Research Paper

A Wrapper feature selection based frequent pattern classification framework for high dimensional microarray datasets

Araja Raja Gopala[a], Dr. M.H.M. Krishna Prasad[b]

[a]Research Scholar, Department of Computer Science, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, INDIA

Email: rajagopal.araja@gmail.com

[b]Professor, Department of Computer Science and Engineering, University College of Engineering Kakinada, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, INDIA

Email: krishnaprasad.mhm@gmail.com

ABSTRACT

Microarray feature ranking and classification is one of the major challenge to scientific and biomedical researchers due to its high dimensional feature space and limited number of samples. Each microarray contains a large numbers of identical DNA molecules which are used to identify a specific gene related disease. Feature transformation, feature ranking and data classification are the essential techniques which are used to classify the high dimensional data with high true positive rate. Feature transformation is the process of normalizing the feature values within the bounded range. Feature transformation helps to improve the feature ranking process in high dimensional feature space. Most of the traditional feature transformation approaches such as log transformation, min-max normalization etc. are independent of data distribution and outliers. Traditional K-means based feature ranking approaches such as information gain, ANOVA, t-test, Signal-to-noise ratio (SNR) are not applicable to wrapper based feature selection approach. Wrapper based feature selection approach is applied on the high dimensional feature space to find and extract the subset of highly correlated features using ensemble classification accuracy. In this paper, a novel filter based wrapper feature selection method and data classification approach are proposed on high dimensional microarray datasets using MapReduce framework. In this model, a kernel based max-min data transformation method is used to normalize the entire microarray data for feature space partitions. Hybrid correlation based clustering method is used to partition the feature space into k-correlated features. Selected k-correlated features are used in proposed wrapper feature selection approach for data classification. Finally, a comparative analysis is performed on the different microarray datasets to study the true positive rate, error rate and runtime of the proposed model and the traditional models.

Keyword: Microarray analysis, Machine learning, Biomedical datasets, Cancer prediction, Feature selection.

How to cite this article: Raja Gopala A, Krishna Prasad MHM. A Wrapper feature selection based frequent pattern classification framework for high dimensional microarray datasets. Int J Drug Deliv Technol. 2026;16(55s): 994-1016. DOI: 10.25258/ijddt.16.55s.99

Source of support: Nil.

Conflict of interest: None

.INTRODUCTION

The microarray technology usually uses the sequence of resources those are constructed by using various genome projects. Various feature selection and classification models are required to find the relations among the huge number of gene sets. There are large numbers of spots on a single microarray. Every individual microarray includes large numbers of identical DNA molecules those are responsible for identification of a particular gene. Traditional models are executed in two phases for gene identification and classification process:- In the initial phase, all mRNA from cells include two numbers of conditions. When the microarray is excited with the help of a laser, every individual spot emits fluorescence. It has the responsibility to measure the exact quantity of sample out of different conditions. Among all applications of microarray technology, gene expression in Cancer disease diagnosis is the most popular one. Unlike classical approaches, microarray technology has the responsibility to detect different patterns of normal as well as abnormal tissues more efficiently and effectively.

Feature Selection in Microarray Data :

Microarray data is generally contain a set of genes and its disease associations. Most of the traditional approaches detect inappropriate and computationally infeasible data. Hence, it is difficult to process all of the genes that are not required during the process of classification. Hence, the overall computational overhead also increases significantly. Unwanted noise is resulted during the process of classification. Hence, it is very much required to select few numbers of genes those usually take part during the classification process. All of the traditional gene selection techniques involve a perfect combination of filter and wrapper schemes. Filtering approaches have the responsibility to rank every individual feature according to their goodness. During the process of ranking, the relationship among every individual gene with respective class label is considered. Univariate scoring metric play a significant role in the above ranking process. The top ranked genes are selected prior to the execution of classification schemes. On the contrary, wrapper schemes require the gene selection approach in order to integrate with a classifier. The prime objective of this technique is to evaluate the classification performance of every individual gene subset. The optimal subset of genes is detected according to the ranking of performance. Traditional filtering schemes are incapable and inefficient to measure the relationship in between different genes. The gene

expression data play significant role during the process of biomedical diagnosis. A microarray instance includes a huge numbers of genes or characteristics. According to the latest research concepts, limited numbers of genes may result high prediction accuracy during the diagnosis process of cancer disease. Large numbers of genes are not relevant to the disease of interest. Therefore, the gene selection procedure is very complicated task during the microarray data processing. Feature selection is considered as the most powerful tool in order to decrease the size of available data.

Microarray Clustering:

Cluster analysis is the most powerful approach in order to explore complicated diseases. Apart from this, it will enhance the process of prognosis. All of the traditional clustering approaches involve the selection process of certain parameters. The total numbers of expected clusters are very much essential in order to operate and quantify the distance among different data vectors. The distance measure has significant importance during the clustering process. The clustering approach has the capability to resolve the issues related to user-defined numbers of clusters. It is also capable enough to search the data in case of any underlying structure.

Clustering can be defined as an unsupervised scheme of data mining. It has the prime objective to group similar objects based on their similarity index. Basically, the clustering analysis contains three major domains, those are:-

Similarity measurement(s)

Clustering technique

Clustering validation

Since last two decades, there have been extensive amount of research works carried out in the field of clustering analysis. Gene is considered as the basic biological information storage unit which is present in each and every living being. Gene expression can be defined as the process through which information from a particular gene takes part during the synthesis of functional gene products. The quality of gene expression clustering scheme is very much important during the refinement of valuable biological data. It is very difficult to store, extract and process huge quantities of genome data. Presently, extreme learning machine is considered as the learning algorithm in single layer feed-forward networks. There are multiple versions of extreme learning machines and these approaches are implemented in different biomedical applications. Because of cost infeasibility and complex nature of microarray data sets, it is very much complicated to predict the disease status. Gene expression is considered as an important process which is responsible for mapping of genes DNA sequence into the corresponding mRNA sequences. It has the responsibility to describe the expression levels of numerous numbers of genes within a particular cell. Cluster analysis is the most powerful approach in order to explore complicated diseases. Apart from this, it will enhance the process of prognosis. All of the traditional clustering approaches involve the selection process of certain parameters. Classification can be defined as the procedure to classify objects of interest to different previously defined categories or classes.

In this contribution of the paper, a novel kernel filtering based classification approach is presented on high dimensional datasets to improve the true positive rate and accuracy using the hadoop framework.

Related works

F. Meng, et.al, developed a new bi-cluster based Bayesian principal component analysis approach in order to evaluate microarray missing value [1]. Bayesian principal component analysis is a popular microarray missing value estimation approach. The overall performance of the above mentioned system is not significant. In this work, a bi-cluster based technique is introduced in order to exploit local structure of the matrix. In the above proposed technique, each and every bi-cluster gene along with experimental conditions is detected. An automatic learning mechanism is introduced in order to get optimal parameters. In future, further research works may be carried out in order to decrease the normalized root-mean-square error remarkably.

M. Yuan et.al,, introduced an intelligent system for the analysis of microarray data using principal components and estimation of distribution algorithms [2]. The theory of microarray permits to measure each and every expression of genes at the same time. Certain conditions are maintained and the numbers of these conditions are total ten. We can mention here that, clustering and bi-clustering are considered as two important tools in order to analyse gene expression data. The above mentioned data are gathered from various microarray experiments. All the genes having similar behaviour are included inside a particular group. In the above process all required biological knowledge are retrieved successfully. A particular gene can play multiple roles, thus non-exclusive grouping mechanisms are essential. Gene shaving approach is implemented in order to detect distinct sets of genes having equivalent expression. On the other hand, genome biology is considered as the most famous clustering approach in which only coherent clusters are identified. In all of these clusters, the value of variance is very high. Therefore, there exist chances of overlapping in between clusters. In this work, they introduced an intelligent system in order to analyse microarray data. The above presented system applies three numbers of non-exclusive techniques in order to carry out the complete process of clustering and bi-clustering. The prime objective of this work is to detect coherent clusters of genes having large between-sample variance.

C. Chuang, developed an advanced feature gene selection approach of adult ALL microarray data along with affinity propagation clustering scheme [3]. Microarray data analysis technique has become more popular and most commonly implemented approach in order to identify diseases. It includes total ten thousands genes for input dimension. It is actually a severe computational issue in the process of data analysis. In this research paper, they introduced an affinity propagation clustering in case of feature gene selection of adult acute Lymphoblastic Leukemias microarray data. Identification of feature genes

completely depends upon the total number of clusters in case of affinity propagation clustering. Affinity propagation clustering can be defined as an advanced clustering scheme that involves message interchange among various data points. Apart from this, it is also responsible for decreasing the dimension of every individual sample. It never requires prior knowledge about the total numbers of clusters. By analysing the outcomes of the above proposed approach, certain genes with AP clustering can provide appropriate learning in classification and prediction.

L. Fan, K. Poh and P. Zhou proposed partition-conditional ICA for Bayesian classification of microarray data [4]. Exact and proper classification of microarray data is a very crucial task in the field of medical decision making. By studying the previous research papers, we can say that, class-conditional independent component analysis (CC-ICA) is efficient enough for enhancing the performance of naïve Bayes classifier in microarray data analysis. In certain cases the microarray dataset contain few numbers of samples for several classes. In the above cases, some applications of CC-ICA appear to be infeasible. This research paper actually extends the traditional CC-ICA approach. The new extended and modified version of CC-ICA is known as partition-conditional independent component analysis (PC-ICA). As compared to the traditional approaches, PC-ICA provides better and most efficient feature extraction strategy. In other words, we can also mention here that, this above presented approach enhances the overall performance of Naïve Bayes classification of microarray data.

M.Sun et.al,, proposed an efficient expert system in order to carry out the classification of microarray gene expression data with the help of decision tree algorithm [5]. Gene selection process has the responsibility to assist during the analysis of microarray gene expression data. It is very complicated task to get appropriate classification outcomes through the implementation of machine learning approaches. The main reasons behind the above mentioned complication are:- issue of dimensionality and issue of over-fitting. Again, we can state that, the dimensions of features are huge but the samples are limited in numbers.

In the above presented research paper, they developed a new technique in order to overcome the above mentioned two issues. Additionally, it is implemented in order to choose a small set of biomarker genes during the process of diagnosis. At last, they used these markers in the classification process of cancer disease. Several numbers of microarray data sets are included in the experimental evaluation phase. We can conclude here that, the above presented approach is both efficient and reliable in nature.

