International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-4, February 2019

# Improving Accuracy For Cancerclassification With Gene Selection

**Geeta Chhabra, VasudhaVashisht, Jayanthi Ranjan**

***Abstract:*** ***The article presents a detail overview of different classification techniques for colon cancer prediction by gene expression dataand evaluated their performance based on classification accuracy, computational time &proficiency to reveal gene information. The gene selection methods have been introduced also and evaluated with respect to their statistical significance to cancer classifier.The purpose is to build a multivariate model for tumour classification with genetic algorithm.The multivariate models were constructed using nearest centroid, k-nearest neighbours, support vector machine, maximum likelihood discriminant functions, neural networks and random forest classifiers combined with genetic algorithm applied to the colon cancer publicly available dataset.It has been observed from the experimental analysis that Maximum Likelihood Discriminant Functions (MLHD) performs better and accuracy has been further been improved by using most frequent genes using the forward selection method. Also, maximum likelihood discriminant functions are cost effective and faster than neural networks (NNET), nearest centroid (Nearcent) and random forest (RF). Thus, the experiments show that classification accuracy is affected with the selection of genes that contributes to the accuracy of the model. It will remove the irrelevant genes thus will reduce the size and make the algorithm fast.***

***Index Terms:*** ***data mining; genetic algorithm; machine learning algorithms.***

## I. INTRODUCTION

The cancer detection and class discovery has recently attracted much attention in the field of medical science. Precise classification of various colon cancer types can help in better understanding of the treatment. For better understanding of the problem, systematic approach depending on gene expression analysis data has been recommended. The genes specific patterns can propose basic problems relating to prevention, cure of disease, drug discovery and biological evolution mechanisms, thus resulted in the classification of gene expression data. The various methods from machine learning and statistics have been used for classification of colon cancer data. Such data are distinct from any related data as: (i) it is very high dimensional and has thousands of genes. (ii) freely available data is either very large or very small (that contains noisy data).

**Manuscript published on 28 February 2019.**
***Correspondence Author(s)
**Geeta Chhabra,** Research Scholar, Amity Institute of Information Technology, Amity University, Noida, Uttar Pradesh, India,
**VasudhaVashisht,** Assistant Professor, Department of Computer Science & Engineering, Amity School of Engineering& Technology, Amity University, Noida, Uttar Pradesh, India,
**Jayanthi Ranjan** Professor, Institute of Management Technology, Ghaziabad, Uttar Pradesh, India,



(iii) most of the genes are not relevant to distinguishcancer class [4].Existing classification techniques cannot effectively tackle this type of data. Some experts recommendgene selection before classification which reduces the data size by removing irrelevant genes and thus improving the running time and classification accuracy. The most critical issue is the prediction of the class of the categorical variable for gene expression data[18]. We have also proposed the gene selection method that is also important in the gene expression data classification. Several other issues are related to cancer classification beside gene selection. These issues are biological vs statistical relevance,the gene contamination and asymmetrical errors related to classification, which are of great concern[10].

The accurate classification of data is difficult due to genes that do not contribute for the cancer classes. This is known as the biological noise.In the gene expression data, the ratio of relevant genes to irrelevant genes is large.Most of the genes are not relevant. The relevant genes are in small number as compared to total number of genes. These irrelevant genes interferethe power of relevant genes andresults in extra computation time for the classifier.Thus, to select the relevant genes, it is necessary to develop a mechanism. Biological relevancy is also an important goal to achieve in cancer classification beside classification accuracy. The biological information disclosed during this process can further help biologists in having better understanding about the genes.The biologists are not only interested in classifiers that have high accuracy butcan also reveal important biological information.

## II. CLASSIFICATION PROBLEM BY GENE EXPRESSION DATA

Analytical Classification problem has extensively been investigated and analysed by experts in the field of databases, machine learning and statistics[13]. In the past, algorithms such as linear discrimination analysis, bayesian network, and decision tree [18] etc have been used but for the last few years classification using gene expression has attracted much attention. Experiments have shown that several types of cancers can be identified by gene expression changes[20]. Most of the proposed classification methods are from machine learning and statistics, ranging from the oldest K nearest neighbour to the support vector machine. There does not exist any classifier which is superior to the other. Some of them works well for binary classification may not work well for multiple classifications, while some are general. Most of the proposed algorithms on gene expression are concerned only with the classification accuracy and does not give much importance to the time taken for computation as most classifiers are expensive computationally.

Classification of gene expression data is different from other data due to its application domain and unique character[4].

Worrawat& Chan(2013) in their article "Apriori gene set-based microarray analysis for disease classification using unlabeled data" haveapplied gene set knowledge, known as transform-approach. They have applied the gene set knowledge to transform the data to another form. They have compared the five sets of gene data. The gene members were ranked by their discriminative power on the basis of most informative gene sets for activity inference to differentiate case.For selection of mostpositive and most negative rank genes, a greedy search algorithm was used.

Guia&Devaraj(2018) in their article "Analysis of Cancer Classification of Gene Expression Data: A Scientometric Review" have discussed the cancer classifier models and evaluated the classifiers using supervised machine learning.

