usually presented in a matrix form.

GSE12288 dataset has 222 samples and 22,282 genes. Patients who went for coronary angiography are selected based on the Duke CAD index. 110 patients are affected with CAD whose CAD i value is greater than 23 and 112 patients without CAD whose CADi = 0.

GSE9820 dataset has 153 samples and 20,859 genes. 86 patients are affected with severe triple-vessel CAD and 67 patients are not affected with CAD.

GSE20681 dataset has 198 samples and 45,015 genes. Expression profiling of all blood cells from patients are taken before cardiac catheterization. From 198 samples, 99 patients have ≥ 50% stenosis by QCA-Quantitative coronary angiography, 99 patients have luminal stenosis of < 50% by QCA. GSE42148 dataset has 24 samples and 62,972 genes. Out of 24 samples, 13 samples with angiographically confirmed coronary artery disease (CAD) between ages 40 - 55 years and 11 population-based asymptomatic controls with normal ECG and matched for age, gender and common risk factors such as diabetes and hypertension to that of the cases The description of gene dataset is tabulated in table 1.

Table 1. Description of dataset

Datasets	Sample size	Genes Count	No.of Classes
GSE12288	222	22,282	2
GSE9820	153	20,589	2
GSE20681	198	45,015	2
GSE42148	24	62,972	2

B. R Language

R is an open source language. It is used to do manipulation and analysis of various data's in the datasets. Different plots can be made using R and it is used for software development activities in machine learning, data mining and in variety of fields. It is an effective, extensible and abundant environment for various statistical computations and graphics. One of the key features of R language is that it supports user-created R packages and we can import data containing variety of file formats such as CSV (Comma Separated Values), XML(), binary files.

R language has various data structures. It includes vectors, matrices, arrays, data frames (similar to tables in relational database in DBMS) and lists. There are many packages available for R and we can use the package whenever we need. There are various interfaces are available for R language. Among them RStudio is commonly used interface.

C. Hierarchical Clustering

Hierarchical Clustering is the best popular method for gene expression data analysis. In hierarchical clustering, genes with similar expression patterns are assembled together and are connected by a series of branches (clustering tree or *dendrogram*). Experiments with like expression profiles can also be grouped together using the same method.

Steps in Hierarchical clustering

1. Cluster both genes and samples
2. Calculate Euclidean Distance
3. Perform Average Linkage

In order to determine the cluster to cluster distant linkage method is used. Three linkage methods are available.

In single-linkage clustering, the distance between one cluster and another cluster is taken and should be be equal to the shortest distance from any

member of one cluster to any member of the other cluster.

In complete-linkage clustering, the distance between one cluster and another cluster is taken and should be equal to the greatest distance from any member of one cluster to any of the member of the other cluster.

In average-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any of the member of the other cluster.

D. Log Transformation of Data

Gene expression levels are greatly skewed in linear scale: half of the data-point (the lower expressed genes) are between 0 and 1 (with 1 meaning no change), and the other part (the higher expressed genes) between 1 and positive infinity. Consider the case where the normalized expression levels are 0.1 (A), 1 (B) and 10 (C) for 3 samples (A-C) under study. We understand that sample A has a ten-fold lower expression compared to sample B, and that C has a ten-fold higher expression compared to B. However, in ***linear scale*** A and B are much closer (similar) to each other than B and C (0.9 units versus 9 units). A parametric statistical test will be biased and not appreciate that A and C are equally

different from B. Upon **log transformation** (I use base 10 here, but any base will do), the distance

between A and B, and between B and C becomes equal (1 log10 unit, as the log10 values of A, B, and C are -1, 0 and 1).

Log transformation makes the data more symmetrical and so the parametric statistical test provides a more accurate and relevant answer.

E. Random Forest

Random Forest is a method that combines various decision trees to predict or classify the datas. Random forests are used to rank the importance of variables in a regression or classification problem. Mean decreaseGini is a measure of variable importance based on the Gini Impurity Index. It is used for calculations of splits during training.