B. Hosseini and K. Kiani introduced a scalable and robust fuzzy weighted clustering technique that is based on MapReduce with application to microarray gene expression [6]. They have named their proposed technique as FWCMR. Data clustering is considered as the most efficient data mining approach in order to detect groups of similar objects available inside a particular dataset. There exist numbers of different issues and challenges, those are mentioned below:-

Scalability in order to manage large volumes ,

Robustness to intrinsic outlier data,

Validity of clustering outcomes, and so on.

To overcome the above mentioned issues, a fuzzy weighted clustering technique is implemented. This approach is parallel and distributed in every individual phase. This technique can be implemented in different data clustering applications. It can be implemented in gene expression clustering in order to obtain the functional relationships of various genes. The above presented approach evaluates the similarity measure by integrating ordered weighted averaging and spearman correlation coefficient techniques. Here In this work, density reachable genes are merged together in order to form sub-clusters. At last, the final cluster outcomes regenerated through the combination of the above mentioned sub-clusters.

After that, an efficient voting mechanism is applied that has the responsibility to identify the best weights. Apart from this, it can also identify valid clusters in between every distinct dataset. The above algorithm is implemented on a distributed processing environment. This system is capable to process every size of data those are saved in different cloud infrastructures. The precision value of clusters is computed with the help of several cluster validity indexes.

S. F. Hussain and Md Ramazan introduced bi-clustering of human cancer microarray data using co-similarity based co-clustering [7]. The bi-clustering of gene expression process has an objective to detect each and every localized pattern within a particular sub-space. A bi-cluster is also known as co-cluster. Co-cluster can be defined as a group of genes those exhibit equivalent expression intensity within a subset of features. Almost all of the bi-clustering approaches have the objective to detect sub-matrices those have some extent of coherence. At first, a sub-matrix is selected and after that, rows and columns of that matrix are added or removed iteratively. The above approaches completely depend upon the initial, complicated selection mechanism of genes and condition clusters. In this research work, they have applied a new technique in order to cluster textual data.

It has an objective to detect bi-clusters within gene expression data. The above presented technique depends upon the theories of co-similarity among genes. It gives rise to weighted higher order routes within a specific bipartite graph. Hence, they constructed various statistical relations among genes and several conditions. The above-mentioned statistical relationships are established through comparing every individual gene and various conditions prior to the final extraction of bi-clusters. The above presented approach is capable enough to detect required non-overlapping bi-clusters in case of both synthetic data and original cancer data. We can mention that, this approach is perfectly resistant to noise within the data. Again, it can extract bi-clusters in case of different medical applications.

N. Iam-On and T. Boongoen introduced a new diversity-driven generation of link-based cluster ensemble approach. Again, they have included the applications of this approach in the classification process [8]. Since last decade, huge amount of research works have been carried out in order to enhance the overall classification accuracy. These classifiers can be implemented in order to resolve various kinds of real world issues just like prior prediction of system failure and microarray dependent cancer disease

diagnosis process. The overall accuracy of the system depends upon uninformative variables in case of latest data. Apart from the feature selection process, the actual data are transmitted to different variations. In the above scenario, just the characteristics are considered. This technique is different from all previously developed classical transformation based approaches in terms of cluster ensembles. The information matrix along with the transform data play vital role during the complete classification process. The link based cluster ensemble approach results more accurate clustering process. Therefore, this research paper includes the basic concepts of link based cluster ensemble approach. This approach is integrated with diversity driven production of ensemble. It produces informative and diverse set of clusterings. Both synthetic and original datasets are included and these data are tested with the help of C4.5, Naive Bayes, kNN, neural network and random forest classification algorithms. By analysing the resulted outcomes, we can say that, this approach can enhance the overall classification accuracy significantly.

C. Lee, W. Lin, Y. Chen and B. Kuo introduced an advanced gene selection and sample classification scheme on microarray data based on adaptive genetic algorithm/k-nearest neighbor method [9]. In the present time, the uses of microarray technology are increasing day by day. Among numbers of different application areas, this technology is mostly implemented in gene expression for cancer disease diagnosis. The most attractive characteristic of microarray technology is that, it is capable of measuring huge number of genes simultaneously. Previously, all researchers applied parametric statistical approaches in order to detect various significant genes. Microarray data never fulfil the constraints of traditional parametric statistical approaches. In other words, we can say that, the Type 1 error is usually increases.

Hence, objective of this research work is to develop an efficient and effective gene selection approach without including assumption restriction. Again, the dimension of dataset is decreased significantly. In this work, adaptive genetic algorithm or k nearest neighbour algorithm is implemented in order to generate required gene subsets. Furthermore, this approach plays significant role during the action reduction process. Each and every test sample is classified accurately and appropriately. By implementing the above-mentioned approach, researchers are capable to detect the relevant genes appropriately from the sub-gene set.

Wahyu et.al, developed a new kernel based nonlinear dimensionality reduction technique for microarray gene expression data analysis [10]. Appropriate recognition of cancers that depends upon microarray gene expressions is very crucial for the healthcare professionals. It helps the healthcare professionals and researchers in order to select the most appropriate treatment strategy. Genomic microarrays are considered as the most effective research tools in the domain of bioinformatics and medical research. A basic microarray evaluation may generate very high dimensional data along with large quantities of information. Because of these huge quantities of data, many researchers tried to extract only the relevant and important features. Apart from this, they also emphasized on reducing the dimensionality problem. This research paper presented a new kernel based technique that depends upon the process of linear embedding. This technique is used to select the optimal numbers of nearest neighbours. Apart from this, it is also useful to build uniform distribution manifold. In this work, an advanced nonlinear dimensionality reduction kernel based approach is presented. Again, support vector machine algorithm is also implemented which has the objective of developing an extended class of learning machines. The above mentioned class has significant role during the classification and recognition of genomic microarray. In this research work, they included two different DNA microarray data sets for the evaluation. We can conclude here that, the proposed approach is no doubt more efficient and effective as compared to all other traditional approaches.

J. Li and F. Wang developed a matrix factorization framework in order to carry out the process of unsupervised gene selection [11]. The latest technology of microarray gene expression made the phenotype classification of different diseases more accurately. During the process of gene expression data analysis, every individual sample is represented through a large number of genes. Among all of these genes, most of them are redundant in nature. Again, some of them are insignificant to the required disease issue. Hence, it is very crucial and complicated task to choose the most relevant genes during the process of gene expression data analysis. In this work, an efficient technique is introduced that has two stage of gene selection. In the initial phase, they implemented the K means technique in order to perform gene clustering.

It has the responsibility to eliminate several redundant genes. In the subsequent phase, they chose the most representative genes out of the rest according to the process of matrix factorization. At last, through the analysis of outcomes the effectiveness of this presented approach can be demonstrated.

B. Liu, C. Yu, D. Z. Wang, C. C. Cheung, and H. Yan designed exploration of geometric bi-clustering for microarray data analysis in data mining [12]. The process of bi-clustering is considered as the most important approach during the process of data mining. It has the responsibility to detect equivalent patterns. Geometric bi-clustering approach is usually applied in order to decrease the complexity of NP complete biclustering approach. This research paper emphasized on three basic and up to date platforms just like multi core CPU, GPU and FPGA. All of these three has the responsibility to improve the traditional GBC approach. By analysing the parallelizing characteristic of mentioned GBC approach, the researchers developed:-

Multithreaded software that can be executed on a server grade multi core CPU system.

CUDA program for GPU in order to improve the GBC approach.

A parameterizable and scalable hardware architecture that can be applied on an FPGA.

They compared the speed and energy efficiency of the above three presented approaches. In future, further

research works can be carried out in order to modify and extend the above presented technique.

Mukhopadhyay and M. Mandal introduced a new technique for identification of non-redundant gene markers from microarray data [13]. Their proposed approach is a multi-objective variable length PSO-based approach. Detecting relevant genes that usually cause different kinds of cancer is an emerging issue. The marker genes are those genes that usually modify their expression level in correlation with the risk or disease progression. The process of gene expression profiling through microarray technology has become popular now a days. Hence, it can be implemented during the classification and diagnosis of cancer diseases. Extraction of marker genes out of large numbers of genes is considered as a severe issue. Almost all of the traditional approaches for detection of marker genes are responsible to detect a set of genes those may be redundant in nature. In order to overcome the above mentioned issues of previously developed traditional approaches, an advanced multi-objective optimisation technique is presented that can detect a small set of non-redundant diseases. This approach is responsible for resulting high sensitivity and specificity at the same time. In this research paper, all the issues of optimisation are identified and an advanced framework is developed. This framework depends upon the concepts of variable length Particle Swarm Optimisation technique. By including several real world data sets, the performance of this approach is analysed. In future, research efforts may be performed to improve the performance of this approach.

T. Nguyen and S. Nahavandi proposed a modified AHP for gene selection and cancer classification using type-2 fuzzy logic [14]. This research work presented an extended version of analytic hierarchy process in order to choose the most informative genes. These genes are included as inputs to type 2 Fuzzy Logic system in order to carry out the process of cancer disease classification. The extended AHP has the responsibility to include quantitative factors those can be included as ranking outcomes of specific gene selection approaches. Some examples of these factors are:- t test, entropy, receiver operating characteristic curve, wilcoxson test and signal to noise ratio.

The above presented technique is implemented to carry out the process of classification in order to manage nonlinear, noisy and outlier data. The above mentioned are some major issues in cancer microarray gene expression profiles. In this work, an advanced unsupervised learning technique is integrated with an efficient fuzzy c-means clustering in order to initialise parameters. Different kinds of classifiers just like multilayer perceptron network, support vector machine and fuzzy ARTMAP are applied for comparative study. The above presented tool is considered as the most powerful tool in order to carry out the process of cancer disease classification. It can also be implemented as an efficient clinical decision support system.

*C. Orsenigo, C. Vercellis*to performed a comparative study of nonlinear manifold learning techniques in order to classify microarray data [15]. This research paper includes an empirical comparison of some of the latest nonlinear manifold learning approaches for the process of dimensionality reduction. This technique can be implemented in case of high dimensional microarray data classification. In this paper they assessed performance of six approaches, those are:- isometric feature mapping, regionally linear embedding, Laplacian eigenmaps, regional tangent space alignment and maximum variance unfolding. They have developed an advanced framework which is responsible for the extension of traditional dimensionality reduction process. This technique can easily eliminate overestimation of classification accuracy and this may cause misleading comparative results. Almost all of the empirical techniques need a very fast and efficient out of sampling scheme in order to map additional high dimensional data points into a reduced space. They have implemented an advanced multi output kernel ridge regression which is actually the modified version of linear ridge regression. This technique depends upon various kernel functions. On the other hand, the above presented approach is integrated with a variant of isometric feature mapping. In future, additional research works may be carried out in order to extend and enhance the above presented model.

