DOI: https://doi.org/10.47391/JPMA.201

Regulatory genes identification within functional genomics experiments for tissue classification into binary classes via machine learning techniques

Bushra Wazir[1], Dost Muhammad Khan[2], Umair Khalil[3], Muhammad Hamraz[4], Naz Gul[5], Zardad Khan[6]

1-5 Department of Statistics, Abdul Wali Khan University, Mardan, Pakistan; **6** Department of Mathematical Sciences, University of Essex, United Kingdom

Correspondence: Zardad Khan **Email:** zkhan@essex.ac.uk

Abstract

Objectives: The aim of this study is to filter out the most informative genes that mainly regulate the target tissue class, increase classification accuracy, reduce the curse of dimensionality, and discard redundant and irrelevant genes.

Methods: This paper presented the idea of gene selection using bagging sub-forest (BSF). The proposed method provided genes importance grounded on the idea specified in the standard random forest algorithm. The new method is compared with three state-of-the art methods, i.e., Wilcoxon, masked painter and proportional overlapped score (POS). These methods were applied on 5 data sets, i.e. Colon, Lymph node breast cancer, Leukemia, Serrated colorectal carcinomas, and Breast Cancer. Comparison was done by selecting top 20 genes by applying the gene selection methods and applying random forest (RF) and support vector machine (SVM) classifiers to assess their predictive performance on the datasets with

selected genes. Classification accuracy, Brier score, and sensitivity have been used as performance measures.

Results: The proposed method gave better results than the other methods using both random forest and SVM classifiers on all the datasets among all the feature selection methods.

Conclusion: The proposed method showed improved performance in terms of classification accuracy, Brier score and sensitivity, and hence, could be used as a novel method for gene selection to classify tissue samples into their correct classes.

Key Words: Gene selection, classification, random forest, cancer, microarray gene expression

Introduction

Cancer is a genetic disease due to changes in some of the genes that control the way how our body cells function (for example, growth and division to make new cells). Therefore, identification of such genes is important for the diagnosis of the disease. High dimensional technologies produce a huge amount of data in many research fields, such as biomedical science[1]. These datasets, such as microarray gene expression data, are known as high dimensional and contain a huge number of irrelevant genes to the corresponding classification problem[1,2]. High dimensional gene expression data pose a number of challenges to the conventional statistical tools, such as logistic regression and chi-square methods, used for their analysis[3]. Analyzing high dimensional data, statistical models lose generalization power and interpretability when applied to unseen data; in that very few genes regulate the target class and rest are redundant (genes with similar expression values) or non-informative[4,5]. Moreover, these analyses also require considerable computational resources[3]. Genes selection techniques are used to overcome these problems with the main theme of selecting a subset of the most informative genes

to be used in model construction and prediction[3,6,7]. Gene selection procedures helps in discovering discreminitive genes that regulate the target class and eliminating irrelevant and useless genes that do not play any role in the regulation of the response class[3]. Gene/feature selection methods are divided into three categories, i.e Wrapper, Filter and Embedded. A brief discussion on these methods is as follows:

Filter methods: These methods identify discreminitive genes by calculating the relevant score for each gene. Genes that have high relevance scores are selected for the purpose of classification by the help of different classifiers. These methods do not require classification algorithem for the selection of important genes. Moreover filter methods deals with big data easily and are computationally fast and simple. Examples of gene selection methods based on filtering approach could be found in various studies[1].

Wrapper methods: In wrapper methods, gene subsets are evaluated by partitioning the gene expression data into training and testing parts and running a predictive model on the training parts corresponding to each gene subset. The predictive model is then assessed by applying it on the testing part for each gene subset and classification accuracy is calculated. Classification accuracy is used as the corresponding score for each gene subset. Gene subset having the highest score is selected as the final gene set to be used in tissue classification[1].

Embedded methods: Embedded methods select informative genes as part of model construction. The model favours those genes that mainly regulate the target class during the training phase of the model. Classification and regression tree is one of the examples of embedded methods[8].