Gini Impurity Index

Gini Importance is the measure of every time a split of a node is made on variable, the gini impurity criterion for the two descendant nodes is less than the parent node. Adding up the gini decrease for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure.

There are several different methods of evaluating attribute importance for Random Forests model such as MeanDecreaseAccuracy, MeanDecreaseGini. We are using MeanDecreaseGini - a measure of variable importance based on the Gini impurity index used for the calculation of splits during training. Each time a split of a node is made on variable, the Gini impurity criterion for the two offspring nodes is less than the predecessor node. Counting up the Gini decreases for individual variables over total amount of trees in the model gives a quick feature importance that is usually consistent with the permutation significance measure.

The Gini impurity index is defined as

$$G = \sum_{i=1}^{n_c} P_i(1 - P_i) \qquad \textbf{(1)}$$

Where n_c is the number of classes in the target variable and P_i is the ratio of this class. Therefore in binary classification case G is maximized for the sample with equal amount of instances of each class and minimized for the homogeneous sets:

The importance is then calculated using the equation. 2 and it is averaged over all splits within the

$$\textbf{Importance(variable)} = G_{parent} - G_{split1} - G_{split2} \quad \textbf{(2)}$$

forest involving the predictor in question.

Variable importance is the average conditional on the variable used and the meanDecreaseGini of the

group would be the mean of these importances and weighted on the share. This variable is used within the forest compared to the remaining variables in the same cluster.

Measures of variable importance: ranking

An important feature of RF is that it provides a rapidly computable internal measure of variable importance (VIMP) that can be used to rank variables. This feature is especially useful for high-dimensional genomic data. Two commonly evaluated importance measures are node impurity indices (such as the Gini index) and permutation importance. In classification, the Gini index importance is based on the node impurity measure for node splitting. The importance of a variable is given as the Gini index reduction for the variable summed over all nodes for each tree in the forest, normalized by the number of trees.

IV. PROPOSED METHODOLOGY

The main objective of this research is to identify the differentially expressed genes ie. marker genes which is responsible for cardiovascular disease.

The following steps are used in this work for cardiovascular disease prediction.

1. Load the dataset
2. Check the behavior of the data
3. Check the behavior of the data after log-transformation
4. Hierarchical clustering of the "samples" based on the correlation coefficients of the expression values.
5. To select genes we iteratively fit random forests, at each iteration building a new forest after discarding those variables (genes) with the smallest variable importances.
6. Find the variable importance using random forest by measuring the Gini Impurity.

Gini Impurity Index $G = \sum_{i=1}^{n_c} P_i(1-P_i)$

Importance(variable) $= G_{parent} - G_{split1} - G_{split2}$

7. Get subset of expression values for 25 most important genes.
8. Visualization of results.

Procedure for gene selection using Random Forest with Giniindex

1.Start: Ranked Gene Set R=[ ]

Selected subset s= [1…d];

2.Repeat until all features(genes) are ranked

(i) Train t decision trees on subsamples of the original training data, with features in Set S as input variables.

(ii) Compute and normalize the data with log transformation.

(iii) Compute the ranking scores f_i for features in S using G(Gini Impurity Index).

(iv) Find the gene with the largest variable importance.

$a=\text{argmax}_i f_i$ and $b=\text{argmin}_i f_i$

(v) Update R= [a,R], S=S- [a]

3. Output : Ranked gene list

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Variable Importance using Random Forest

GINI importance measures the average gain of purity by splits of a given variable. If the variable is worthwhile, it tends to split mixed labeled nodes into pure single class nodes. Splitting by a permuted variables tend neither to extend nor decrease node purities. Permuting a suitable variable, tend to give relatively large decrease in mean gini-gain. GINI importance is closely related to the local decision function and the random forest uses to select the best available split. Therefore, it doesn't take a lot of time to compute. Gini importance is overall

inferior to variable importance as it is relatively more biased, more unstable and tend to answer a more indirect question.