S. S. Ray, A. Ganivada, and S. K. Pal emphasized on a granular self-organizing map for clustering and gene selection in microarray data [16]. They have introduced an advanced granular self-organising map through combination of fuzzy rough set along with SOM. During the training phase, the weights of winning neuron and its neighbour neurons are updated with the help of an advanced learning mechanism. The neighbour neurons are again defined with the help of fuzzy rough sets. The clusters which are generated with the help of GSOM are used in a decision table as decision classes. According to the above mentioned decision table, an efficient scheme of gene selection is introduced.

The effectiveness of the above proposed technique can be demonstrated with the help of both clustering samples and designing an unsupervised fuzzy rough feature selection strategy. The mentioned strategy is implemented during the process of gene selection along with microarray data. The outcomes of this approach are compared with other traditional clustering schemes. This technique is better in terms of classification accuracy and feature evaluation index. We can conclude here that, this approach is statistically improved as compared to other classical unsupervised schemes.

Saha, U. Maulik, concentrated on the improvement of new automatic differential fuzzy clustering using SVM classifier for microarray analysis [17]. Presently, the issue of clustering in case of microarray data is considered as the primary concern of various researchers. Almost all of the clustering approaches have an objective to detect effect of genes. In the above case, the number of cluster is also called as a priori. An advanced real coded enhanced fuzzy clustering approach is developed that is responsible for automatic evolution of the total number of clusters and partitioning of gene expression data set.

In order to enhance the outcomes, the above clustering technique is merged with support vector machine algorithm which is considered as the most common approach for supervised learning. The partial gene expression data points are chosen from various clusters according to the proximity. The clustering assignments of rest of the gene expression data points are identified with the help of an advanced trained classifier.

T. Wong and K. Liu introduced a probabilistic algorithm that depends upon clustering analysis and distance measure for subset gene selection [18]. Numbers of different gene selection approaches for microarray data can be implemented as the most popular classification tools in order to evaluate the discernability of a specific gene subset of a particular disease. The above-mentioned evaluation process has relatively high computational complexity. In this work, they presented a probabilistic mechanism that depends upon density based clustering scheme and a distance measure in order to carry out the process of individual and gene replacement. Researchers are allowed to select appropriate values as the parameters of the mentioned probabilistic scheme in order to set the computational complexity. They have included microarray data sets in order to perform the process of evaluation. Further research efforts are required to enhance classification accuracy and performance of the above mentioned approach.

T. Wong and D. Chen developed a new gene selection methodology for microarray data based on risk genes [19]. Numbers of different gene selection approaches have been introduced in order to choose an appropriate subset of genes that can result high prediction accuracy in case of cancer disease classification. Almost all of the genes have equivalent preference levels. Some researchers termed mutated or flawed genes as risk genes. It is considered as the primary cause of a particular disease. This research paper explains an advanced gene selection approach that depends upon the risk genes. The information produced by risk genes can decrease the overall time complexity for the process of gene selection. Additionally, it enhances the accuracy of cancer disease classification. The complete gene selection approach can be divided into two phases. We can mention here that, all risk genes are required to be selected. In the initial phase, all the genes those have equivalent expression levels or functions must be eliminated. In the subsequent phase, the process of gene selection and gene replacement is implemented. In this process, they analysed outcomes which is responsible for the decomposition of the remaining genes into clusters. The above presented technique outperforms all other conventional approaches.

Proposed Model

The overall architecture of the proposed model is represented in fig 1. Initially, each microarray gene disease dataset is processed to find the synonym of the gene feature for efficient gene-symbol to gene-name mapping. Each microarray training dataset is pre-processed using the data transformation function to remove the variation among the data distribution. In the proposed work, a kernel based data transformation function is used to normalize the input training data for clustering and wrapper feature ranking process in the mapper phase.

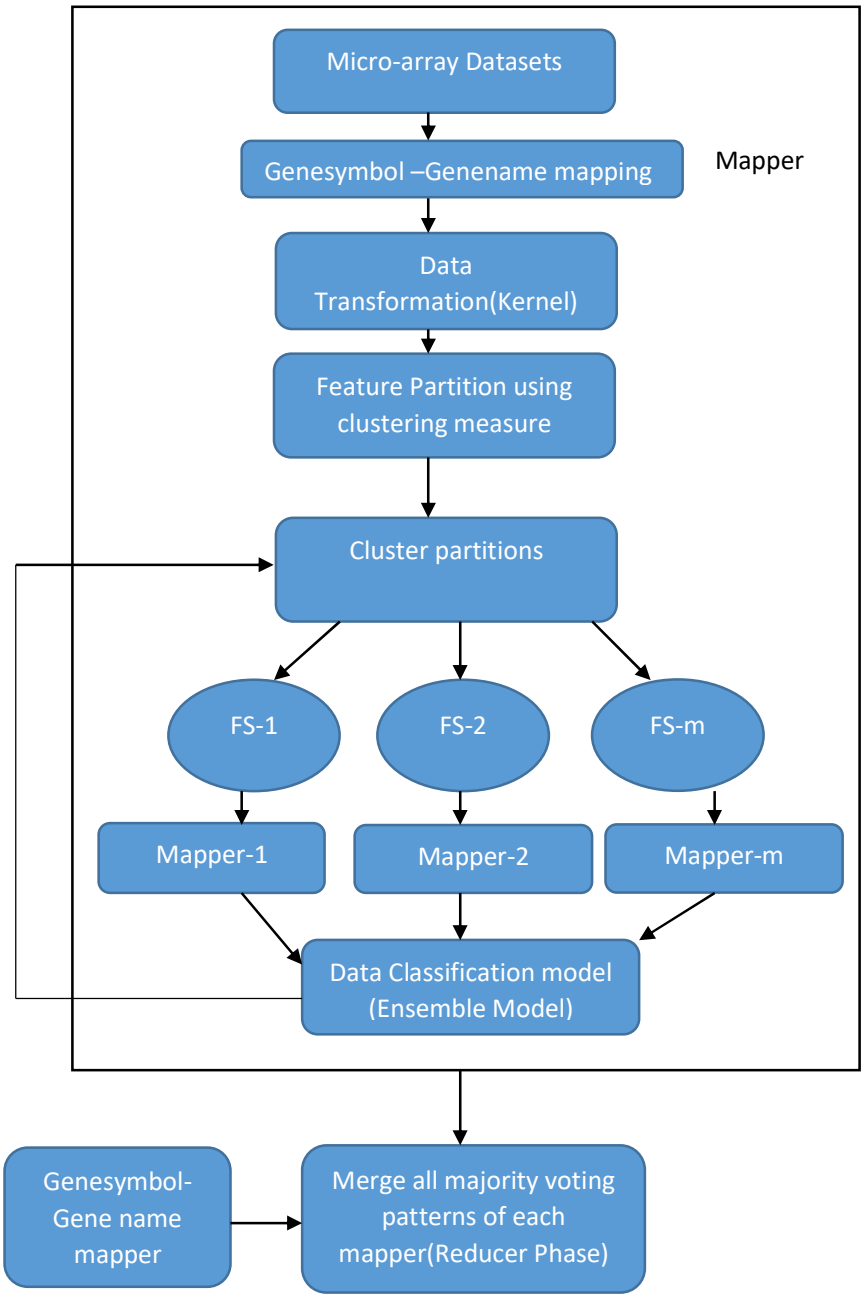

Figure 1: Proposed Map-Reduce based Filtered feature ranking and data classification model

Algorithm 1: Micro-array Gene Symbol to Gene name Mapping

Input: Training microarray dataset D, Gene synonym database GD, F(D): Feature space of D, FS(D,i): ith Feature symbol of D, Max similarity MaxSim[]; GMap={};

Output: Filtered data with gene symbol to gene name mapping.

Procedure:

Read input data D.

For each gene symbol in feature space F(D)

Do

For each gene name GN in GD

Do

Gsim[] = {};

Compute Similarity between gene symbol and gene name using the similarity measure.

$$Gsim[] = \max\{\frac{1}{3}\{\alpha(\frac{1}{|FS(D,i)|}+\frac{1}{|GD[j]|})+(1-\frac{t}{\beta})\}, \frac{\sqrt{|FS(D,i).GD[j].Max\{|FS(D,i)|,|GD[j]|}}}{Max\{|FS(D,i)|,|GD[j]|}\}$$

done

GMap{FS(D,i)→GD[j];// Map gene symbol to Gene name at jth maximum similarity.

Done

Clustering algorithm is applied on the kernel normalized data to find efficient features partitions for wrapper feature ranking. In this approach, k-feature partitions are extracted using the clustering algorithm to k-mappers. These k-data partitions are given to k-mappers for wrapper feature selection using the ensemble learning model. Proposed wrapper feature selection is based on the classification accuracy to find the optimal microarray gene features. Finally, all the patterns of the mappers are integrated in the reducer phase. These integrated patterns are mapped to gene-names for real-time gene-disease document mapping process.

Algorithm 2: Kernel based Data Pre-processing

Input : Training microarray dataset D, F(D): Feature space of D, Max similarity MaxSim[], Threshold T.

Output: Kernel Filtering or Transformed data KD.

Procedure:

Read input data D.

For each feature F[i] in feature space F(D)

Do

Apply Kernel transformation on I as

$$\text{GeneKernelTransform}(F[i]) = \text{GKer} = \frac{1}{1+\phi^2 / \cos(\text{SumSquares}(D)}$$

Where $\phi = \sqrt{F[i](1+e^{-F[i]^2})}$

If(GKer !=0)

then

$$V[] = \frac{F[i]-\min\{F[]\}}{\text{Max}\{F[]\}-\min\{F[]\}}(1-\text{GKer})$$

End if

Done

Algorithm 3: Exponential Clustering approach on kernel filter data

Input : Kernel dataset KD, Number of clusters k

Output: Cluster partitions

Procedure:

Load kernel transformation data KD.

Initialize number of input clusters parameter k.

Randomly select k cluster features FC[] among the input kernel transformed data.

For each cluster C^K in FC[]

Do

Select a mean distance object of the cluster as representative object as C_r.

$$C_r = \sum_{i=0}^{|C^K|} C_i^K / N$$

Compute the gene similarity index between the representative object to the remaining cluster objects as

$$\text{GeneSim}(C_r, C_i^K) = | C_r - C_i^K | / \text{cosine}(C_r.C_i^K)$$

End for

Sort the gene similarity feature index values in each cluster. Select top m ranked cluster features as partitions in each cluster.