Kourou(2015) in their article "Machine learning applications in cancer prognosis and prediction" presented a review on recent machine learning methodsfor modelling of cancer progression. The main aim is to develop a model that can be used to perform prediction, estimation, classification or similar task.

We have worked on binary classification problem. The data from two groups of patients, which is gene expression data has been selected with tumour and normal class. The dataset contains the expression set for 2000 genes[12]. The aim of the research is to compare the different classifiers and select the best combination.

## III.OVERVIEW OF DIFFERENT CLASSIFIERS

We have compared the accuracy of genetic algorithm with various classification techniques (4,18,19) such as nearest centroid (Nearcent), support vector machine (SVM), k nearest neighbours (kNN), maximum likelihood discriminant functions (MLHD), neural network (NNET) and Random Forest (RF).For this, we have used R package GALGO, having tools for statistical models. For supervised classification, it uses a genetic algorithm search method along with statistical modelling. It has number of statistical modelling techniques for classification. It is variable search method based on principle of evolution by natural selection. From initial random population, it selects the evolving sets of variables that satisfy the certain criteria.Steps followed in GALGO package are:

Step 1: Initially it creates a number of random variables.

Step 2:By training a statistical model each variable is checked for its ability to predict the class, based on the fitness function and gives the score to each variable.

Step 3: The variable with a valuemore than predefined score is selected and iterations stops else it continues to the next step.

Step 4: The population variables gets replicated and variables with higher fitness score generates more offspring.

Step 5: The replicated parents contain the information through crossover and mutations. The crossover and mutations are introduced in the variables randomly. Their cycle is known as generations.

To implement genetic algorithm with various classification techniques using GALGO package we have used the wrapper function which specify the data for classification and stores the results of the analysis. We have configured the wrapper function that will store 300 variables having 5 genes that corresponds to the following models with classification accuracy of 90%.

### *A. K-Nearest Neighbors(kNN)*

It is a supervised classification method where the distinct patterns are grouped together. The nearest neighbour classification searches for the K number of samples close to the pattern for classification [7,11,17]. To determine, closeness between samples, a distance measure is required which is by default Euclidean distance. It is preferred as distance measure because of its generality, which is mathematically defined as;

$$d_{xy} = \sqrt{\sum_{i=1}^{K}(z_{ix} - z_{iy})^2} \quad (1)$$

In equation (1), Krepresents the number of genes in chromosome, x &y are two samples and Z is the total sample of genes. The distance d between the known and unknown sample is calculated and sorted. The first nearest neighbour is with the smallest distance, second nearest neighbour is the second smallest and so on. The unknown sample is assigned with the class where majority of the k nearest samples lies.Predictions of new samples are performed by associating them to the samples analysed with those that are likewise similar. The algorithm has three datasets, one for learning classifier which is a training dataset, another dataset is for validation and testing dataset, whichis used for prediction of unknown samples. The groups can be made using the class distribution between these K nearest neighbours. The training dataset has usually irregular class distribution means that each group can have different numbers of samples. The algorithm calculates the number of groups in data samples. To buildthe classifier, training process is repeated with validation data and accuracy is evaluated.

The parameters needed for wrapper function[11,14] are; classification. method="knn", knn. krepresentsthe number of neighbours required; knn. distance is the distance method to search for neighbours. The values by default for these parameters are 3 and "Euclidean" respectively. It is a non-parametric method. It does not requirethe data to follow a normal distribution but itstill requires the data to standardize before analysis [19].

### *B.Nearest Centroid(Nearcent)*

It is the simplest supervised hybrid classification as it is the combination of an instance based and statistical methods. The method has two datasets namely one for training and another for testing. The algorithm learns classifier patterns from the training dataset and it evaluates the accuracy using testing dataset. The target classes correspond to individual group. It calculates mean or median of the individual group known as centroids. When all the samples are assigned, centroid is recalculated; this process is repeated unless the desired goal is achieved. When the final centroid has been calculated, each unknown sample is assigned the class with minimum Euclidean distance [7,17].

The parameters needed for wrapper function [11,14] are; classification.method="nearcent" for "mean" or "median" as centroids.


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Though it is anon-parametric method, still it is desirable to normalize the data before analysis. Genetic algorithm in R uses nearcent.R.predict and nearcent.C.predict methods for nearest centroid classification [19].

### C.*Maximum Likelihood Discriminant Functions(MLHD)*

To distinguish between sample clusters, it is the most powerful classifier. The means and covariance of multiple variables are considered to distinguish the clusters. The original variables are organised in linear combination to maximize the separation between different clusters.It refers to bay's rule[19] that designate a sample to a group with maximum conditional probability using discriminant function which is based on the means and the pooled covariance for every gene in the chromosome. The genetic algorithm in R works with non-standardized data. The parameter for the wrapper function [11,14] is classification.method="mlhd" [19].

### *D.Support Vector Machines(SVM)*

In this method, each sample is plotted in the convenient plane known as kernel function which is then transformed into higher dimensional space for better and easier separation. The kernel function in SVM can be customized. The best line for separation has maximum distance between the closest samples and the line within that class [3,6, 9,16, 21,23].

The package e1071 in R is used to build the classification tree using svm.R.predict function for genetic algorithm[24]. The parameters required for the wrapper function[11, 14] are; classification method="svm" and svm.kernel which is kernel transformation (by default is "radial") [19].