In association with classification problems, gene selection concentrate on the selection of the most informative and regululatory genes. In this connection many gene selection methods have been employed. Among them is random forest (RF)

to solve the two issues of variable selection, i.e., to choose the most important attributes and try to construct the best parsimonious predictive model[9]. Xu et al, suggested an improved RF method, which exploits a new feature weighting tool for selecting a gene subspace and hence boost classification accuracy on microarray data[10]. Vladimir et al, used random forest as a classification and regression techniques for compound classification and quantitative structure-activity relationship (QSAR) modeling where they considered the technique for categorical biological activities[11]. They run models for six data sets and presented three additional features of random forest. They claimed that random forest is one of the most precise and dominant tool of delivering best performance. Diaz-Uriarte et al, used random forest for classification of microarray data and proposed a new method of feature selection based on random forest[12]. Tran et al, proposed the idea of combining bagging and feature selection[13]. In the selection of relevant gene subset for bagging, a wrapper based feature selection method is employed.

Besides bagging and random forest, several other methods have also been employed by various researchers for gene/feature selection. Apiletti et al, proposed a method based on filtering gene selection approach where at first stage they detect outliers in gene expression data for each gene[14]. Several gene selection techniques evaluate the importance of genes in distinguishing the tissue samples in a given target class by deciding a cut point or by fitting a statistical model to microarray gene expression data[15].

This idea was exploited to an expression range to build a gene mask[14]. The idea of set covering approach was used to minimize gene subset and select the most informative genes in the analysis of tumor and normal colon tissues probed by oligonucleotide arrays[16]. The minimum gene subset was selected by replacing the set covering approach with the greedy search approach[14]. To cope with the issue of dimensionality and outliers in genes selection, the greedy search approach was

considered together with proportional overlapping analysis for classifying tissue sample into binary classes[17].

Most of the tree based featue selection methods discussed above use the idea given in random forest to select genes, i.e., a random sample of genes is chosen to select the splitting gene at each node of the tree[9]. In situations where the number of samples is small compared to the number of genes, the random forest idea might not give satisfactory results, as some of the genes might not get a chance to be used as a splitting variable. Therefore, this study used bagged tree forest for feature selection where all the features are assessed to decide on the best possible split. This ensures that every gene plays its role in the construction of the tree model, thus reducing the chance of missing out important genes.

Datasets and Methods:

A total of 5 data sets have been used to compare the methods by calculating classification accuracies, Brier scores and sensitivities. A brief description of each of these datasets is given as follows:

Colon: Colon is a type of cancer which is also known as colorectal cancer. This cancer commonly initiates when strong tissues in the line of rectum change and develop abnormally, resulting in a mass called tumor. The dataset comprised of 2000 genes with 62 colon tissues, out of which 22 tissues were normal and 42 were cancerous. Colon dataset given by Alon et al, and Ben-Dar et al, was utilized for binary tissue classification[16, 18]. Since then, this dataset is progressively utilized and analyzed by many researchers.

Breast Cancer: Breast cancer is a type of cancer occuring in the breast cells. This cancer is common in many parts of the world[19]. This dataset comprised of observations on 4869 genes from 77 tissue samples, of which 33 were noncancerous and 44 cancerous. The data matrix is a 77 × 4869 binary class problem. This data was taken from the study conducted by Michiels et al.[20]

Lymph node breast cancer: This data consisted of 144 lymph node breast cancer patients. Gene expression measurements of 70 genes were signals for metastasis-free survival. A class variable was represented by "event". This data was used by Marc et al, to discover the most reliable means of signals in breast cancer that help in selection of patients for systemic treatment[21].

Leukemia: Leukemia is a group of cancerous blood forming cells. It usually starts in the bone marrow of human body. This cancer occurs due to a mass production consisting of abnormal white blood cells, which fight against infection and toxicities[22]. The leukemia data was published by Golub et al,[23] consisting of observations on 72 tissue samples and 7129 genes.

Serrated colorectal carcinomas: Serrated colorectal carcinomas (CRCs), that are morphologically different from usual CRCs, have been proposed to follow a unique pathway of CRC formation. The dataset has been taken from a study to examine the gene expression profiling of 37 serrated CRCs against conventional CRCs, and to identify differentially expressed genes representing potential biomarkers for serrated CRC[24]. Observations on a total of 22215 genes from the tissue samples have been made.