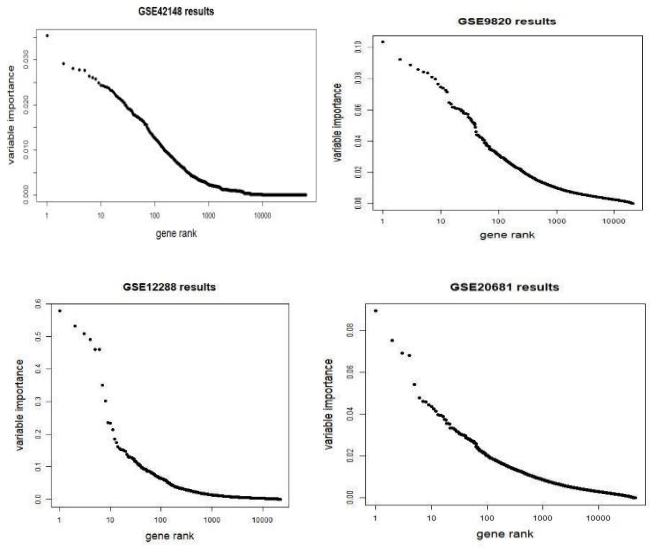

Fig. 1. Variable Importance using Random Forest

The above graph shows the gene ranking based on the variable importance. This variable importance is the measure of gini impurity index. Based on the gini impurity index, the genes are ranked. The ranking is done based on the value of gini index, if the gini index value is small then that gene is given higher rank.

For Asian Dataset (GSE42148) the top 25 genes which are responsible for causing the heart disease is plotted in the figure 2.
The top ranked genes for GSE12288, GSE9820, GSE20681 is shown in the figure 3, figure 4 and figure 5 respectively. The plotting is based on the value of meanDecreaseGini. Mean decreaseGini is a measure of variable importance based on the Gini Impurity Index. It is used for calculations of splits during training. Gini Impurity measures how frequently a randomly chosen record from the data set used to train the model. It will be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset (e.g., if half of the records in a group are "A" and the other half of the records are "B", a record randomly labeled based on the composition of that group has a 50% chance of being labeled incorrectly).

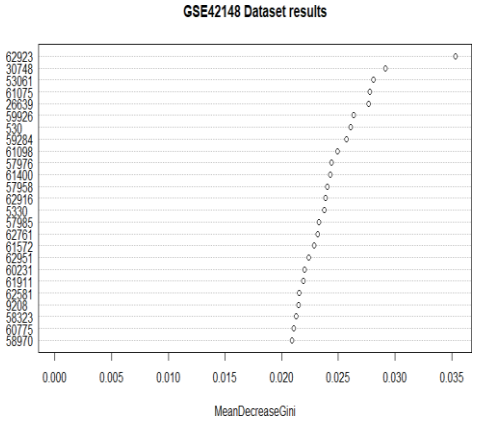

Fig 2. Top Ranked Genes for GSE42148

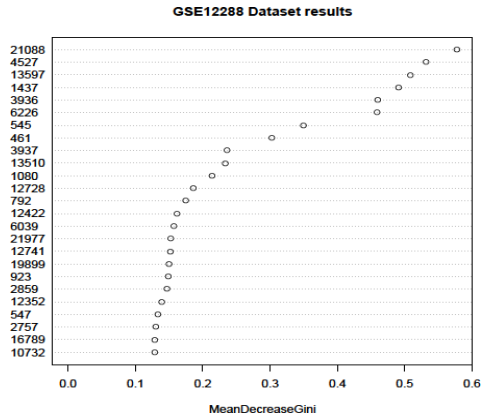

Fig 3. Top Ranked Genes for GSE12288

Gini Impurity reaches zero when all records in a group fall into a single category (i.e., if there is only one possible label in a group, a record will be given that label 100% of the time). This measure is basically the probability of a new record being incorrectly classified at a given node in a Decision Tree which is based on the training data. Gini importance can be leveraged to calculate Mean Decrease in Gini, which is a measure of variable importance for estimating a target variable. Mean Decrease in Gini is the average (mean) of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. This is effectively a measure of how important a variable is for estimating the

value of the target variable across all of the trees that make up the forest. A upper Mean Decrease in Gini shows higher variable importance. Variables are sorted and displayed in the Variable Importance Plot created for the Random Forest by this measure. The most important variables to the model will be highest in the plot and have the largest Mean Decrease in Gini Values, conversely, the least important variable will be lowest in the plot, and have the smallest Mean Decrease in Gini values.