### *E.Neural Networks(NNET)*

It is a classification method that mimics the learning pattern of natural biological neural networks. In it, neurons in the brain used to communicate through axons and dendrites ends. If the stimulus signals in dendrite ends is greater than a potential action, then neurons cells produces a signal response in the axons ends. Artificial neurons are the mathematical function which produces an output if the sum is greater than threshold which is weighted sum of inputs.Neural network is a combination of several artificial neurons [5, 7, 17].

The nnet package for neural network in R uses the function nnet.R.predict as classifiers for genetics algorithm. The required parameters for wrapper function (11, 14) are; classification.method="nnet" and nnet.size isthe number of units in hidden layer [19].

### *F.Random Forest(RF)*

It is an ensemble classifier for classification problem based on random feature selection and bagging. It gives class as output as it consists of many decision trees. Decision trees are built on leaf nodes and split nodes.Leaf nodes store the information about the sample that can be used for future prediction [7,17].

The genetics algorithm in R uses the package ranforest to build trees using the function ranforest.R.predict.The required parameter for wrapper function[11, 14]is; classification method=" ranforest" [19].

## IV. EXPERIMENTAL ANALYSIS

We used the dataset maintained by Merk S (2018). The 62 samples from colon cancer patients representing 2 different diseases subclasses have been processed. The data consists of 40 samples with tumour and 22 samples which are normal from colon-cancer patients. The article summarizes the results for a binary problem. The data from two groups of patients, which is gene expression data has been selected, 40 samples with t class and 22 samples with n class. The dataset contains the expression set for 2000 genes.The colon cancer dataset is in matrix form which has columns as samples and rows as genes. For the prediction of gene expression data, genetic algorithm combined with machine learning classifiers have been used[3].

### *A. Software Used*

The R package, GALGO is user friendly and is quite useful to develop statistical models for large dataset. It includes methods for supervised classification without any coding for its usage which makes it easy to use for biologists. The object-oriented nature of GALGO in the R environment makes it an ideal framework for any model that uses genetic algorithms as search strategy combined with statistical analysis [1, 19].

The GALGO package has wrapper function, by default;it saves all the variables even if they are not getting the fitness goal. The success of the algorithm is assessed by the fitness value across the generations. The evolution of fitness across generation and time for all the models have been summarized in Table 1.

**Table 1: Evolution of fitness across generation and time**

| Model | Generations | Time Taken in secs |
|---|---|---|
| k-nearest neighbours (KNN) | 25 | 2346 |
| Nearest Centroid (NEARCENT) | 139 | 8683 |
| Maximum Likelihood Discriminant Functions (MLHD) | 35 | 3517 |
| Support Vector Machines (SVM) | 11 | 3371 |
| Neural Networks (NNET) | 57 | 5189 |
| Random Forest | 87 | 21045 |

Table 1 presents evolution of fitness parameters for algorithm predication across generations and time. Parameters 'generation' here refers to genetic algorithm. It shows that Maximum Likelihood Discriminant Functions (MLHD) is faster than nearcent and it is computationally more expensive as it is taking less time. Figure 1 presents the evolution of maximum fitness across generations in MLHD model. It shows that in average, fitness is achieved in 35 generations in case of Maximum Likelihood Discriminant Functions (MLHD) classifier. The lines indicate that the average fitness for all the variables & for those variables that have not achieved the goal.