Table 1 gives a brief summary of the datasets showing the number of tissue samples, number of genes and classwise distribution of tissue samples as noncancerous and cancerous in each dataset.

Gene expression data are commonly given in the form of an expression matrix, $X = [x_{ji}]$, such that $X \in \mathbb{R}^{n \times d}$ and $[x_{ji}]$ is the expression value of gene j for the ith observation where $j = 1, \dots, d$ and $i = 1, \dots, n$. Each tissue sample also has a response class label, y_i, that showed the phenotype of the observation (tissue sample) being observed. Let $Y \in \mathbb{R}^n$ be the set of class labels given that its element, y_i, has a unique value c which is either 1 or 0.

The method exploited the idea of bagging and variable importance using random forest mechanism to select the relevant genes. Bagged classification trees used to ensure that all genes are given the chance to contribute in tree construction. At each node all the "d" features were considered for choosing the best split. On the other hand, trees in random forest were grown on bootstrap samples by considering a random set of $p < d$ features for node splitting. A large number of bagged forests, each consisting of a small number of trees, were grown on bootstrap samples from the training part and the most accurate forests were selected based on out-of-bag error estimates. The selected forests are combined together and used to rank genes in the random forest style. Gene scoring was done as follow:

In each tree of the forest, a binary split was made on the gene that gives the best possible partition of the tissue samples. This was done by computing an impurity measure (Gini index here) of class distribution in the original sample set and the sets formed due to the binary split. Gene that gives the least value of the Gini index is chosen for splitting the tissue samples. This process was iterated recursively on each consequent partition until there remains a single tissue sample in the resulting node. Each time a partition of a node is made on a gene, the Gini impurity measure for the two resulting nodes is less than the parent node. Gini decreases for each individual gene were added over all the trees in the forest. Genes with the highest added decrease were selected as the final set of genes.

Let M be the total number of forests each consisting of B trees. After estimating the out-of-bag errors of all the M forests, top L forests with the smallest errors were selected and combined to form a bagged tree ensemble. The final ensemble thus consist of $T = L \times B$ trees. To calculate the importance of a gene x_i, i.e., $Imp(x_i)$, for predicting Y, weighted impurity decreases $p(t)\Delta i(s_t, t)$ for all nodes t where x_i is used as the split variable, were added and averaged over all the T trees in the final forest, i.e.,

$$Imp(x_i) = \frac{1}{T}\sum_T \sum_{\forall t: v(s_t = x_i)} p(t)\Delta i(s_t, t),$$

where $p(t)$ is the proportion T_t/n of tissues reaching t and $v(s_t)$ is the gene used in split s_t. $\Delta i(s_t, t)$ is the maximum decrease in impurity at node t for the split s_t that divides the n_t node samples into t_L and t_R and is given by

$$\Delta i(s_t, t) = i(t) - p_L i(t_L) - p_R i(t_R),$$

where $i(t)$ is the Gini impurity mesure, $p_L = n_{t_L}/n_t$ and $p_R = n_{t_R}/n_t$.

The proposed BSF method took the following steps for genes selection.

1. Grow a large number M of bagged tree forests each consisting of a small number of trees, say B and rank them with respect to their out-of-bag error.
2. Select a certain number of the top ranked forests, say L.
3. Combine the top ranked L forests to form a final ensemble of L × B bagged trees.
4. Rank genes using the selected forest by calculating the Gini score of each gene.

To evaluate different gene selection approaches, it is necessary to check the accuracy of a classifier applied after the gene selection procedure, where classification is done only on the selected genes. By this evaluation, one could check the ability of genes that regulate the tissue target class. Authors have used different gene selection methods and have shown that gene selection methods have significant effect on a classifier accuracy[25,26]. The same approach has been utilized in another study[14].

Random forest and support vector machine approaches were utilized to assess the predictive power based on the selected genes in comparison with three other state-of-the-art methods, i.e., Wilcoxon rank sum test[1], masked painter and proportional overlapping analysis. A brief description of the classifiers is given as follows:

Support vector machine (SVM)

Support vector machine (SVM) is one of the most commonly used classifiers[27]. The simplest kind of SVM is the linear SVM classifier. In linear classification, a separating plane is said to be the best if it has the maximum margin from both the classes. The margin line is the space between two equivalent hyper planes, each of which goes into the support vectors of one class.