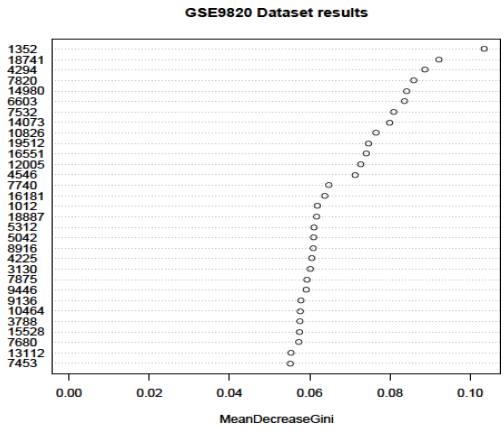

Fig 4. Top Ranked Genes for GSE9820

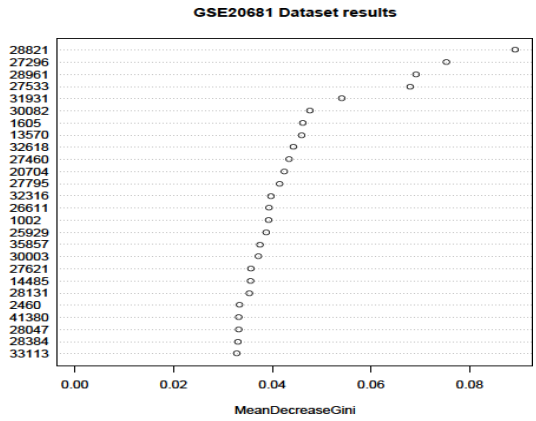

Fig 5. Top Ranked Genes for GSE20681

The plot shows the variable on the y-axis, and their importance on the x-axis. They are ordered from top-to-bottom.The top variables are most important and the bottom one are least-important. Therefore, the most important variables are shown at the top and their importance is assessed by the position of the dot on the x-axis.

The most important variables obtained from the variable importance plot is used for PCA, CDA or other analyses. We should look for a large break between variables, to decide how many important variables to choose. This is an important tool for reducing the number of variables for other data analysis techniques, but you should be careful not to have either too few variables (that won't separate the data) or too many variables (that will over explain the differences).

VI. CONCLUSION

Gene selection is an important part of microarray data analysis due to the large dimension of data. Therefore, the selection of marker genes among thousands to diagnose heart disease is very important. It can help the clinicians to make correct decisions to treat the patients with heart disease. The objective of our work is to identify the top most candidate genes responsible for heart disease. The top most 25 candidate genes are highly expressed in nature. Random Forest is used for finding the important attributes which contribute more towards the diagnosis of heart disease which indirectly reduces the number of tests taken by the patient. It combines the concept of random forest and gini impurity index in order to get the candidate genes. The algorithm proposed in this paper is for gene selection for a particular training set in high dimensional data, such as microarrays .In the future, incorporating pathways into gene selection should help enhance the predictive ability and interpretability of the findings.

References

[1] A.Anaissi, P.J. Kennedy, M. Goyal, and D.R. Catchpoole, "A balanced iterative random forest for gene selection from microarray data", BMC Bioinformatics, vol. 14,no.1, 261,2013

[2] A. Ben Ishak, "Variable selection using support vector regression and random forests: A comparative study", Intelligent Data Analysis, vol. 20,no.1,pp. 83–10,2016.

[3] H. Ishwaran, "Variable importance in binary regression trees and forests", Electronic Journal of Statistics, vol.1 , pp. 519–537,2007.