Algorithm 4: Wrapper feature ranking based classification

Traditional feature selection measures such as t-statistic, significance analysis of microarray (SAM) and signal to noise ratio (SNR) are used in [20] to rank the k-means clustered features on microarray datasets. The main problem in these feature ranking measures is the selection of appropriate genes from the high dimensional feature space using wrapper method. These ranking measures incorporate the classification accuracy and true positive rate on the selected features (>50) using the t-test, SAM and SNR measures. The modified version of SAM, SNR and t-test measures are summarized below.

Hybrid T-statistic feature ranking measure:

T-statistic ranking measure is used to find the variation in the gene features using the standard deviation of the class labels. Basically, it is the ratio of difference of the means of the class labels to the maximized standard deviation.

$$\text{HT} - \text{test} = \frac{2(\mu_P - \mu_N)}{\sqrt{\min\{\sigma_P^2/|P|, \sigma_N^2/|N|\}}} \quad \text{-----(1)}$$

where μ_P is the mean of the positive cluster class samples μ_N is the mean of the negative cluster class samples.

Hybrid Signal to noise Ratio(SNR) feature ranking measure:

It is the ratio of difference of the means of the class labels to the sum of the standard deviation of the positive and negative gene disease classes. Here, the genes with highest signal to noise ratio measure is selected as highest ranking measure for data classification.

$$\text{HSNR} = \frac{|\mu_i - \mu_j| * \max\{\sigma_P, \sigma_N\}}{2(\sigma_P + \sigma_N)} \quad \text{-----(2)}$$

where μ_P and σ_P are the mean and standard deviation of the cluster positive class samples μ_N and σ_N are the mean and standard deviation of the cluster negative class samples.

Maximized Correlation, T-test ,SNR based ranking

It is the maximization of the correlation between the features, hybrid t-test and hybrid SNR ratio. This ranking measure is used to select the optimal binary class features in each cluster.

$$MCTSNR = Max\{Correlation(ClusterFeatures:CF), \frac{2(\mu_P - \mu_N)}{\sqrt{\min\{\sigma_P^2/|P|, \sigma_N^2/|N|\}}}, \frac{|\mu_i - \mu_j| * \max\{\sigma_P, \sigma_N\}}{2(\sigma_P + \sigma_N)}\} \quad -(3)$$

For each pair of candidate features CF
Do

Hybrid Gene PCA Feature Ranking

In the proposed feature ranking model, traditional PCA algorithm is enhanced to find the most essential features in the given feature space. PCA is used to find the most significant features using the covariance between the features and their Eigen vectors. If the covariance between the features represents positive, then it indicates the strong relationship among the features. Similarly, if the covariance between the features represents negative, then it indicates the weak relationship among the features. Also, the variation in the principal components decrease as we move from the first PC to the last PC, hence the significance.

Compute covariance between features as

$$Cov(CF\{x,y\}) = \frac{\sum_{i=1}^{n}(CF[x_i] - \mu_{CF[x_i]})(CF[y_i] - \mu_{CF[y_i]})}{(n-1)} \quad -(4)$$

Compute the Eigen vector and values using the eq.(5) and eq.(6)

$$EigenValues[] = Det(\lambda I - COV(CF)) = 0 \quad ---- (5)$$

Here I is the identity matrix of same dimension as COV(CF). The corresponding Eigen vector is given as

$$(\lambda I - COV(CF))v = 0 \quad ------ (6)$$

Here the optimal eigen sum is computed as

$$OptimalEigenSum = \frac{\sum Eigenvalues[i])}{Ksmallestval(Eigenvalues[], 0.5*(Maxindex(Eigenvalues[]) + Mnindex(Eigenvalues[])} \quad ----(7)$$

Compute correlation between features as

$$Corr(CR\{x,y\}) = \frac{\sum_{i=1}^{n}(CF[x_i] - \mu_{CF[x_i]})(CF[y_i] - \mu_{CF[y_i]})}{(n-1)} \quad --(8)$$

Done

Algorithm 5: Wrapper Feature Ranking based Ensemble Classification

Proposed hybrid measure is used to minimize the runtime (ms) and to improve the true positivity and false negative rate on large datasets. Since, network training data have nominal attributes and numerical attributes, proposed measure is used to find the nominal association between the numerical and nominal attributes using the following measures for node selection.

Proposed hybrid attribute selection measure is given as

$$Gene\ Feature\ Selection = GFS(D_i) = Max\{\rho_1, \rho_2\} ----- (9)$$

$$Conditional\ \ Entropy\ \ of\ B\ on\ A\ : CE(B[]/A[]) = \sum P(A[], B[]).\log(\frac{P(A[])}{P(A[], B[])}$$

$$Conditional\ \ Entropy\ \ of\ A\ on\ B : CE(A[]/B[]) = \sum P(A[], B[]).\log(\frac{P(B[])}{P(A[], B[])}$$

$$\rho_1 = \frac{-CE(B[]/A[])^3}{(\sum A[]_i)^3\ Chi\_square(D_i)^3}$$

$$\rho_2 = \frac{-CE(B[]/A[]).CE(A[]/B[])}{\sum A[]_i)^3 \sqrt{Chi\_square/(N(m-1)}}$$

N = total observations

m = minimum(#rows, #columns)

$$MGFS = \max\{GFS(D), HPCA(D), MCTSNR(D)\} ---(10)$$

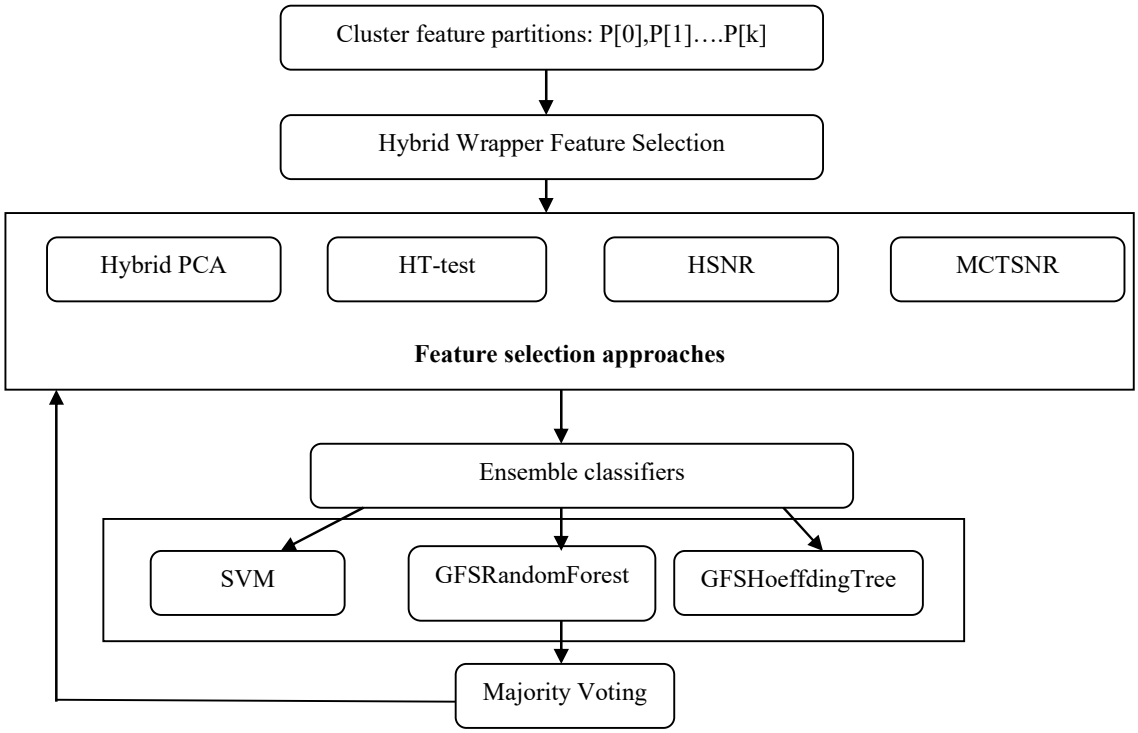

Figure 2: Ensemble feature selection and classification process

Wrapper feature selection based Ensemble Classifier for gene pattern discovery

Input: Cluster partitions;Maximum features Max; Ranked Features RF.

Output: Gene patterns

Procedure:

For each cluster partition in Mapper M[i]

Do

Let the set of gene feature ranking measure are represented as GF.

GF[]={"HPCA","HT-test","HSNR","MCTSNR"};

Let the set of base classifiers are denotes as C[]={"SVM","GFSRandomForest","GFSHoeffdingTree"};

Apply feature selection method GF[] on the cluster partition using equations (1)-(8).

Sort features using the gene ranking values.

Select Max ranked features RF[] from the sorted list for ensemble gene pattern discovery.

Apply the classification models C[] on the partition using ClassPredictions CP[]={};

For each classifier C[i] do

Do

If(CP[0]=="SVM")

CP[0]=Classify(C[0],RF[]);

Else if(CP[1]=" GFSRandomForest")

CP[1]=Classify(C[1],RF[]) using equation (9) and (10) as attribute selection measure for decision tree construction.

Else if(CP[2]=" GFSHoeffdingTree")

CP[2]=Classify(C[2],RF[]) using equation (9) and (10) as attribute selection measure for decision tree construction.

To select the optimal gene for disease prediction the majority voting of the CP[0],CP[1] and CP[2] are considered to improve the highest true positive rate and accuracy.

End for

End for

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed model to the existing models, different microarray datasets were selected from the biomedical repository. Different dataset used for experimental evaluation are summarized in Table 1. In the experimental results, 10% of the training data are used as testing data for performance evaluation. Proposed feature selection based ensemble methods increase the performance of true positive rate and accuracy on entire high dimensional datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore the prediction accuracy of each cross

validation tends to be more accurate than the traditional ensemble classification models. From the experimental results, it is clear that proposed ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

Micro array Datasets	Gene sets	Data-Type
Prostate	2136	Continuous/Numeric
Lymphoma	5000	Continuous/Numeric
DLBCL-Stanford	4000	Continuous/Numeric
Breast cancer	24481	Continuous/Numeric
Leukemia	7129	Continuous/Numeric

Table. 1 Datasets and Its Characteristics

Proposed model increase the performance of true positive rate and accuracy on entire high dimensional microarray datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models.