Random Forest (RF)

Random forest is an ensemble learning procedure for classification and regression problems that consisted of many decision trees, where each tree is grown on bootstrap sample from the training data[9]. A new tissue sample is classified on the basis of majority voting from the decision trees in the forest.

Several other classification methods could be considered in addition to random forest and support vector machine to evaluate the performance of the genes selection methods[28,29].

The given datasets were divided into 70% training part (for feature selection and model fitting) and 30% testing (for performance evaluation) part. In the first phase, feature selection methods were applied on the training part. Top 20 genes were selected by all the gene selection methods. In the second phase, the two classifiers were applied on the training part of each dataset with selected set of genes and then the required metrics were calculated on the testing parts. This process was iterated 500 times and final results were the average from all the combination of the runs. For all the analysis, R programming language has been used. For the SVM and RF classifiers, '*kernlab*' and '*randomForest*', R packages were used, respectively. The default values of the paramethers as given in the corresponding packages have been used. For random forest, number of trees have been fixed at 500, node size was set to 1 and the number of genes selected randomly at each node of the tree was the square root of the total number of genes. In the case of SVM, the linear

kernel has been used along with the default automatic selection for the alpha parameter.

Performance Evaluation

To assess the predictive performance of the classifiers based on the selected set of genes identified by the gene selection methods, classification accuracy, sensitivity and Brier score have be used as performance measures. These measures are explained as follows:

Classification accuracy: Classification accuracy was obtained by dividing the number of correct classifications on the total number of tissue samples in the test data. This measure could be easily obtained from a matrix formed by cross tabulating true tissue status vs prediction made by a model. This matrix was known as confusion matrix and is given below.

	True Status	
Prediction	Postive	Negative
Postive	TP (n_{11})	FP (n_{12})
Negative	FN (n_{21})	TN (n_{22})
	$Sensitivity = \frac{n_{11}}{n_{11}+n_{21}}$	$Specificity = \frac{n_{22}}{n_{12}+n_{22}}$

In the above table, TP indicated the number of positive cases classified as positive in the test data; FP was the count of negative cases labelled as positive; FN was the count of positive cases labelled as negative and TN was the count of negative cases labelled as negative. Based on this, classification accuracy was given as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Sensitivity: Sensitivy, also called true positive rate (TPR) is the proportion of positive instances that are accurately classified as positive. It is also called recall and hit rate and is given by

$$Sensitivity = \frac{TP}{TP + FN}$$

Brier Score: Brier score is a scoring criteria that measures the accuracy of probabilistic prediction[30]. A probabilistic prediction indicated an exact outcome/event. This score is commonly used for binary class problems. For true predictions, Brier score would be closed to 0 and close to 1 otherwise.

Brier score could be computed as follows

$$BS = \frac{\sum_{i=1}^{\#\ of\ test\ point}(y_i - \hat{p}_i)^2}{\#\ of\ tissue\ samples\ in\ test\ data},$$

where,

y_i = A particular tissue class value in 0, 1 form;

$\hat{p}_i$ = Estimated class probability

Results

Tables 2-6 gives the results of all the gene selection methods i.e. the proposed BSF method, Wilcoson, proportional overlapping score (POS) method and masked painter (MP), on the 5 datasets via random forest (RF) and support vector machine (SVM). In the given tables, the first column showed the classifiers, the second was the performance measure used against each of the classifiers and the subsequent columns were the values obtained for the measures for each gene selection method. The result of the best performing method is shown in bold.

Table 2 gives the results of the methods for colon data, Table 3 demonstrates the results for breast cancer data and Table 4 gives the results of all the methods for

nki dataset. Table 5 and Table 6 give the results of the gene selection methods on GSE4045 and leukemia datasets, repectively.