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

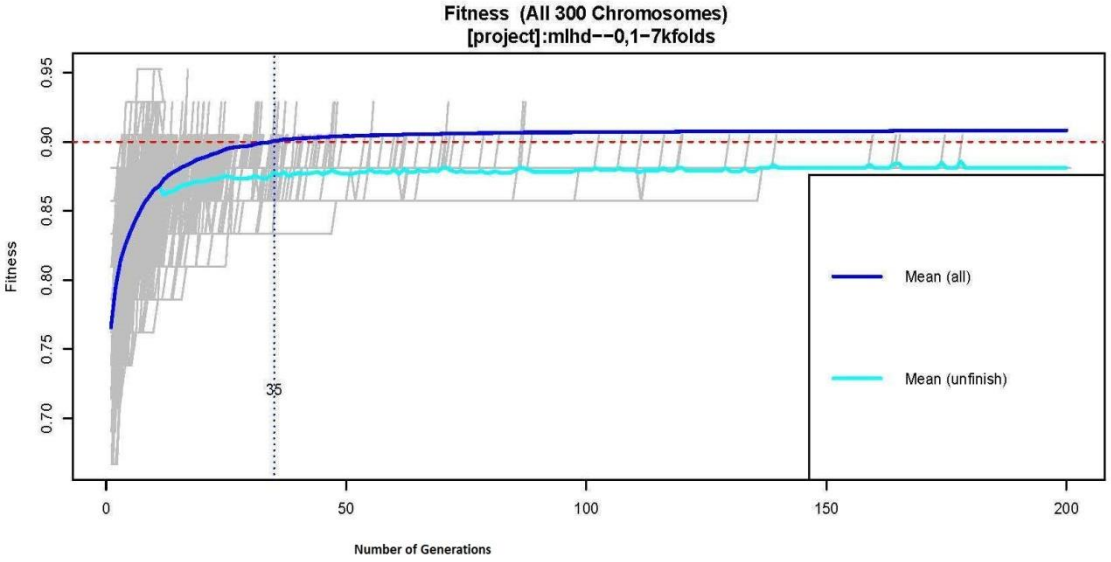


**Figure1: Evolution of maximum fitness across generations in MLHD model**

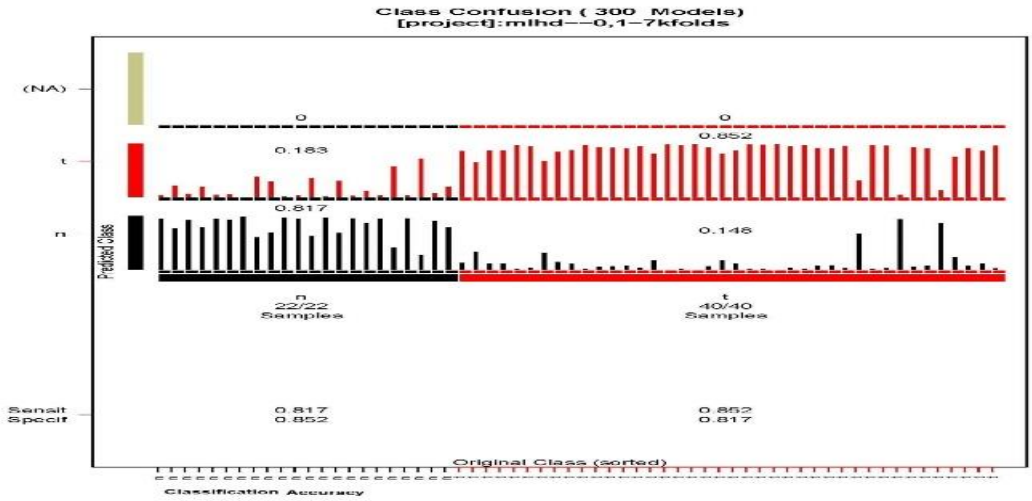


**Figure 2: Overall classification Accuracy By MLHD**

Figure 2 graphically depicts the accuracy of Maximum Likelihood Discriminant Functions. The individual samples are along X axis grouped according to disease t or n. The predicted class is along the Y axis. Bar charts give the percentage of models to classify each sample. The samples in first column marked in black belong to n with 81.7 % accuracy and on average wrongly classified as t with 18.3 %. Similarly, the samples in the second column are classified as t with 85.2% accuracy and wrongly classified as n with 14.8 %.

**Table 2: Overall classification Accuracy**

| **Model** | **Accuracy** |
|---|---|
| k-nearest neighbours (KNN) | 0.7855 |
| Nearest Centroid (NEARCENT) | 0.8160 |
| Maximum Likelihood Discriminant Functions (MLHD) | **0.8345** |
| Support Vector Machines (SVM) | 0.7790 |
| Neural Networks (NNET) | 0.7940 |
| Random Forest | 0.7910 |

Accuracy is one metric for evaluating classification models. It is the ratio of number of correct predictions to the total. [15].Accuracy for all the classifier has been calculated and summarised in Table2. From Table 2, it is clear that Maximum Likelihood Discriminant Functions is most accurate model for classification followed by nearest centroid. Since the chromosome size is fixed in the wrapper function initially, which means some genes does not add to the accuracy of colon cancer prediction. It needs to be identified and should be removed from the chromosomes using backward selection method. The procedure is to eliminate the gene from chromosome and accuracy of the shorted chromosome is calculated, if the accuracy is not decreased, another gene is eliminated and again accuracy is calculated. If the accuracy is decreased, the gene is replaced. Another series of genes are tested for their accuracy until all the selected genes in the chromosome have been checked for their accuracy.