Leukaemia Data Results

Attribute	Statistical properties
AFFX-BioB-5_at	
mean	-101.2368
std. dev.	88.4871
AFFX-BioB-M_at	
mean	-163.7105
std. dev.	112.4806
AFFX-BioB-3_at	
mean	-4.0789
std. dev.	112.6976
AFFX-BioC-5_at	
mean	199.5789
std. dev.	113.5188
AFFX-BioC-3_at	
mean	-265.3421
std. dev.	117.838
AFFX-BioDn-5_at	
mean	-390.8421
std. dev.	148.1833
AFFX-BioDn-3_at	
mean	-73.1316
std. dev.	279.7015
AFFX-CreX-5_at	
mean	-188.6842
std. dev.	107.76
AFFX-CreX-3_at	
mean	81.7368
std. dev.	97.2059
AFFX-BioB-5_st	
mean	126.5526
std. dev.	197.4741
AFFX-BioB-M_st	
mean	-7.9211
std. dev.	158.111
AFFX-BioB-3_st	
mean	-666.5526
std. dev.	286.5725
AFFX-BioC-5_st	
mean	-505.5263
std. dev.	273.1509
AFFX-BioC-3_st	
mean	-190
std. dev.	149.741
AFFX-BioDn-5_st	
mean	110.0789
std. dev.	125.5154
AFFX-BioDn-3_st	
mean	169.9211
std. dev.	152.4278
AFFX-CreX-5_st	
mean	-71.8684
std. dev.	86.4755
AFFX-CreX-3_st	
mean	-330.1842
std. dev.	211.5202
hum_alu_at	
mean	25185.9474
std. dev.	10636.1532
AFFX-DapX-5_at	
mean	-11.3158
std. dev.	147.0344
AFFX-DapX-M_at	
mean	134.5526
std. dev.	122.7517
AFFX-DapX-3_at	
mean	-87.5789
std. dev.	65.6869

AFFX-LysX-5_at
mean 22.1579
std. dev. 43.3593

AFFX-LysX-M_at
mean -174.2368
std. dev. 270.3656

AFFX-LysX-3_at
mean 21.7105
std. dev. 451.2567

AFFX-PheX-5_at
mean -90.5526
std. dev. 59.576

AFFX-PheX-M_at
mean -129.1053
std. dev. 59.2625

AFFX-PheX-3_at
mean -3.7632
std. dev. 59.7845

AFFX-ThrX-5_at
mean -24.5
std. dev. 49.6046

AFFX-ThrX-M_at
mean -41.6316
std. dev. 56.4467

AFFX-ThrX-3_at
mean -303.9474
std. dev. 154.8755

AFFX-TrpnX-5_at
mean 9.5
std. dev. 56.4571

AFFX-TrpnX-M_at
mean -623.3158
std. dev. 438.8245

AFFX-TrpnX-3_at
mean -294.7105
std. dev. 288.4411

AFFX-HUMISGF3A/M97935_5_at
mean -258.6316
std. dev. 303.6769

AFFX-HUMISGF3A/M97935_MA_at
mean -60.4474
std. dev. 536.8284

AFFX-HUMISGF3A/M97935_MB_at
mean 204.3158
std. dev. 321.5426

AFFX-HUMISGF3A/M97935_3_at
mean 727.7895
std. dev. 654.4552

AFFX-HUMRGE/M10098_5_at
mean 3000.5263
std. dev. 5364.5012

AFFX-HUMRGE/M10098_M_at
mean 1814.4211
std. dev. 3808.2352

AFFX-HUMRGE/M10098_3_at
mean 1864.0263
std. dev. 4888.2072

AFFX-HUMGAPDH/M33197_5_at
mean 15206
std. dev. 5983.6448

AFFX-HUMGAPDH/M33197_M_at
mean 13265.0789
std. dev. 4299.9562

AFFX-HUMGAPDH/M33197_3_at
mean 18231.4474
std. dev. 3929.4593

AFFX-HSAC07/X00351_5_at
mean 15089.7895
std. dev. 5852.6113

PROPOSED PATTERNS

AFFX-CreX-5_at <= -202
| AFFX-BioC-5_at <= 88: ALL (3.21)
| AFFX-BioC-5_at > 88
| | AFFX-BioC-5_at <= 241: AML (13.5)
| | AFFX-BioC-5_at > 241
| | | AFFX-BioB-5_st <= -144: ALL (2.89)
| | | AFFX-BioB-5_st > -144: AML (9.96/1.29)
AFFX-CreX-5_at > -202
| AFFX-BioC-5_at <= 38: AML (8.04/1.29)
| AFFX-BioC-5_at > 38
| | AFFX-BioB-5_st <= -56: AML (11.25/4.5)
| | AFFX-BioB-5_st > -56
| | | AFFX-BioB-5_st <= 506: ALL (20.25/0.64)
| | | AFFX-BioB-5_st > 506: AML (2.89/0.64)

Number of Leaves : 8

Size of the tree : 15

Weight: 2.03

PROPOSED PATTERNS

AFFX-BioC-5_at <= 328
| AFFX-BioDn-5_at <= -246: ALL (58.5/13.27)
| AFFX-BioDn-5_at > -246
| | AFFX-BioC-5_at <= 141: AML (5.09)

```
|  |  AFFX-BioC-5_at > 141: ALL (3.82/1.27)
AFFX-BioC-5_at > 328: AML (4.59)
```

Lung-cancer Michigan

```
PROPOSED PATTERNS
------------------

AB000114_at <= 91
|  AB000220_at <= 616.6: Tumor (24.6)
|  AB000220_at > 616.6: Normal (4.09/0.88)
AB000114_at > 91
|  AB000114_at <= 118: Normal (48.33/2.19)
|  AB000114_at > 118: Tumor (18.99/3.36)

Number of Leaves  :  4

Size of the tree :  7


Weight: 2.64

PROPOSED PATTERNS
------------------

AB000114_at <= 91: Tumor (21.44/1.72)
AB000114_at > 91
|  AB000460_at <= 942.6
|  |  AB000449_at <= 99: Tumor (3.9/0.86)
|  |  AB000449_at > 99
|  |  |  AB000449_at <= 130.9: Normal (35.18/0.58)
|  |  |  AB000449_at > 130.9
|  |  |  |  AB000449_at <= 164.4: Tumor (14.62)
|  |  |  |  AB000449_at > 164.4: Normal (13.54/0.08)
|  AB000460_at > 942.6: Tumor (7.32/0.9)

Number of Leaves  :  6

Size of the tree :  11


Weight: 3.1

PROPOSED PATTERNS
------------------

AB000409_at <= 502.8
|  AB000220_at <= 702.8
|  |  AB000409_at <= 236.6: Tumor (3.83)
|  |  AB000409_at > 236.6
|  |  |  AB000449_at <= 151.9: Normal (64.06/5.67)
|  |  |  AB000449_at > 151.9: Tumor (3.43/0.48)
|  AB000220_at > 702.8
|  |  AB000114_at <= 228.3: Tumor (9.61)
|  |  AB000114_at > 228.3: Normal (6.56)
AB000409_at > 502.8: Tumor (8.51)

Number of Leaves  :  6

Size of the tree :  11


Weight: 2.68

PROPOSED PATTERNS
------------------

AB000449_at <= 151.5
|  AB000409_at <= 248.7: Normal (12.32/1.35)
|  AB000409_at > 248.7
|  |  AB000220_at <= 523.9: Tumor (38.9/0.51)
|  |  AB000220_at > 523.9
|  |  |  AB000220_at <= 702.8
|  |  |  |  AB000409_at <= 371.9: Tumor (3.99)
|  |  |  |  AB000409_at > 371.9
|  |  |  |  |  AB000114_at <= 54.1: Tumor (2.22)
|  |  |  |  |  AB000114_at > 54.1: Normal (14.71/0.32)
|  |  |  AB000220_at > 702.8: Tumor (8.43)
AB000449_at > 151.5: Normal (15.43/2.86)

Number of Leaves  :  7

Size of the tree :  13

Weight: 2.89

PROPOSED PATTERNS
------------------

AB000409_at <= 236.6: Tumor (19.46)
AB000409_at > 236.6
|  AB000460_at <= 578.8
|  |  AB000114_at <= 53.4: Tumor (2.06)
|  |  AB000114_at > 53.4: Normal (10.72/0.22)
|  AB000460_at > 578.8
|  |  AB000449_at <= 118: Tumor (18.67)
|  |  AB000449_at > 118
|  |  |  AB000449_at <= 132.2: Normal (9.76/1.99)
|  |  |  AB000449_at > 132.2
|  |  |  |  AB000460_at <= 599.3: Normal (2.81)
|  |  |  |  AB000460_at > 599.3
|  |  |  |  |  AB000449_at <= 178.3: Tumor (17.65)
|  |  |  |  |  AB000449_at > 178.3
|  |  |  |  |  |  AB000449_at <= 188.1: Normal (5.27/1.45)
|  |  |  |  |  |  AB000449_at > 188.1: Tumor (9.6)

Number of Leaves  :  9

Size of the tree :  17
Average Classificatin Accuracy :0.972
Average TP Rate :0.969
Average Recall :0.972
Mean Absolute Error :0.047
Average Runtime of Each partition :2932.13
```

BRCA Cancer Results:

PROPOSED PATTERNS

---

dbSNP = true: false (229.72/19.06)

dbSNP = false

| SeqContent = ATT: false (18.85/2.97)

| SeqContent = CTT: true (4.75)

| SeqContent = GTT: false (14.28/4.75)

| SeqContent = TAT: false (45.06/0.59)

| SeqContent = AAA: true (10.3/3.18)

| SeqContent = CAA: false (15.67/2.97)

| SeqContent = AAC: false (23.21/4.15)

| SeqContent = CAC: false (18.26/2.37)

| SeqContent = GAA

| | mutAss = neutral: false (10.12/0.59)

| | mutAss = low: true (3.56)

| | mutAss = medium: false (5.55/2.37)

| | mutAss = high: false (0.0)

| | mutAss = stopgain: true (0.59)

| | mutAss = stoploss: false (0.0)

| SeqContent = AAG: false (27.37/8.31)

| SeqContent = CAG: false (28.76/6.53)

| SeqContent = GAC: false (19.44/3.56)

| SeqContent = GAG: true (21.4/9.53)

| SeqContent = TGA: true (26.14/9.53)

| SeqContent = TGC: false (22.62/3.56)

| SeqContent = TCA: false (0.0)

| SeqContent = AAT: false (21.22/5.34)

| SeqContent = TCC: false (0.0)

| SeqContent = TGG: true (6.53)

| SeqContent = CAT

| | mutAss = neutral: false (20.04/4.15)

| | mutAss = low: true (4.15)

| | mutAss = medium: true (2.37)

| | mutAss = high: false (0.0)

| | mutAss = stopgain: false (0.0)

| | mutAss = stoploss: false (0.0)

| SeqContent = TCG: false (0.0)