Top 20 genes were selected by each method and the rest were discarded. Classification accuracy, sensitivity and Brier score (to know the degree of belief in classification) were calculated on tissue samples in the test set. This process has been repeated 500 times on different random partitions of the datasets. The BSF method gave better result than the other competitors. The proposed BSF method achieved the smallest Brier score values in most of the cases. Further discussion on individual dataset is as follows:

For colon dataset, the random forest and SVM classifiers gave the highest accuracy on genes selected via the proposed BSF method as compared to the Wilcoxon, POS and Masked Painter methods. Accuracy of the proposed method by using random forest classifier was 0.8508, whereas accuracies of the competitors, i.e., Wilcoxon, POS and Masked Painter methods were 0.8495, 0.8391 and 0.7371, respectively. Likewise, random forest attained the highest sensitivity for the proposed method, which is 0.6698. The sensitivity for Wilcoxon, POS and masked painter by using random forest classifier were 0.6479, 0.6345 and 0.5341, respectively. Similar conclusion was drawn from the SVM classifier results. In terms of Brier score, the proposed method has also outperformed the other three methods via both the classifiers.

Likewise Table 3 reflected the highest accuracy and sensitivity for breast cancer of the proposed BSF method via random forest classifier i.e. 0.7931 and 0.8372 respectively, with minimum Brier score of 0.1453. Moreover on SVM classifier the proposed BSF method outperformed all the competitors in terms of Brier score and accuracy, while, Wilcoxon excelled other methods, with regard to sensitivity.

For lymph node breast cancer dataset the proposed method has achieved the best results in terms of classification accuracy, sensitivity and Brier score among all the other methods, on both the classifiers, as can be seen in Table 4.

The results on the cerated colerectral carcinoma dataset were tabulated in Table 5. Substantial difference could be noted in terms of accuracy, sensitivity and Brier score of the proposed method via both the classifiers. The Accuracy and sensitivity of the proposed method via random forest classifier are 0.9177 and 0.6842, respectively, which is much higher than the other methods. The Accuracy of Wilcoxon, POS and masked painter were 0.7847,0.8025 and 0.7331, respectively, via random forest classifier. On random forest classifier the Brier score of the proposed method was less than all the other methods. The SVM classifier too yielded the highest accuracy, sensitivity and the smallest Brier score for the proposed method.

Likewise, it could be noted from Table 6, that the accuracy, sensitivity, and Brier score of the proposed method via the random forest and SVM classifiers were comparatively better than all the other methods on the leukemia dataset. Using random forest classifier the accuracy and sensitivity of the proposed BSF method was 0.9214 and 0..8753, surpassing all the other methods. The accuracy of the competitors, i.e., Wilcoxon, POS and masked painter via random forest is 0.829, 0.88871 and 0.7536, respectively. The proposed method has the smallest Brier score as compared to all the other methods via random forest classifier. Similarly SVM classifier has given the highest accuracy, sensitivity and the smallest Brier score for the proposed method, which is 0.7552 , 0.7661 and 0.0332, respectively.

Discussion

This work has proposed a gene selection method using gene ranking via Gini score method in conjunction with bagged tree ensemble. The proposed method is based

on a greedy search approach that selects the best bagged tree forests from a large pool of forests based on their out-of-bag error. The proposed method achieves improved gene selection in two folds, i.e. selecting the best tree forests while discarding those that do not perform well in the training phase, and considering all the genes for finding the best splitting variable at each node of the trees in the forest. On the other hand, the standard random forest[9] considers node splitting on a randomly selected subset of genes/features and might have the likelihood of ignoring important features/genes in the model building process. The proposed method has outperforemed all the methods on almost all the datasets considered in this paper. In addition to classification accuracy and sensitivity, Brier scores[30] has also been used as performance measure to know the degree of belief in classifying observations to their correct target classes based on the selected set of genes. The proposed method has outperformed all the state-of-the-art methods on all the datasets given in this paper in terms of Brier score. This suggests that the proposed method could be used for gene selection to classify observations into their correct target classes with higher degree of belief as compared to the other methods, i.e. POS[17], Wilcoxon rank sum test[1] and MP[14]. Genes selected via the proposed method also give higher values of sensitivey for majority of the datasets considered using random forest[9] and support vector machine[27] classification algorithms. This means that the proposed method could be more effective in avoiding false positives as compared to the rest of the methods. Moreover, as the proposed method selects the best performing sub-forests and discards those with poor performance, this reduces the size of the final ensemble which in turn saves computational costs in terms of storage resources.