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

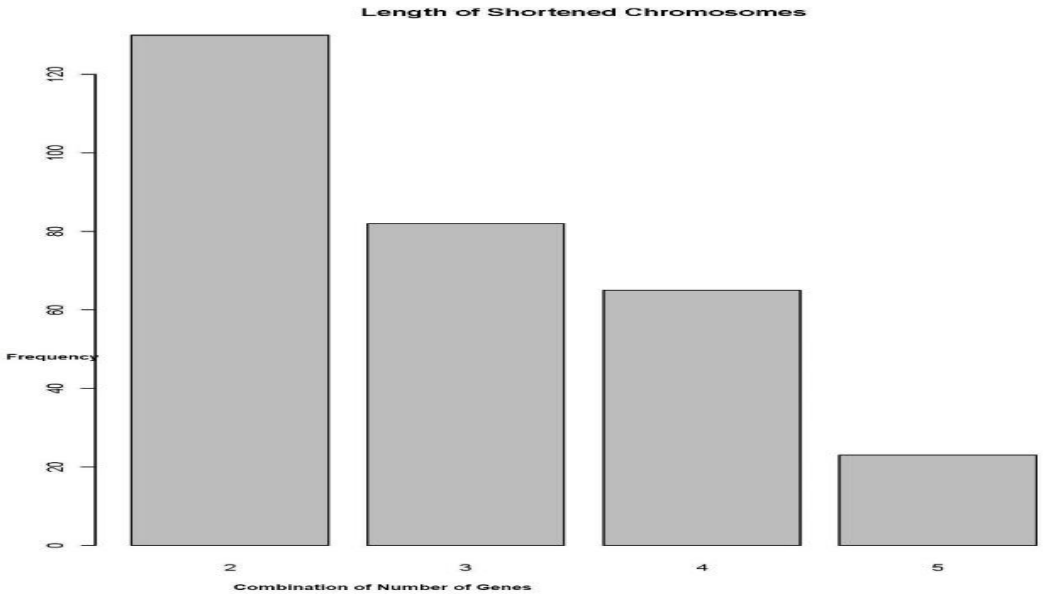


**Figure3: Chromosome refinement using backward selection method**

Here we are trying to solve the two-class problem,tumour or normal and from the Figure 3, it is clear that the majority of the models actually require 2 genes to accurately classifythe samples as the chromosome refinement using backward selection method is the highest for two genes.The genetic algorithm provides an enormous collection of chromosomes, all of which are good solutions of the problem. But one needs to build a representative model based on the clinical importance or for biological interpretation which uses most frequent genes using the forward selection method in the chromosome as the basis of inclusion. The model with the lesser number of genes and the highest accuracy is the representative model.

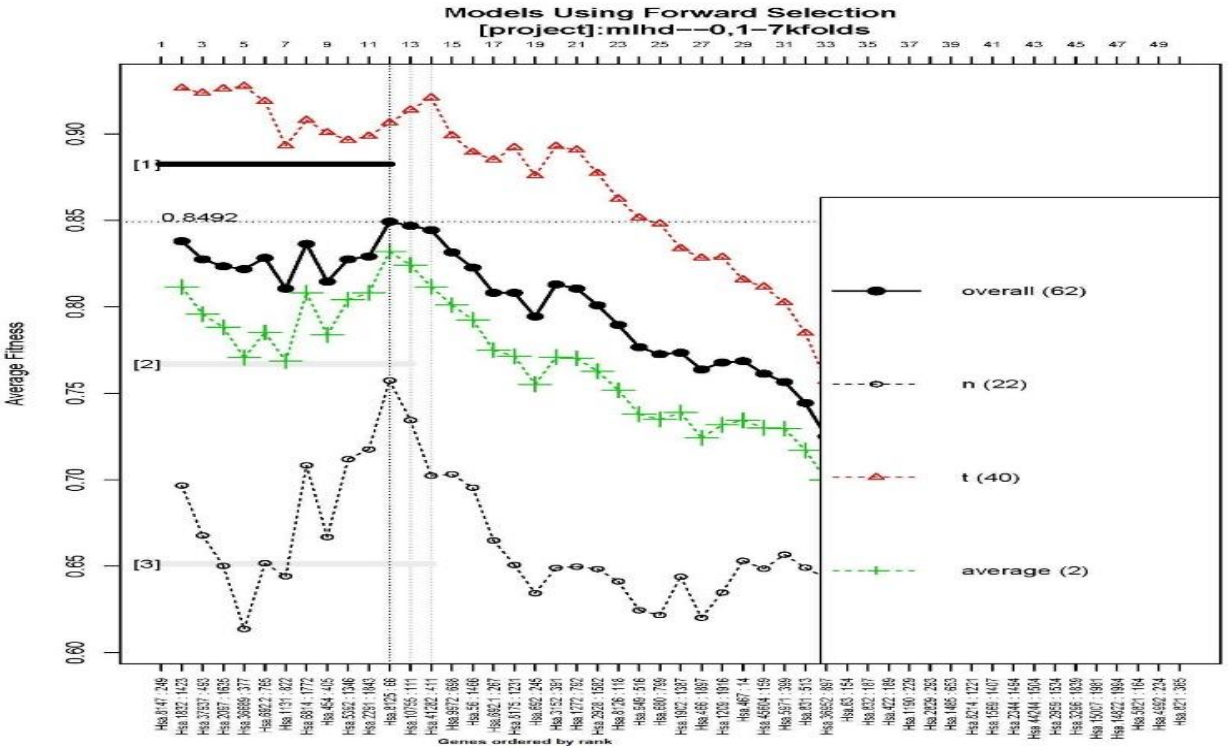


**Figure 4: Forward Selection using most frequent genes in Maximum Likelihood Discriminant.Y axis represents the classification accuracy. X axis have the genes ordered by their rank.Solid line depicts the overall accuracy.Accuracy per class is the coloured dashed lines. Model 1 which is resulted from the selection whose fitness value is maximum is shown in black thick line.**


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The Figure4. shows the Maximum Likelihood Discriminant Functions using forward selection method with the accuracy of about 85%. With a slight increase of 1% which was 84% with Maximum Likelihood Discriminant Functions alone. In this case, the forward selection method has 3 best models shown along the vertical axis. The most accurate model 1 as shown by the dark thick line is having 12 most frequent genes. The heat map plot to visualize the best model is shown at Figure 5 with 12 most frequent genes.

