Classification and Selection of Microarray Gene Expression Data Using an Supervised Algorithm

R J Anil Kumar, M N Veena, T Sheela

[1]Maharanis Science College for Women, Mysore, Karnataka
[2] PET Research Foundation, PES College for Engineering Mandya, Karnataka
3. Maharanis Science College for Women, Mysore, Karnataka

Abstract

Microarray data classification is a supervised learning task that is used to predict the class of any given sample from the gene expression phenotype. Here, the gene expression data of the samples are used to develop a classification model or classifier that is used to classify new samples based on different classes. It is through in two stages; in the first stage, a classifier is built describing a predefined set of classes and in the second stage, the constructed model is used for classification of new samples. This work focuses on classification of microarray cancer data that contains binary class labels. Artificial Neural Network (ANN),, , , , Support Vector Machine (SVM), Optimal SVM are applied in this Paper for classifying microarray data. Microarray data samples contain large number of genes most of which are either redundant or useless or sometimes both. To improve the accuracy in classification and detection of cancer samples, relevant genes should be selected. This work proposes a combined gene selection approach for different types of microarray cancer data. The proposed approach uses a novel and efficient gene selection approach that makes use of Significance analysis, T-test statistics and Signal-to-Noise Ratio (SNR) for identifying the most promising genes. The proposed gene selection can be used separately or combined with a clustering algorithm.

1. Introduction

Gene expression analysis deals with a massive database by tracking the gene expressions of numerous genes in a parallel fashion. Due to several advantages, microarray technology is employed widespread by the healthcare professionals to analyse and study the gene expressions of any living organism. Hence, this technology helps the healthcare professionals to handle the expression levels of any number of genes in a single attempt. Additionally, the microarray technology enables the beneficiary about the behaviour and growth of genes in the living organism. Though there are numerous existing data mining techniques in the literature, they are not suitable for microarray gene data analysis. Gene expression analysis can be carried out in several ways and the most popular technique for gene expression analysis is the DNA arrays. The main reason is that the DNA arrays provide better throughput and cost efficient. Initially the DNA explores several genes and they are placed on an active surface of a substrate. Generally, a microarray experiment is carried out by clubbing the messenger Ribonucleic Acid (mRNA) molecule with the corresponding DNA template. The expression level of the genes is quantified by considering the level of mRNA associated with the array. Based on this expression levels, the gene expression pattern can be finalized and the abnormality can be detected.