| SeqContent = GAT

| | mutAss = neutral: false (4.36/1.19)

| | mutAss = low: true (4.75)

| | mutAss = medium: false (15.08/2.37)

| | mutAss = high: true (0.59)

| | mutAss = stopgain: false (0.0)

| | mutAss = stoploss: false (0.0)

| SeqContent = TGT: false (17.84/8.31)

| SeqContent = TTA: true (1.19)

| SeqContent = TTC: false (11.9/2.37)

| SeqContent = TCT: false (0.0)

| SeqContent = TTG: true (0.59)

| SeqContent = TTT: false (7.54/1.19)

| SeqContent = AGA

| | mutAss = neutral: true (4.75)

| | mutAss = low: true (10.09)

| | mutAss = medium: false (27.16/11.27)

| | mutAss = high: true (1.19)

| | mutAss = stopgain: true (0.59)

| | mutAss = stoploss: true (0.0)

| SeqContent = CGA: true (13.47/6.35)

| SeqContent = AGC: true (2.97)

| SeqContent = CGC: false (8.13/1.78)

| SeqContent = ACA: false (0.0)

| SeqContent = CCA: false (0.0)

| SeqContent = GGA

| | mutAss = neutral: true (3.82/0.86)

| | mutAss = low: false (10.48/1.78)

| | mutAss = medium: false (10.21/4.75)

| | mutAss = high: true (1.53/0.34)

| | mutAss = stopgain: true (2.29/0.51)

| | mutAss = stoploss: false (0.0)

| SeqContent = ACC: false (0.0)

| SeqContent = AGG: true (11.27)

| SeqContent = CCC: false (0.0)

| SeqContent = CGG: true (1.78)

| SeqContent = GGC: false (31.94/6.53)

| SeqContent = GCA: false (0.0)

| SeqContent = ACG: false (0.0)

| SeqContent = CCG: false (0.0)

| SeqContent = GCC: false (0.0)

| SeqContent = GGG

| | mutAss = neutral: false (3.77/0.59)

| | mutAss = low: false (8.13/1.78)

| | mutAss = medium: true (2.37)

| | mutAss = high: true (1.78)
| | mutAss = stopgain: true (1.78)
| | mutAss = stoploss: false (0.0)
| SeqContent = GCG: false (0.0)
| SeqContent = AGT: false (12.5/2.97)
| SeqContent = ATA: false (10.72/1.19)
| SeqContent = ATC: false (4.96/1.78)
| SeqContent = CGT
| | mutAss = neutral: false (4.05/0.88)
| | mutAss = low: true (0.64)
| | mutAss = medium: true (3.83)
| | mutAss = high: true (0.0)
| | mutAss = stopgain: true (0.0)
| | mutAss = stoploss: true (0.0)
| SeqContent = CTA: true (3.56)
| SeqContent = ACT: false (0.0)
| SeqContent = CTC: true (5.34)
| SeqContent = ATG: false (11.1/4.75)
| SeqContent = CCT: false (0.0)
| SeqContent = GGT: false (19.23/6.53)
| SeqContent = GTA: false (16.06/6.53)
| SeqContent = TAA: false (11.9/2.37)
| SeqContent = CTG
| | mutAss = neutral: false (4.73/1.55)
| | mutAss = low: true (2.57)
| | mutAss = medium: false (7.52/1.17)
| | mutAss = high: false (0.0)
| | mutAss = stopgain: false (0.0)
| | mutAss = stoploss: true (0.64)
| SeqContent = GTC
| | mutAss = neutral: true (1.78)
| | mutAss = low: true (2.37)
| | mutAss = medium: false (7.54/1.19)
| | mutAss = high: false (0.0)
| | mutAss = stopgain: false (0.0)
| | mutAss = stoploss: false (0.0)
| SeqContent = TAC: true (13.26/3.18)
| SeqContent = GCT: false (0.0)
| SeqContent = GTG
| | mutAss = neutral: false (7.22/0.87)
| | mutAss = low: true (6.39/3.18)
| | mutAss = medium: true (1.85)
| | mutAss = high: false (0.0)
| | mutAss = stopgain: false (0.0)
| | mutAss = stoploss: false (0.0)
| SeqContent = TAG: true (7.92/3.18)

Number of Leaves : 115

Size of the tree : 127

Weight: 1.47

PROPOSED PATTERNS

---

fre <= 0.01

| SeqContent = ATT: true (30.21/13.78)
| SeqContent = CTT: true (4.01/1.09)
| SeqContent = GTT
| | mutAss = neutral: true (11.84/2.32)
| | mutAss = low: false (2.32)
| | mutAss = medium: true (3.17)
| | mutAss = high: false (1.95)
| | mutAss = stopgain: true (0.0)
| | mutAss = stoploss: true (0.0)
| SeqContent = TAT
| | dbSNP = true: true (11.41/2.92)
| | dbSNP = false: false (28.94/1.59)
| SeqContent = AAA
| | polyphen = benign: true (2.38/0.73)
| | polyphen = probably: true (1.59)
| | polyphen = possibly: false (9.63/1.14)
| SeqContent = CAA: false (16.84/7.93)
| SeqContent = AAC: false (25.02/11.1)
| SeqContent = CAC
| | mutAss = neutral: false (5.66/1.39)
| | mutAss = low: false (0.48/0.12)
| | mutAss = medium: true (6.79/1.95)
| | mutAss = high: false (0.0)
| | mutAss = stopgain: false (0.0)
| | mutAss = stoploss: false (0.0)
| SeqContent = GAA
| | mutAss = neutral: false (7.81/1.59)

```
|   |   mutAss = low: true (3.28/1.09)
|   |   mutAss = medium: true (8.3/1.95)
|   |   mutAss = high: true (0.0)
|   |   mutAss = stopgain: true (0.36)
|   |   mutAss = stoploss: true (0.0)
|   SeqContent = AAG: true (51.51/12.32)
|   SeqContent = CAG
|   |   dbSNP = true: false (5.11)
|   |   dbSNP = false: true (31.12/13.67)
|   SeqContent = GAC: false (20.38/9.52)
|   SeqContent = GAG
|   |   dbSNP = true: true (9.59/1.09)
|   |   dbSNP = false: false (32.77/7.3)
|   SeqContent = TGA: false (37.15/10.22)
|   SeqContent = TGC: false (21.24/9.52)
|   SeqContent = TCA: true (0.0)
|   SeqContent = AAT
|   |   dbSNP = true: false (6.2)
|   |   dbSNP = false
|   |   |   mutAss = neutral: false (7.45/1.59)
|   |   |   mutAss = low: true (4.76)
|   |   |   mutAss = medium: true (11.84/3.91)
|   |   |   mutAss = high: true (0.0)
|   |   |   mutAss = stopgain: true (0.0)
|   |   |   mutAss = stoploss: true (0.0)
|   SeqContent = TCC: true (0.0)
|   SeqContent = TGG: true (5.47/1.46)
|   SeqContent = CAT
|   |   dbSNP = true: false (4.74)
|   |   dbSNP = false: true (24.89/9.77)
|   SeqContent = TCG: true (0.0)
|   SeqContent = GAT
|   |   dbSNP = true: false (4.01)
|   |   dbSNP = false
|   |   |   polyphen = benign: true (4.54)
|   |   |   polyphen = probably: true (5.97/1.95)
|   |   |   polyphen = possibly: false (8.52/2.66)
|   SeqContent = TGT: true (29.53/7.32)
|   SeqContent = TTA: true (1.46/0.73)
|   SeqContent = TTC
|   |   mutAss = neutral: false (4.27)
|   |   mutAss = low: true (1.95/0.36)
|   |   mutAss = medium: true (7.08/2.32)
|   |   mutAss = high: false (0.0)
|   |   mutAss = stopgain: false (0.0)
|   |   mutAss = stoploss: false (0.0)
|   SeqContent = TCT: true (0.0)
|   SeqContent = TTG: false (2.55/0.36)
|   SeqContent = TTT: false (7.44/3.17)
|   SeqContent = AGA: true (51.22/10.86)
|   SeqContent = CGA: false (21.73/4.38)
|   SeqContent = AGC: true (2.55/0.73)
|   SeqContent = CGC: false (10.86/4.76)
|   SeqContent = ACA: true (0.0)
|   SeqContent = CCA: true (0.0)
|   SeqContent = GGA: true (41.55/11.96)
|   SeqContent = ACC: true (0.0)
|   SeqContent = AGG: true (8.03/1.09)
|   SeqContent = CCC: true (0.0)
|   SeqContent = CGG: false (2.92/1.09)
|   SeqContent = GGC
|   |   mutAss = neutral: false (18.43/6.35)
|   |   mutAss = low: true (7.44/2.68)
|   |   mutAss = medium: true (6.71/1.95)
|   |   mutAss = high: true (0.0)
|   |   mutAss = stopgain: true (1.59)
|   |   mutAss = stoploss: true (0.0)
|   SeqContent = GCA: true (0.0)
|   SeqContent = ACG: true (0.0)
|   SeqContent = CCG: true (0.0)
|   SeqContent = GCC: true (0.0)
|   SeqContent = GGG: true (17.68/7.69)
|   SeqContent = GCG: true (0.0)
|   SeqContent = AGT
|   |   mutAss = neutral: false (8.18/1.59)
|   |   mutAss = low: true (4.76)
|   |   mutAss = medium: true (1.95/0.36)
|   |   mutAss = high: true (0.0)
|   |   mutAss = stopgain: true (0.0)
|   |   mutAss = stoploss: true (0.0)
|   SeqContent = ATA: false (10.49/3.17)
|   SeqContent = ATC
```

```
|   |   mutAss = neutral: true (0.0)
|   |   mutAss = low: true (5.49/0.73)
|   |   mutAss = medium: false (2.32)
|   |   mutAss = high: true (0.0)
|   |   mutAss = stopgain: true (0.0)
|   |   mutAss = stoploss: true (0.0)
|   SeqContent = CGT: true (6.46/1.95)
|   SeqContent = CTA: true (3.28/1.09)
|   SeqContent = ACT: true (0.0)
|   SeqContent = CTC: true (4.38/1.09)
|   SeqContent = ATG
|   |   dbSNP = true: false (2.55)
|   |   dbSNP = false: true (16.6/3.91)
|   SeqContent = CCT: true (0.0)
|   SeqContent = GGT: true (24.04/6.59)
|   SeqContent = GTA: true (23.68/6.23)
|   SeqContent = TAA
|   |   mutAss = neutral: false (2.32)
|   |   mutAss = low: false (0.0)
|   |   mutAss = medium: false (7.44/3.17)
|   |   mutAss = high: false (0.0)
|   |   mutAss = stopgain: true (3.17)
|   |   mutAss = stoploss: false (0.0)
|   SeqContent = CTG: true (17.44/7.69)
|   SeqContent = GTC: true (10.73/5.0)
|   SeqContent = TAC: false (16.16/6.2)
|   SeqContent = GCT: true (0.0)
|   SeqContent = GTG: false (19.09/6.09)
|   SeqContent = TAG: false (11.78/2.92)
fre > 0.01: false (64.4/1.59)

Number of Leaves  :  121

Size of the tree :  141

Number of Leaves  :  78
Size of the tree :  86
Average Classificatin Accuracy :0.984
Average TP Rate :0.979
Average Recall :0.985
Mean Absolute Error :0.026
```

Average Runtime of Each partition :2802.97

Table 2: Comparison of present model to the traditional models for runtime(ms) and feature ranking on on DLBCL dataset.