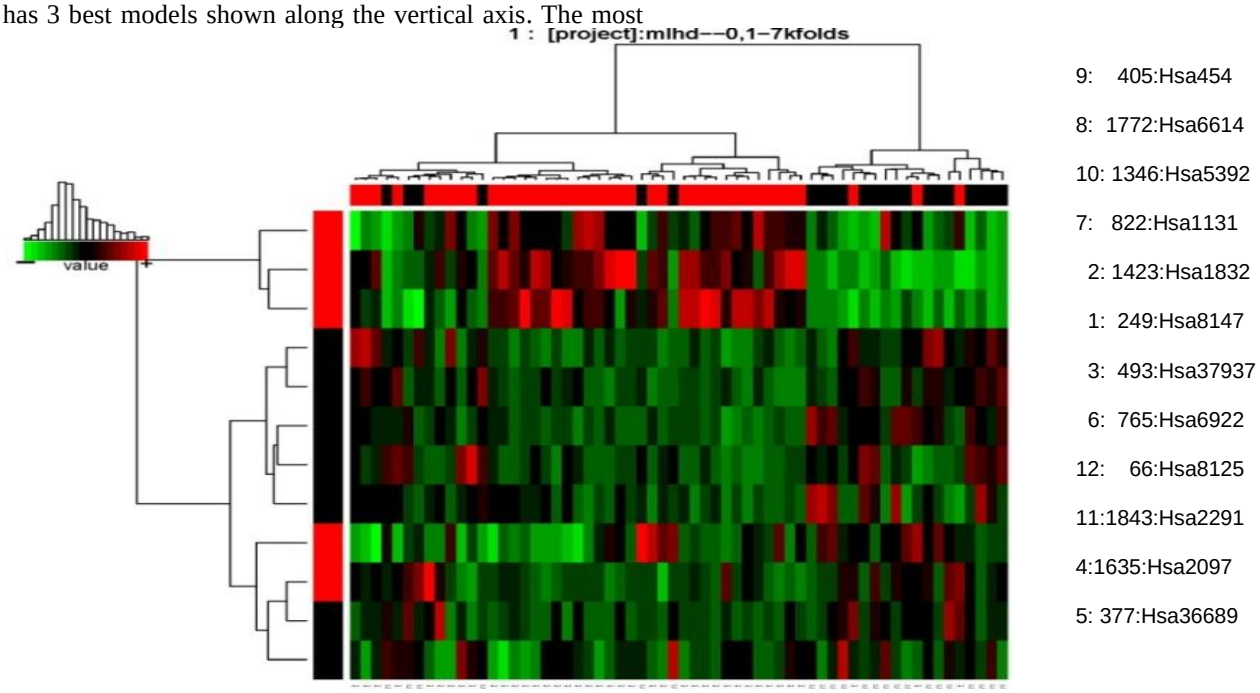


**Figure 5: Heatmap from Maximum Likelihood Discriminant Functions model resulted from forward selection method**

The unique feature of the model is to predict class membership of unknown samples. A new dataset of 15 samples has been temporally appended to the original dataset. The result is shown in the Figure6, in which, the data to be predicted is labelled as "unknown". The black lines show that the 51 % of them has been predicted as normal with 81.9% accuracy and red lines shows that 49% are predicted with the tumour with 84.8 % accuracy.