[4] A. Hapfelmeier, and K. Ulm, "A new variable selection approach using Random Forests. Computational Statistics & Data Analysis", vol.60, pp.50–69,2013.

[5] C. L. Huang, and J. Wang, ," A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with Applications, vol. 31,no.2, pp. 231–240,2006.

[6] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui, "On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data", Journal of Computational Biology, vol. 8no. 1, pp. 37–52, 2001.

[7] C.H. Park, and S.B Kim, "Sequential random k-nearest neighbor feature selection for high-dimensional data", Expert Systems with Applications, vol. 42,no. 5, pp. 2336–2342, 2015.

[8] Y.Qi, H. Sun, Q.Sun, and L.Pan, "Ranking analysis for identifying differentially expressed genes", Genomics, vol.97,no.5, pp.326–329,2011.

[9] P.C Roberts, "Gene expression microarray data analysis demystified", Biotechnology Annual Review, vol. 14, pp. 29–61,2008.

[10] D. Yao, J. Yang, X. Zhan, X. Zhanand and Z.Xie, "A novel random forests-based feature selection method for microarray expression data analysis". International Journal of Data Mining and Bioinformatics, vol.13,no.1, 84,2015.

[11] Arpita Nagpal, Vijendra Singh, "A Feature Selection algorithm based on Qualitative Mutual Information for Cancer Microarray Data",Procedia Computer Science", vol. 132 ,pp. 244–252, 2018.

[12] S. Zhang, and J. Cao, "A close examination of double filtering with fold change and t test in microarray analysis", BMC Bioinformatics, vol. 10. no.1 ,402,2009.

[13] Gerard Biau,"Analysis of a Random Forests Model", Journal of Machine Learning Research,pp. 1063-1095,2012.

[14] H. Pang, S. L.George, , Ken Hui and Tiejun Tong. "Gene Selection Using Iterative Feature Elimination Random Forests for Survival Outcomes", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 5, pp. 1422–1431, 2012.

[15] L. Sun,X. Zhang, J. Xu, W. Wang, and R. Liu, "A Gene selection approach based on the fisher linear discriminant and the neighborhood rough set". Bioengineered, vol. no.1 ,pp. 144–151,2017.

[16] J. Ramos, J. A. Castellanos-Garzón, A. González-Briones, J. F De Paz, and J. M. Corchado, "An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray". Interdisciplinary Sciences: Computational Life Sciences, vol.9,no.1, pp.1–13,2017.

[17] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, , "A review of microarray datasets and applied feature selection methods", Information Sciences, vol.282, pp.111–135,2014.

[18] H. Deng, and G. Runger, "Gene selection with guided regularized random forest". Pattern Recognition, vol.46,no. 12, pp. 3483–3489,2013.

[19] Y. Kong, and T. Yu, "A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification". Scientific Reports, vol.8,no. 1,2018

[20] M.B. Kursa,"Robustness of Random Forest-based gene selection methods", BMC Bioinformatics, vol.15,no.1,2014.

[21] Yunming Ye , Qingyao WuWu, Q., Zhexue Huang, J., Ng, M. K., & Li, X. Stratified sampling for feature subspace selection in random forests for high dimensional data", Pattern Recognition, vol. 46, no.3, pp.769–787,2013.

[22] L. Song, and S. Horvath, "Predicting COPD status with a random generalized linear model", Systems Biomedicine, vol.1,no.4, pp. 261–267,2013.

[23] T. Trinh, D. Wu, S. Salloum, T. Nguyen, and J.Z. Huang, "A frequency-based gene selection method with random forests for gene data analysis", 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF),2016

[24] T.T. Nguyen,J.Z. Huang, and T.T. Nguyen, "Unbiased Feature Selection in Learning Random Forests for High-Dimensional Data", The Scientific World Journal, pp.1–18,2015.

[25] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification", ISPRS Journal of Photogrammetry and Remote Sensing, vol.67, pp.93–104,2012.

[26] M.R. Haque, M.M. Islam, H. Iqbal, M.S. Reza, and M.K. Hasan, " Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder", 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2),2018.