2. Related Work

Gene selection also features certain interesting hybrid algorithms such as the one proposed by Alshamlan et al. [1] that use a unique Genetic Bee Colony (GBC) algorithm for selecting informative genes for classification by combining the Genetic Algorithm (GA) and Artificial Bee Colony (ABC) algorithms. By combining the characteristics of both the algorithms, the GBC algorithm was able to handle both binary-class and multiple-class cancer datasets with better accuracy in both. Here also they have used 3 class dataset called Lymphoma. The dataset contains 3 category of cancer, namely DLBCL, CLL and FL. Han et al. [2] proposed an improved gene selection approach using Binary Particle Swarm Optimization (BPSO). Gene-to-Class Sensitivity (GCS) is identified using the BPSO and this sensitivity value is used for gene selection. The GCS is extracted from microarray data using Extreme Machine Learning (EML). EML based gene selection was tested using K-Nearest Neighbor and Support Vector Machine classifiers for better accuracy. Yang et al. [3] applied two filter approaches, and one wrapper approach for Feature selection, namely information gain, correlation-based feature selection method and binary particle swarm optimization feature selection. The selected features were used to evaluate the performance of classification. Hybrid approaches resulted in better classification accuracy than the individual approaches. Clustering is a useful technique for finding a natural grouping of a given data set with the help of distance or similarity function. Datta and Datta [4] introduced the grouping of genes into clusters based on a similarity measure. By calculating the distance between them, the similarity between objects is defined. Gene expression values are continuous quantities for which, depending on the specific problem, various distance measurements (Euclidean, Mahalanobis, Manhattan, etc.) can be calculated. The clustering algorithms are divided on the basis of the approach used to form the clusters. Mainly, partitioning and hierarchical algorithms are the two categories. The K-means clustering algorithm is an iterative approach that splits the given dataset into K partitions or clusters. Sensitivity of K-means algorithm and determining the value of K are the limitations of K-means clustering. To overcome these limitations, more new clustering algorithms have been proposed. Fang-Xiang et al. [5] proposed a clustering approach to cluster gene expression data that makes use of the Genetic Algorithm based K-means Clustering Algorithm, otherwise called as the GKMCA. In a similar way Kalyani et al. [6] has proposed a hybrid cluster algorithm combining the K-Means clustering technique with Particle Swarm Optimization (PSO) for efficient clustering. Apart from this Chandrasekar et al. [7] proposed a hybrid clustering algorithm to cluster gene expression data by combining K-Means algorithm with Cluster Centre Initialization Algorithm (CCIA). Wu [8] suggested the genetically weighted K-means algorithm as genetic algorithm hybridization and weighted K- means algorithm. In this approach, each individual is encoded by a partitioning table which uniquely determines a clustering, and genetic operators. The author shows that in terms of cluster performance and cluster quality it performs better than the K-means algorithm. Many conventional clustering algorithms have been applied on gene expression data; Thalamuthu et al. [9] compared the various clustering strategies followed to cluster gene expression data. An evaluation has been done using six popular techniques by simulating them with real-time datasets. The clustering techniques that are evaluated are K-means clustering, hierarchical clustering, model-based clustering, SOM, PAM and tight clustering. As per the evaluation a recommendation has been given to use tight clustering technique and model-based clustering technique to cluster gene expression data. To extract the biological knowledge from microarray data, Fuzzy C-Means (FCM) clustering tool has been used. FCM can get trapped in local minima when started with poor initialization and limited to clusters contained in linear subspaces of the data space because they use a fixed distance norm. Ravi et al [10] proposed two new Fuzzy Clustering (FC) algorithms based on Differential Evolution (DE): Differential Evolution fuzzy clustering 1 and Differential Evolution fuzzy clustering 2. They have compared the results with Fuzzy C- Means (FCM) algorithm and Threshold Accepting Based Fuzzy Clustering algorithms. Application of fuzzy c-means (FCM) algorithm for microarray data was described by Dembele and Kastner [11] that can link each gene with all the clusters by making use of indexes containing real-valued vectors. Use of FCM can overcome the issue that any given gene can be associated with many clusters. The values of these vector components range between 0 and 1. Index value closer to 1 denotes a strong association. The membership of the gene respect to each cluster is defined as the vector associated with the corresponding indices. The FCM technique however suffers from the parameter estimation problem. Dae-won et al. [12] proposed the Gustafson Kessel (GK) algorithm that can be used to detect clusters having different shapes. When applied to a high dimensional dataset having really large number of genes, the Gustafson Kessel (GK) algorithm becomes computationally inefficient. Eisen et al. [13] proposed an agglomerative based approach called the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and graphically represented the clustered data set. Alon et al. [14] applied a divisive based approach called deterministic-annealing algorithm to split the genes. Jiang et al. [15] proposed clustering algorithms for identifying biologically relevant groups of genes. They have discussed about the basic of clustering technique used on gene expression data and the clusters are divided into three categories for analysis such as: Sample-based clustering, Gene-based clustering and Subspace clustering. Each category is used for specific type of applications and provides specific challenges for the clustering task. Prognostic risk scores from gene expression have shown prominent clinical values as they are promising biomarkers. They can be used for the prediction of prognosis, including the identification of mortality and metastasis risks in patients. They can also be used to determine the response of patients to treatment [25]. Identifying the risk of cancer recurrence or metastasis in patients can help clinicians strategically recommend effective treatment. Furthermore, the determination of response to treatment can identify the overall survival of the patients and intuitively develop novel drugs or appropriate treatment based on each patient’s classification. In the majority of cancer types, HLA gene expression has been shown to prolong overall survival [26]. On the other hand, an increase in the expression of Human Endogenous Retrovirus K mRNA in the blood is linked to the presence of breast cancer, which shows it is a biomarker [27]. BRCA2 is another gene whose expression is associated with highly proliferative and aggressive breast cancer. The higher the expression of BRCA2, the more aggressiveness the breast cancer [5].