Conclusion and Future Work

A novel gene selection method has been proposed in this paper that is based on an ensemble of the most accurate forests chosen from a large pool of small size forests grown by the method of bagging. The method has been compared with other state-of-the-art methods used in literature for gene selection on a total of 5 gene expression datasets. The analyses have revealed that the proposed BSF method has outperformed the other methods in almost all the cases. This means that the proposed method could effectively identify those genes that have the highest discriminative ability to classify a given tissue sample to its correct target class. The main limitation of the method is computational complexity. This could be avoided by using parallel computing.

Disclaimer: None

Conflict of Interest: None

Funding Sources: None

References

1. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507-17.
2. Mahmood MS, Kureshi N, Frossard PM. Gene markers and complex disorders: a review. J Pak Med Assoc. 2004;54:584-9.
3. Khan Z, Naeem M, Khalil U, Khan DM, Aldahmani S, Hamraz M. Feature Selection for Binary Classification Within Functional Genomics Experiments via Interquartile Range and Clustering. IEEE Access. 2019 Jun 12;7:78159-69.
4. Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc. 2010;105:713-26.

5. Mahmood MS, Kureshi N, Frossard PM. Gene markers and complex disorders: a review. J Pak Med Assoc. 2004;54:584-9.
6. Khalid M, Khan S, Ahmad J, Shaheryar M. Multivariate Covariance using Principal Component Analysis for Reconstruction of Bidirected Gene Regulatory Networks. In2017 International Conference on Frontiers of Information Technology (FIT) 2017 Dec 18 (pp. 229-234). IEEE.
7. Khalid M, Khan S, Ahmad J, Shaheryar M. Identification of self-regulatory network motifs in reverse engineering gene regulatory networks using microarray gene expression data. IET Systems Biology. 2018 Dec 13;13(2):55-68.
8. Breiman L. Classification and regression trees. Routledge; 2017.
9. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
10. Xu B, Huang JZ, Williams G, Wang Q, Ye Y. Classifying very high-dimensional data with random forests built from small subspaces. Int J Data Warehous. 2012;8:44-63.
11. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43:1947-58.
12. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;3.
13. Tran CT, Zhang M, Andreae P, Xue B. Bagging and Feature Selection for Classification with Incomplete Data. In Europ Conf Ap Evol Comp 2017; 471-486.
14. Apiletti D, Baralis E, Bruno G, Fiori A. Maskedpainter: feature selection for microarray data analysis. Intel Data Anal. 2012;16:717-37.

15. Marczyk M, Jaksik R, Polanski A, Polanska J. Adaptive filtering of microarray gene expression data based on gaussian mixture decomposition. BMC Bioinformatics. 2013;14:101.
16. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci. 1999;96:6745-50.
17. Mahmoud O, Harrison A, Perperoglou A, Gul A, Khan Z, Metodiev MV, Lausen B. A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. BMC Bioinformatics. 2014;15:274.
18. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. J Comput Biol. 2000;7:559-83.
19. Mujtaba S, Haroon S, Faridi N, Lodhi FR. Correlation of human epidermal growth factor receptor 2 (HER-2/neu) receptor status with hormone receptors Oestrogen Receptor, Progesterone Receptor status and other prognostic markers in breast cancer: an experience at tertiary care hospital in Karachi. J Pak Med Assoc. 2013;63:854-8.
20. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet. 2005;365:488-92.
21. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. Gene Pattern 2.0. Na Genet. 2006;38:500.
22. Khan A, Shafiq I, Shah MH, Khan S, Shahid G, Arabdin M. Chronic myeloid leukaemia presenting as priapism: A case report from Khyber Pakhtunkhwa. J Pak Med Assoc. 2018;68:942-4.

23. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531-7.

24. Laiho P, Kokko A, Vanharanta S, Salovaara R, Sammalkorpi H, Järvinen H, Mecklin JP, Karttunen TJ, Tuppurainen K, Davalos V, Schwartz Jr S. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. Oncogene. 2007;26:312.

25. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics. 2005;6:148.

26. Mahmoud O, Harrison A, Gul A, Khan Z, Metodiev MV, Lausen B. Minimizing redundancy among genes selected based on the overlapping analysis. In Analysis of Large and Complex Data 2016 (pp. 275-285). Springer, Cham.