Avg Time and Number of Gene features on DLBCL		
Model	**Time(ms)**	**#Features Ranking**
PSO	8739	20
ACO	8343	20
Fuzzy PSO	8174	20
KM+SNR	8083	20
KM+t-test	8435	20
KM+SNR	7973	20
HPCA	7698	20
HSNR	7573	20
HT-test	7973	20
MCTSNR	7291	20

Table 2, describes the comparison of the present model to the traditional models for runtime comparison on selected number of feature set(20). From the table, it is observed that the present model is better than the traditional models in terms of runtime(ms).

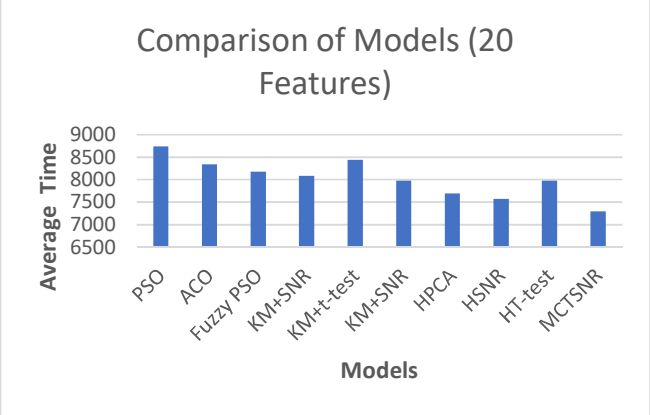

Figure 3: Comparison of 20 feature subsets.

Table 3: Comparison of present model to the existing models in terms of 50 features and its runtime on DLBCL

Avg Time and Number of Gene features on DLBCL		
Model	**Time(ms)**	**#Features Ranking**
PSO	8113	50
ACO	8634	50

Fuzzy PSO	8074	50
KM+SNR	8187	50
KM+t-test	7833	50
KM+SNR	8528	50
Proposed Model(s)		
HPCA	6867	50
HSNR	6335	50
HT-test	6363	50
MCTSNR	5464	50

Table 3, describes the comparison of the present model to the traditional models for runtime comparison on selected number of feature set(50). From the table, it is observed that the present model is better than the traditional models in terms of runtime(ms).

Note:
HRFDT= Hybrid Random Forest Decision Tree (GFSRandom Forest)

HHDT=Hybrid Hoeffding Decision Tree (GFSHoeffding Tree)

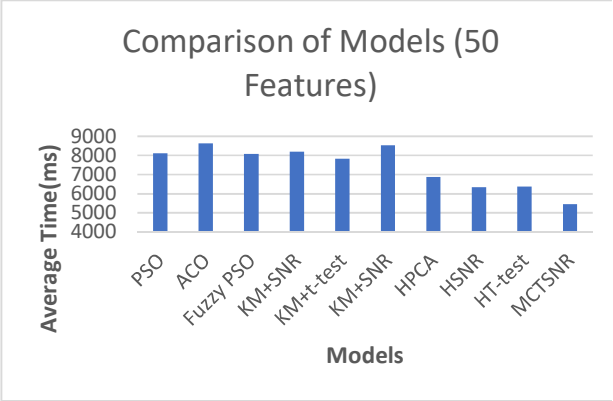

Figure 4: Comparison of 50 feature subsets

Table 4: Comparison of different Microarray datasets and its accuracy using present model to the traditional models

Model	**#Features**	**DLBCL**	**Prostate**	**Lymphoma**	**BreastCancer**
PSO+Ensemble	20	87.45	91.43	90.54	85.35
ACO+Ensemble	20	86.35	89.13	91.33	82.54
Fuzzy PSO+ Ensemble	20	83.24	87.34	89.23	84.67
KM+SNR+ Ensemble	20	94.65	91.43	92.53	86.75
KM+t-test +Ensemble	20	95.74	94.64	95.09	87.14
KM+SAM + Ensemble	20	91.53	89.13	87.45	91.43
Proposed Model(s)					
HPCA+SVM+ HRFDT+HHDT	20	96.85	95.81	96.98	93.64
HSNR+SVM+ HRFDT+HHDT	20	97.98	95.14	96.13	93.18
HT-test +SVM+ HRFDT+HHDT	20	97.15	94.92	95.93	94.82
MCTSNR+ SVM+HRFDT+ HHDT	20	98.93	96.35	97.14	94.92

Table 4, describes the comparison of the present model to the traditional models for accuracy comparison on microarray datasets. From the table, it is observed that the present model is better than the traditional models in terms of accuracy.
Table 3 and 4, describes the performance of the proposed model on all cancer datasets. Here, all the cancer datasets are evaluated using the proposed model to find the average true positive rate and precision rate on the high dimensional datasets. From the table, it is visualized that the proposed model has high true positive rate and precision over the existing models.

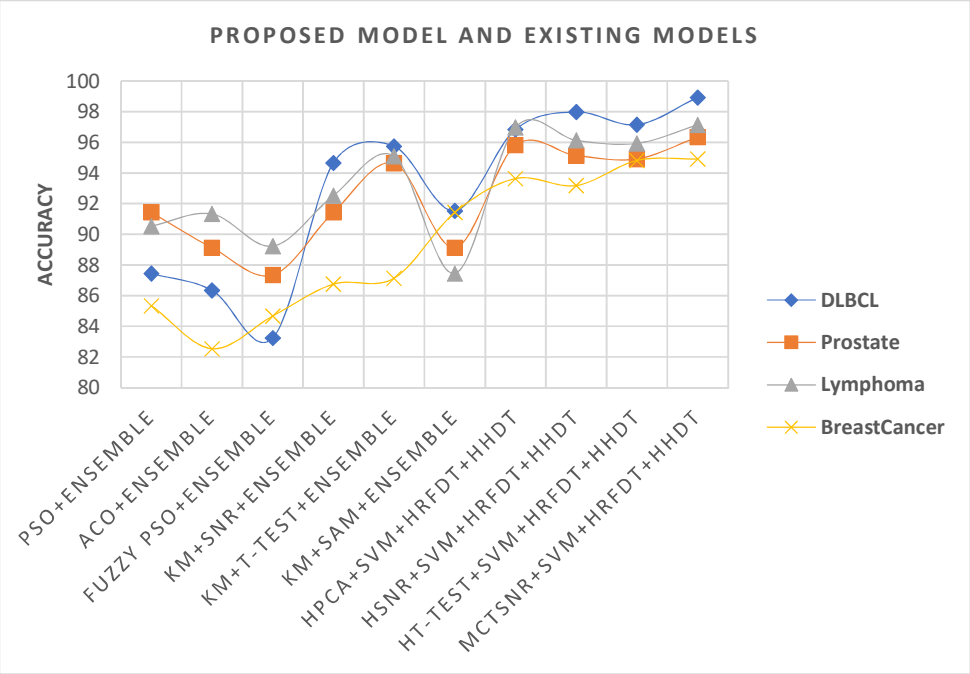

Figure 5: Comparison of proposed model to existing models on average feature seleciton

Model	DLBCL
PSO+Ensemble	87.45
ACO+Ensemble	86.35
Fuzzy PSO+Ensemble	83.24
KM+SNR+Ensemble	94.65
KM+t-test+Ensemble	95.74
KM+SAM+Ensemble	91.53
HPCA+SVM+HRFDT+HHDT	96.85
HSNR+SVM+HRFDT+HHDT	97.98
HT-test+SVM+HRFDT+HHDT	97.15
MCTSNR+SVM+HRFDT+HHDT	98.93

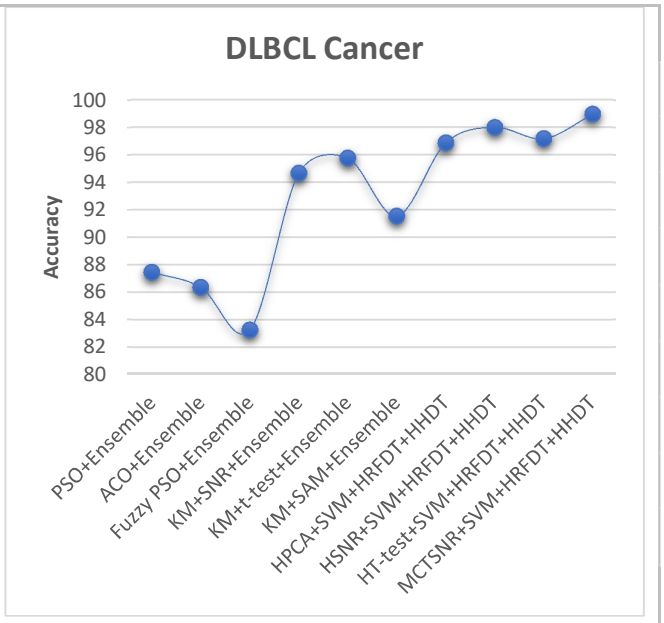

Figure 6: Comparison of proposed model to existing approaches on DLBCL cancer classification

Model	Prostate
PSO+Ensemble	91.43
ACO+Ensemble	89.13
Fuzzy PSO+Ensemble	87.34
KM+SNR+Ensemble	91.43
KM+t-test+Ensemble	94.64
KM+SAM+Ensemble	89.13
HPCA+SVM+HRFDT+HHDT	95.81
HSNR+SVM+HRFDT+HHDT	95.14
HT-test+SVM+HRFDT+HHDT	94.92
MCTSNR+SVM+HRFDT+HHDT	96.35