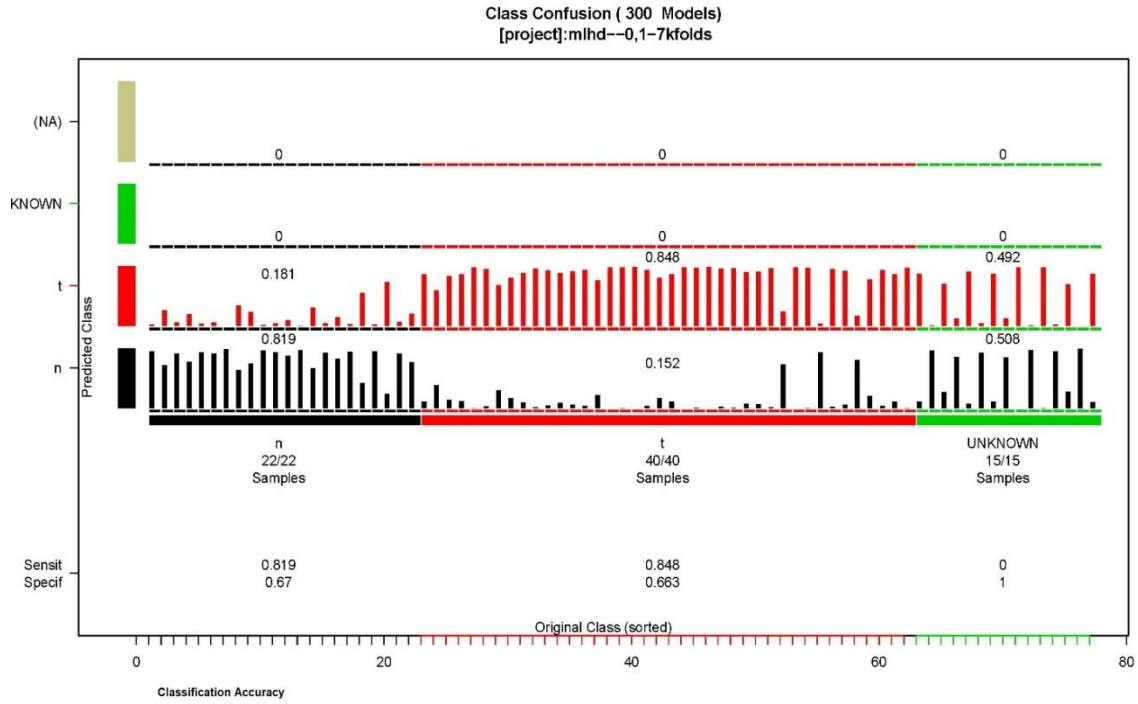


**Figure 6: Prediction for unknown samples**


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

## V. CONCLUSION & FUTURE WORK

The wrapperfunction approach used search for the optimal feature set. It maximizes the classification performance in terms of an evaluation function. There is certain limitation of this approach. One of the limitations is that search for optimal feature subsets needs to be conducted separately for each classification algorithm.The feature subset selected,works well for one algorithm does not work well for the other algorithm. With the small sample size, there may be an over fitting problem due to the estimation of the evaluation function in feature subset selection. In wrapper function methods, itneeds effective searching through all possible combinations of gene sets.

We have used Genetic algorithm usingk-nearest neighbours (kNN), nearest centroid (Nearcent), neural networks (NNET), random forest (RF) and support vector machines (SVM),maximum likelihood discriminant functions (MLHD).

It has been observed from the experimental analysis that Maximum Likelihood DiscriminantFunctions (MLHD) has the highest accuracy of 84 % and which have been further improved to 85 % by using most frequent genes using the forward selection method.

Also, in maximum likelihood discriminant functions, fitness is achieved in 35 generations and it is faster than neural networks (NNET), nearest centroid (Nearcent) and random forest (RF). Thus, the experimentshows that classification accuracy is affectedwith the selection of geneswhich contribute to the accuracy of model.It will remove the irrelevant genes and will reduce the size & makes the algorithm fast. From mean sensitivity and specificity, we have concluded that the selected model i.e.model 1 with 12 genes is far more accurate than any original evolved model. Thus, we can conclude that,by selecting appropriate genes not only we can refine the accuracy of the model to predict the unknown data but also make it computationally inexpensive.

The cancer classification is classification accuracy and also toexplore the biological relevance information for gene expression data. The methods like neural network makes the classification on the basis of the distribution of data values instead of the context meaning of data. The methods like SVM and KNN make use of the correlation between the gene expression values and do not focus on structure of the data. The decision is based on the process of selection and splitting, provides some insight into the data structure.It provides the correlation information among the genes. There are chances when the decision-tree method fails to give good accuracy due to over fitting and noise. Most of the classifiers lacks the biological relevance aspect in case of cancer classification. In order to achieve the classification goalswith accuracy and bio relevancy, it is necessary to develop new classification algorithms or to modify the existing.

Further research is required based on more public datasets with patients who have been diagnosed with the disease. Other different classifiers can also be used. Moreover, this system can also be applied to multiclass datasets. For the last few year researchers have started exploring cancer classification using gene expression data.The results show that gene expression changes are related to types of cancer. Recent experiment shows that no single classifier is better than other(2). Few methods that work well on binary-class problems are not extensible to multi-class problems while others are more flexible and general. The exploitation of such data by the researchers would facilitate studies resulting in more valid results and integrated clinical decision making.