27. Vapnik V, Golowich SE, Smola AJ. Support vector method for function approximation, regression estimation and signal processing. InAdvances in neural information processing systems 1997 (pp. 281-287).

28. Gul A, Perperoglou A, Khan Z, Mahmoud O, Miftahuddin M, Adler W, Lausen B. Ensemble of a subset of kNN classifiers. Advances in data analysis and classification. 2018 Dec 1;12(4):827-40.

29. Khan Z, Gul A, Perperoglou A, Miftahuddin M, Mahmoud O, Adler W, Lausen B. Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification. 2019 Jun 2:1-20.

30. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78:1-3.

Table 1: Summary of the datasets.

Datasets	Samples	Genes	Class sizes
Lymph node breast cancer	144	77	96/48
Colon	62	2000	40/22
Breast Cancer	77	4948	33/44
Leukaemia	72	7129	47/25
Serrated colorectal carcinomas	37	22215	29/8

Table 2: Performance of the methods on colon dataset.

Classifier		Methods BSF	Wilcoxon	POS	MP
RF	Sensitivity	**0.6698**	0.6479	0.6345	0.5341
	BS	**0.1431**	0.1575	0.1457	0.1511
	Accuracy	**0.8508**	0.8495	0.8391	0.7371
SVM	Sensitivity	0.4211	0.2396	**0.4474**	0.4131
	BS	**0.1691**	0.1924	0.1694	0.1832
	Accuracy	**0.7554**	0.7109	0.7542	0.6632

Table 3: Performance of the methods on breast cancer dataset.

Classifier		Methods BSF	Wilcoxon	POS	MP
RF	Sensitivity	**0.8372**	0.8298	0.7746	0.6351
	BS	**0.1453**	0.1828	0.2069	0.2113
	Accuracy	**0.7931**	0.7482	0.6309	0.6310
SVM	Sensitivity	0.7798	**0.8214**	0.7876	0.6115
	BS	**0.1521**	0.1592	0.2377	0.2641
	Accuracy	**0.7807**	0.7759	0.6414	0.6192

Table 4: Performance of the methods on lymph node breast cancer dataset.

Classifier		BSF	Methods Wilcoxon	POS	MP
RF	Sensitivity	**0.8124**	0.8030	0.7599	0.7991
	BS	**0.0996**	0.1095	0.1149	0.2013
	Accuracy	**0.8754**	0.8630	0.862	0.7539
SVM	Sensitivity	**0.7697**	0.7475	0.6227	0.6422
	BS	**0.1194**	0.1329	0.15671	0.2110
	Accuracy	**0.8536**	0.8515	0.8113	0.7566

Table 5: Performance of the methods on serrated colorectal carcinoma dataset.

Classifier		BSF	Methods Wilcoxon	POS	MP
RF	Sensitivity	**0.6842**	0.2997	0.2043	0.2110
	BS	**0.0720**	0.1603	0.1611	0.1562
	Accuracy	**0.9177**	0.7847	0.8025	0.7331
SVM	Sensitivity	**0.3100**	0.0069	0.0124	0.2741
	BS	**0.0241**	0.1730	0.1799	.0.1526
	Accuracy	**0.8215**	0.7828	0.7848	0.6317

Table 6: Performance of the methods on leukaemia dataset.

Classifier		BSF	Methods Wilcoxon	POS	MP
RF	Sensitivity	**0.8753**	0.7701	0.7831	0.6119
	BS	**0.1103**	0.2110	0.203	0.2142
	Accuracy	**0.9214**	0.8391	0.8871	0.7536
SVM	Sensitivity	**0.7661**	0.7335	0.6552	0.6112
	BS	**0.0332**	0.1130	0.2411	0.1142
	Accuracy	**0.7552**	0.7315	0.7112	0.6351