3. Proposed System

The proposed system is based on machine learning approaches, namely ANN and SVM to classify the microarray data. The Figure 1 shows the proposed system model. To develop the classifier, the microarray data is divided into training and test data. First k-means clustering algorithm is applied to group the microarray data. Next, the proposed system uses t-test statistics, signal-to-noise ratio and significance analysis for selecting the most informative genes. Then the selected genes are used to train the Artificial Neural Network/Support Vector Machine. Once the training is over, the test data is used to evaluate the performance trained network. The details of the gene selection and classification algorithm are presented below.

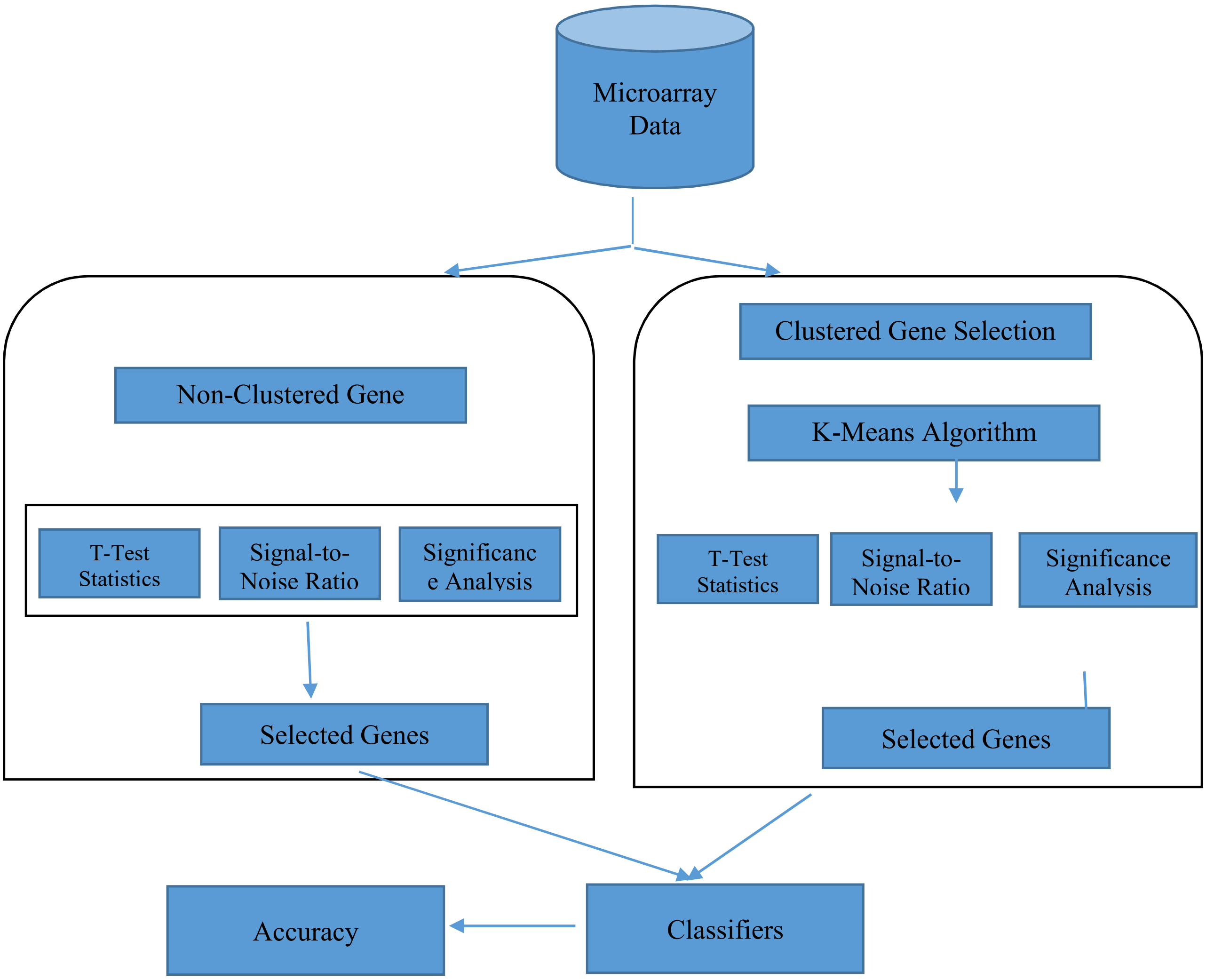


Figure 1: shows the proposed model

3.1 Gene Selection Algorithm

The efficiency of gene selection process decides the accuracy of classification. This work proposes a novel and efficient gene selection approach that effectively identifies best genes from the large collection of gene samples. First genes are clustered using the clustering algorithm [16]. Next, the algorithm uses Significance analysis, T-test statistics and Signal-to-noise ratio for selecting the suitable genes. Here two approaches have been proposed, Clustered gene selection and Non-clustered gene selection. In clustered gene selection, the genes are clustered before applying those three parameters. In non-clustered gene selection, the genes are not clustered. The three parameters are applied directly on genes for selection. In case of Non-clustered gene selection, the genes are ranked based on the three selection parameters using an effective ranking method. Efficiency in gene selection depends on the ranking method and the clustering method used [17].

3.1.1 Selection parameters

3.1.1.1 T-test Statistic

The T-test statistics is a statistical analysis that is used to identify differently expressed genes based on the positive and negative groups of the classes. In case of microarray cancer dataset, positive class denotes the sample with no cancer and negative class denotes the sample with cancer. The t value is calculated for each gene using the t-Test statistics which is the ratio of the difference in means of the positive and negative class to the standard deviations. The t value of a gene is given by equation (1).

$$t = \frac{\mu_p - \mu_n}{\sigma} \qquad (1.1)$$

Where