## REFERENCES

1. AdamsL. J., BelloG. A., Dumancas G.Development and Application of a Genetic Algorithm for Variable Optimization and Predictive Modeling of Five-Year Mortality Using Questionnaire Data.Bioinformatics and Biology Insights.2015;3(3):31-41.
2. Amancio D.R., Comin C.H., Casanova D., Travieso G., Bruno O.M., Rodrigues, A.F., Costa L. F. A Systematic Comparison of Supervised Classifiers. PLoS ONE. 2014; 9(4): e94137. Available from Doi:10.1371/journal.pone.0094137
3. Bennet J., Ganaprakasam C.,Kumar N. A. Hybrid Approach for Gene Selection and Classification using Support Vector Machine. The International Arab Journal of Information Technology. 2015;12(6A):695-700.
4. Bhola A., Tiwari A. K. Machine Learning Based Approaches for Cancer Classification Using Gene Expression Data. Machine Learning and Applications:An International Journal.2015;2(3/4). Available from DOI:10.5121/mlaij.2015.2401.
5. Chen H., Zhao H., Shen J., Zhou R., Zhou Q. Supervised Machine Learning Model for High Dimensional Gene Data in Colon Cancer Detection. IEEE International Congress on Big Data.2015;134-141.
6. Dagliyan O.,Uney-YuksektepeF., Kavakli IH, Turkay M. Optimization Based Tumor Classification from Microarray Gene Expression Data. PLoS ONE. 2011; 6(2). Available from https://doi.org/10.1371/journal.pone.0014579.
7. Galván-TejadaC., Zanella-Calzada L., Galván-Tejada J., Celaya-Padilla J.M., Gamboa-Rosales H., Garza-Veloz I., Martinez-Fierro M.L. Multivariate Feature Selection of Image Descriptors Data for Breast Cancer with Computer-Assisted Diagnosis. Diagnostics. 2017;*7*(1)**:**9.Available from https://doi.org/10.3390/diagnostics7010009
8. Guia J. M. De, Devaraj M. Analysis of Cancer Classification of Gene Expression Data: A Scientometric Review. International Journal of Pure and Applied Mathematics. 2018; 119(12):12505-12513.
9. Kourou K., Exarchos T. P., Exarchos K. P., Karamouzis M. V., Fotiadis D. I. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal.2014; 13:8-17.
10. Lu Y., Han J., Cancer classification using gene expression data.Information Systems. 2003; 28: 243–268.
11. Maher B. A., Mahmoud A. M., El-Horbaty El-S., SalemM. Abdel-B. Classification of Two Types of Cancer Based on Microarray Data. Egyptian Computer Science Journal. 2014; 38(2):56-66.
12. Merk S.colonCA: exprSet for Alon et al. (1999) colon cancer data. R package version 1.22.0. 2018.
13. Moorthy K., Mohamad M. S., Deris S. A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data.Current Bioinformatics.2014;9:18-22.
14. Mashhour M. E.,Houby E.M.F, Wassif T. K.,Salah A.I. Survey on different Methods for Classifying Gene Expression using Microarray Approach. International Journal of Computer Applications.2016; 150(1):12-21.
15. Novakovic J. Dj., Veljovic A., Ilic S.S., Papic Z., Tomovic M. Evaluation of Classification Models in Machine Learning. Theory and Applications of Mathematics & Computer Science. 2017; 7(1):39 – 46.
16. Reena S. G., Rajeswari P. A Survey of Human Cancer Classification using Micro Array Data. International Journal of Computer Technology and Applications. 2011; 2 (5):1523-1533.


*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*

17. Siang T. C., Soon T.W., Kasim S., Mohamad M. S., Howe C. W., Deris S.,Zakaria Z., Shah A.Z., Ibrahim Z. A review of cancer classification software for gene expression data. International Journal of Bio-Science and Bio-Technology.2015;7(4):89-108.
18. Tarek S., Elwahab R. A., Shoman M. Gene expression-based cancer classification. Egyptian Informatics Journal.2016; 18:151-159.
19. Trevino V., Falciani F.GALGO:An R package for Genetic Algorithm Searches. Bioinformatics.2006.
20. Torrente A., Lukk M., Xue V., Parkinson H., RungJ., Brazma A. Identification of Cancer Related Genes Using a Comprehensive Map of Human Gene Expression. PLOS ONE. 2016;11(6):1-20. Available from DOI:10.1371/journal.pone.0157484.
21. Venkatesan E.V., Velmurugan T.Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Indian Journal of Science and Technology.2015;8(29). Available from DOI: 10.17485/ijst/2015/v8i29/84646.
22. Worrawat E., Chan J.H. Apriori gene set-based microarray analysis for disease classification using unlabeled data. Procedia Computer Science. 2013; 23:137-145.
23. Zhang H., Wang H., Dai Z., Chen M.,Chen M.S., Yuan Z. Improving accuracy for cancer classification with a new algorithm for genes selection. BMC Bioinformatics.2012;13:298. Available from https://doi.org/10.1186/1471-2105-13-298.
24. Meyer D., DimitriadouE., Hornik K., WeingesselA., LeischF. e1071: Misc Functions of the Department of Statistics, Probability TheoryGroup. 2017. Available fromhttps://CRAN.R-project.org/package=e1071.

**AUTHORS PROFILE**

**Ms. Geeta Chhabra**, is Deputy Director in Ministry of Statistics & PI, Govt. Of India and also a Researcher at Amity University, Noida, Uttar Pradesh. She has more than eight years' experience in the field of Data Processing of Surveys. Also, a long experience in training & teaching at various organisation. She has more than 20 years' experience in the field of information technology. She is masters in Statistics & Software Systems. During her service in various organisation she guided many students/participants in various computer related projects. She also analysed, developed and implemented many important projects in Govt. of India & Govt of NCT of Delhi, India.

**Dr.VasudhaVashisht** is Assistant Professor and Researcher at Amity University, Noida, Uttar Pradesh. She has more than 25 Research Publications in national and international Conference & Journals. She has filed two patents for BCI research work and one copyright from Govt. of India. She has also received one award of Industrial Project. She is Microsoft certified Faculty Fellow. Reviewer and Editorial Board Member of one International Journal. Received many Letters of Appreciation for academic/research and other works. She organized and attended many Workshops & Training Programs/Seminars etc. She has more than 10 years of experience in University teaching & Research.

**Dr Jayanthi Ranjan** is a PhD from Jamia Millia Islamia central university, India in data analytics and has more than 24 years of teaching & research experience. She has been ranked 6th in ALL INDIA RESEARCH PRODUCTIVITY IN MIS AREA. Her course Big Data and Business Intelligence is ranked top ten in India. She has the highest google scholar itations in the entire IMT Group. She is HKUST certified BIG DATA Analytics expert. She has published more than 150 papers and guided 10 PhD students. She is a very well-known expert in Indian business schools for international linkages and collaborations.


*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*