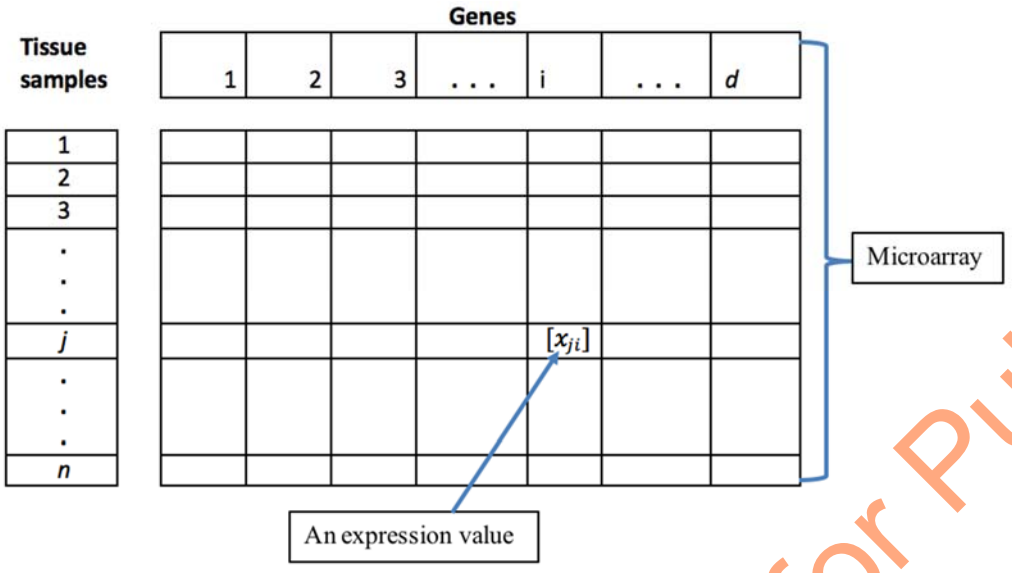

Figure 1: Microarray gene expression dataset.

A flowchart of the method is given in Figure 1.

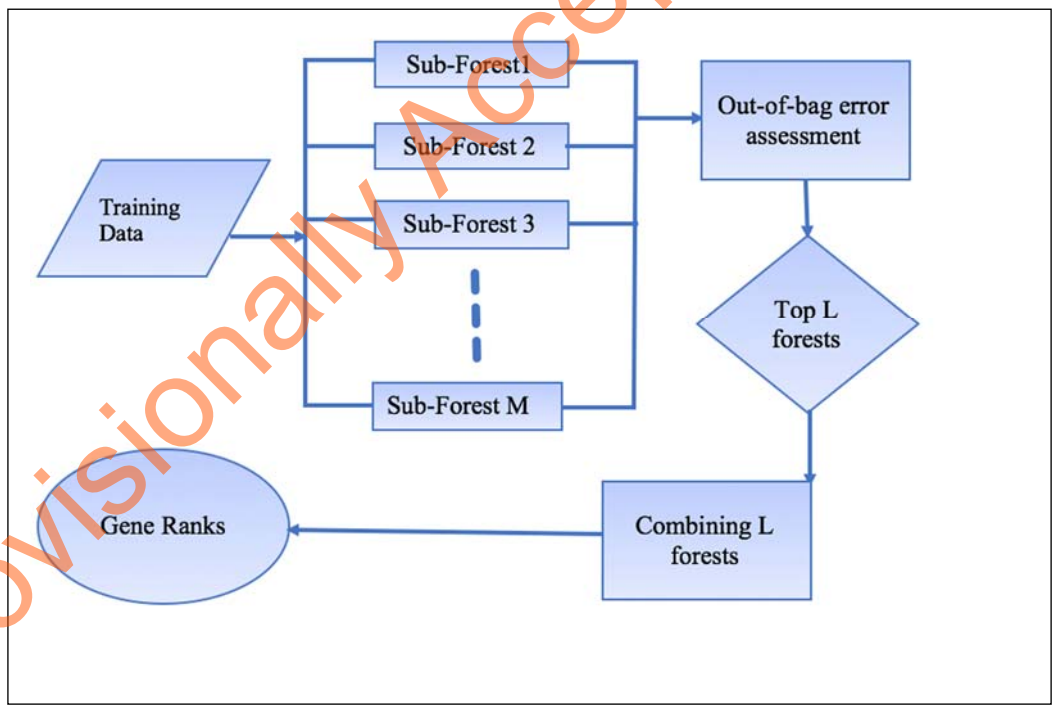

Figure 2: Flow chart of the proposed method.