$$\sigma = \sqrt{\frac{\sigma_p^2}{N_p} + \frac{\sigma_n^2}{N_n}} \qquad (1.2)$$

Here, Np and Nn are the number of positive and negative instances or samples with respect to the gene. σp and σn are the standard deviations of the positive and negative classes with respect to the gene. μp and μn are the means of the positive and negative classes with respect to the gene. σ is the combined standard deviation of both positive and negative classes.

3.1.1.2 Significance Analysis

The next parameter used for identifying genes is the significance analysis that identifies the suitable genes based on the significance calculated using difference between positive and negative classes. The significance analysis provides a statistical measure that is done at gene level that can be used to differentiate between each gene. This can be used to improve the overall efficiency in gene selection. The significance s for a given gene is calculated using equation (3).

$$s = \frac{\mu_p - \mu_n}{\sigma_s + \sigma_0} \quad (1.3)$$

Where $$\sigma_s = \sqrt{\left(\frac{1}{N_p} + \frac{1}{N_n}\right)\frac{\sigma_p + \sigma_n}{N_p + N_n + 1}} \quad (1.4)$$

3.1.1.3 Signal-to-Noise Ratio (SNR)

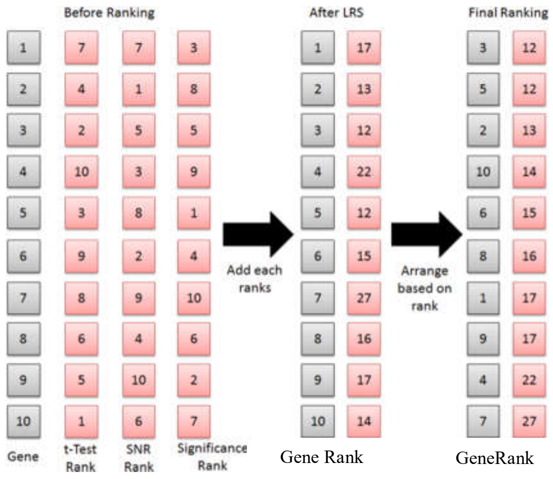
The third parameter used for the gene selection is the signal-to-noise ratio. The SNR is defined as the ratio of signal to noise. Here in our case, the signal is considered to be the mean of the class differences and the noise is considered as the standard deviation difference between the two classes. By representing the positive and negative classes of genes in the microarray data, the suitable genes can be identified. The SNR ratio is expressed as

$$SNR = \frac{\mu_p - \mu_n}{\sigma_p + \sigma_n} \quad (1.5)$$

The most suitable genes will have high significance, high SNR ratio and low t-Test value

3.2 Non-Clustered Gene Selection

In case of Non-clustered gene selection, each gene or feature is evaluated based on the calculated parameters. For each gene, the significance, SNR and the T-test are calculated. After calculating these values the genes are ranked based on a ranking scheme. In the proposed raking scheme called as Logical based Ranking Scheme (LRS) first the genes are ranked separately based on the above three parameters [18]. Next, for each gene the three ranks are added up to obtain a single rank number. Finally the genes are re-arranged based on this final ranking and then the top genes are selected. This process of Logical based Ranking Scheme has been illustrated in Figure 2.