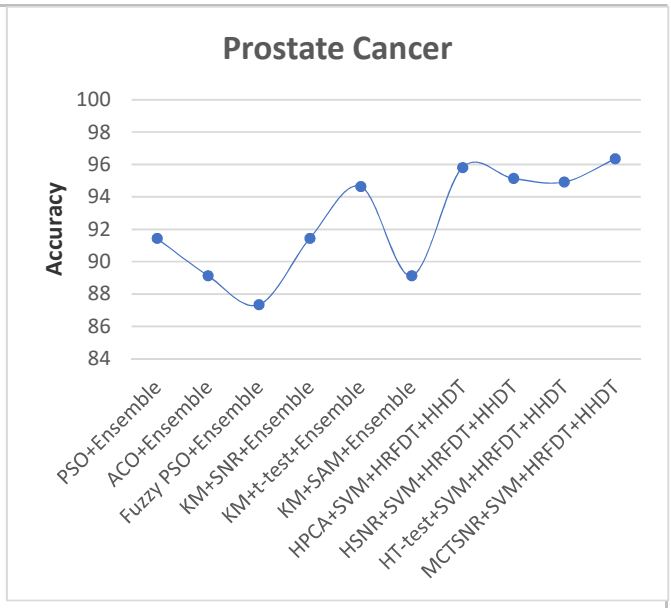

Figure 7: Comparison of proposed model to existing approaches on Prostate cancer classification

Model	Lymphoma
PSO+Ensemble	90.54
ACO+Ensemble	91.33
Fuzzy PSO+Ensemble	89.23
KM+SNR+Ensemble	92.53
KM+t-test+Ensemble	95.09
KM+SAM+Ensemble	87.45
HPCA+SVM+HRFDT+HHDT	96.98
HSNR+SVM+HRFDT+HHDT	96.13
HT-test+SVM+HRFDT+HHDT	95.93
MCTSNR+SVM+HRFDT+HHDT	97.14

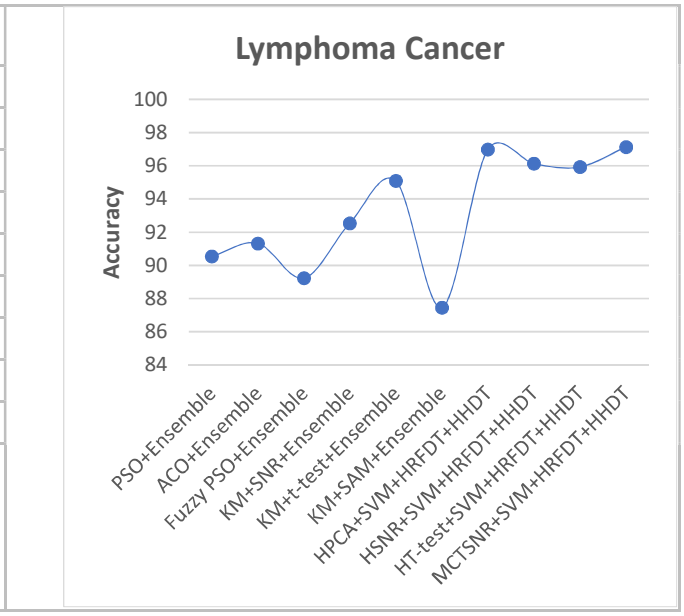

Figure 8: Comparison of proposed model to existing approaches on Lymphoma cancer classification

Model	Breast Cancer
PSO+Ensemble	85.35
ACO+Ensemble	82.54
Fuzzy PSO+Ensemble	84.67

KM+SNR+Ensemble	86.75
KM+t-test+Ensemble	87.14
KM+SAM+Ensemble	91.43
HPCA+SVM+HRFDT+HHDT	93.64
HSNR+SVM+HRFDT+HHDT	93.18
HT-test+SVM+HRFDT+HHDT	94.82
MCTSNR+SVM+HRFDT+HHDT	94.92

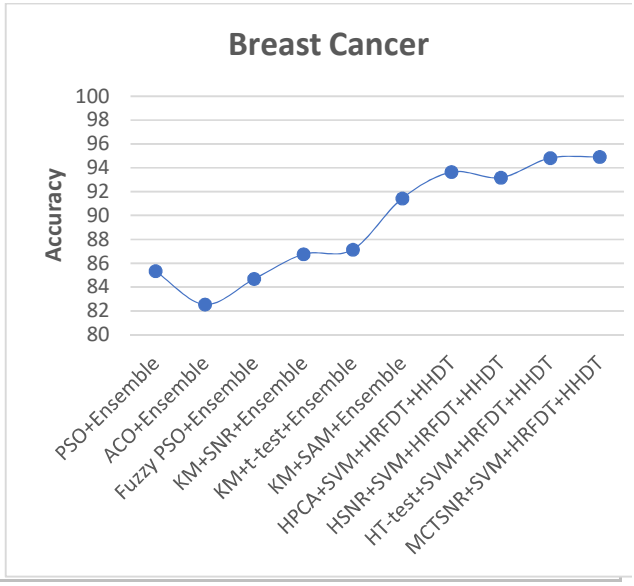

Figure 9: Comparison of proposed model to existing approaches on Breast cancer classification

Conclusion

Ensemble classification algorithm with weighted function is used to find the essential feature sets from the large number of feature space. Since, the weights in the deep neural network is optimized using the weighted function and the logistic function, proposed model efficiently classifies the large data with high dimensionality. Most of the traditional feature transformation approaches such as log transformation, min-max normalization etc. are independent of data distribution and outliers. Traditional PSO based ensemble learning and ABC based ensemble learning are improved using the heuristic activation function and ensemble classification measures. Proposed hybrid feature selection model is applied on multiple classification models to improve the true positive rate and error rate for different dimensional datasets. Experimental results are simulated on different microarray disease dataset and these results proved that the hybrid feature selection model has high true positive rate and minimal mean squared error rate compared to the traditional models. Traditional K-means based feature ranking approaches such as information gain, ANOVA, t-test, Signal-to-noise ratio (SNR) are not applicable to wrapper based feature selection approach. Wrapper based feature selection approach is applied on the high dimensional feature space to find and extract the subset of highly correlated features using ensemble classification accuracy. In this paper, a novel filter based wrapper feature selection method and data classification approach are proposed on high dimensional microarray datasets using MapReduce framework. In this model, a kernel based max-min data transformation method is used to normalize the entire microarray data for feature space partitions. Hybrid correlation based clustering method is used to partition the feature space into k-correlated features. Selected k-correlated features are given to proposed wrapper feature selection approach using data classification. Finally, a comparative analysis is performed on the different microarray datasets to study the true positive rate, error rate and runtime of the proposed model and the traditional models.

References

[1] F. Meng, C. Cai, and H. Yan, “A Bicluster-Based Bayesian Principal Component Analysis Method for Microarray Missing Value Estimation”, “IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 18, NO. 3, MAY 2014”, pp. 863-871.

[2] M. Yuan, Z. Yang and G. Ji, "Partial maximum correlation information: A new feature selection method for microarray data classification", Neurocomputing, vol. 323, pp. 231-243, 2019. Available: 10.1016/j.neucom.2018.09.084 [Accessed 19 April 2019].

[3] C. Chuang, Y. Li, J. Jeng, C. Chang and Z. Wang, “Feature Genes Selection of Adult ALL Microarray Data with Affinity Propagation Clustering”, “2015 International Conference on Consumer Electronics-Taiwan (ICCE-TW)”, pp. 230-231.

[4] L. Fan, K. Poh and P. Zhou, “Partition-conditional ICA for Bayesian classification of microarray data”, Expert Systems with Applications 37 (2010) 8188–8192.

[5] M. Sun, K. Liu, Q. Wu, Q. Hong, B. Wang and H. Zhang, "A novel ECOC algorithm for multiclass microarray data classification based on data complexity analysis", Pattern Recognition, vol. 90, pp. 346-362, 2019. Available: 10.1016/j.patcog.2019.01.047 [Accessed 19 April 2019].

[6] B. Hosseini and K. Kiani, FWCMR: A scalable and robust fuzzy weighted clustering based on MapReduce with application to microarray gene expression, Expert Systems With Applications 91 (2018) 198–210.

[7] S. F. Hussain and Md Ramazan, Biclustering of human cancer microarray data using co-similarity based co-clustering, Expert Systems With Applications 55 (2016) 520–531

[8] N. Iam-On and T. Boongoen, Diversity-driven generation of link-based cluster ensemble and application to data classification, Expert Systems with Applications 42 (2015) 8259–8273.

[9] C. Lee, W. Lin, Y. Chen and B. Kuo, Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method, Expert Systems with Applications 38 (2011) 4661–4667.

[10] R. Wahyu Pratama, S. Purnami and S. Rahayu, "Boosting Support Vector Machines for Imbalanced Microarray Data", Procedia Computer Science, vol. 144, pp. 174-183, 2018. Available: 10.1016/j.procs.2018.10.517 [Accessed 19 April 2019].

[11] J. Li and F. Wang, Towards Unsupervised Gene Selection: A Matrix Factorization Framework,2016.

[12] B. Liu, C. Yu, D. Z. Wang, C. C. Cheung, and H. Yan, Design Exploration of Geometric Biclustering for Microarray Data Analysis in Data Mining, IEEE TRANSACTIONS ON

PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 10, OCTOBER 2014, pp. 2540-2551.

[13] A. Mukhopadhyay and M. Mandal, Identifying Non-redundant Gene Markers from Microarray Data: A Multiobjective Variable Length PSO-based Approach, IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[14] T. Nguyen and S. Nahavandi, Modified AHP for Gene Selection and Cancer Classification using Type-2 Fuzzy Logic

[15] C. Orsenigo, C. Vercellisto , A comparative study of nonlinear manifold learning methods for cancer microarray data classification, Expert Systems with Applications 40 (2013) 2189–2197.

[16] S. S. Ray, A. Ganivada, and S. K. Pal, A Granular Self-Organizing Map for Clustering and Gene Selection in Microarray Data, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, pp. 1-17.

[17] I. Saha, U. Maulik, S. Bandyopadhyay and D. Plewczynski , Improvement of new automatic differential fuzzy clustering using SVM classifier for microarray analysis, Expert Systems with Applications 38 (2011) 15122–15133.

[18] T. Wong and K. Liu, A Probabilistic mechanism based on clustering analysis and distance measure for subset gene selection, Expert Systems with Applications 37 (2010) 2144–2149.

[19] T. Wong and D. Chen , A gene selection method for microarray data based on risk genes, Expert Systems with Applications 38 (2011) 14065–14071.

[20] Sahu, B., Dehuri, S., & Jagadev, A. (2017). Feature selection model based on clustering and ranking in pipeline for microarray data. Informatics In Medicine Unlocked, 9, 107-122. doi: 10.1016/j.imu.2017.07.004