Gene	t-Test Rank	SNR Rank	Significance Rank
1	7	7	3
2	4	1	8
3	2	5	5
4	10	3	9
5	3	8	1
6	9	2	4
7	8	9	10
8	6	4	6
9	5	10	2
10	1	6	7

Gene Rank	
1	17
2	13
3	12
4	22
5	12
6	15
7	27
8	16
9	17
10	14

GeneRank	
3	12
5	12
2	13
10	14
6	15
8	16
1	17
9	17
4	22
7	27

Figure 2: Logical Based Ranking Scheme to Rank Genes

3.3 Clustered Gene Selection

The next variant of the proposed gene selection approach is the Clustered gene selection that makes use of a clustering algorithm. In the proposed approach K-means clustering algorithm is used. Initially the number of clusters is decided and then the microarray dataset is given as input to the K-means clustering algorithm. Here the clusters are formed with respect to the genes and not the samples. By clustering with respect to genes, each gene will go into a cluster based on their similarity or distance. After the clustering process, each cluster will have many genes and using this, duplicate genes can be filtered. That is, all the genes within the same cluster will be almost similar and they can be considered as duplicates. So, only one gene from a cluster is selected. For this purpose the genes within a cluster are ranked based on the three calculated parameters and the one gene at the top that has the best value (high for SNR, significance and low for t-Test) is selected. In case, within a cluster a gene has best value for only two of the parameters and another gene has best value for the third parameter, then the gene with best value in two parameters is selected from that cluster. Only in case if three different genes have the best values of these three parameters, the LRS scheme is used for gene selection within that particular cluster. The overall process of Clustered gene selection is shown in Figure 3.

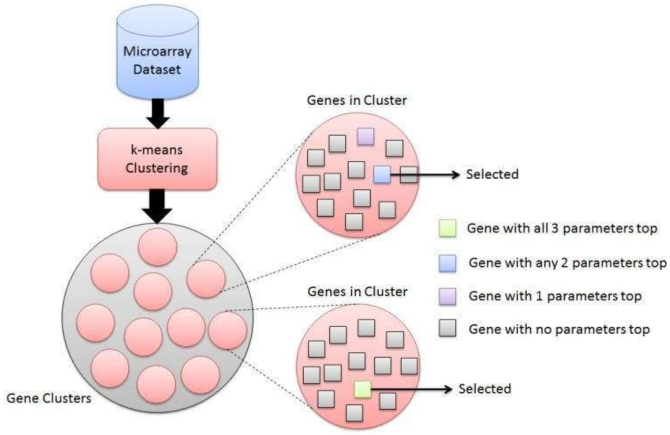

Microarray Dataset
k-means Clustering
Gene Clusters
Genes in Cluster
Selected
Gene with all 3 parameters top
Gene with any 2 parameters top
Gene with 1 parameters top
Gene with no parameters top
Genes in Cluster
Selected

Figure 3: Clustered Gene Selection Process

3.4 Classification Model for Microarray Data Classification

The final step in microarray is classification. In this work, only binary classification of microarray cancer data has been taken. In this case, there are two classes. One is “0” and other is “1” where 0 is “No Cancer” and 1 is “Cancer”. In some cases they are represented as “Benign” (No Cancer) and “Malignant” (Cancer). In case of a categorical value such as Yes or No, pre-processing is done to convert it to binary class. The selected features from the gene selection algorithm are given as input to the KNN/SVM/Optima SVM classifier.

3.4.1 K- Nearest Neighbor (KNN)

Zanaty et.al [19], Ektefa et.al [20]. KNN classifier has been both a workhorse and benchmark Classifier. Given an inquiry vector Xo and a bunch of N marked examples {xi, yi} Nl, the errand of the classifier is to foresee the class name of xO on the predefined P classes. The k-closest neighbor (k-NN) grouping calculation attempts to locate the k closest neighbors of Xo and utilizations a dominant part vote to decide the class name of X0. A measurement separation between any two elements is called as a thought of Proximity. On the off chance that two elements are in the nearness, at that point they are said to have a place with a similar class or gathering. The closest neighbor search is a technique to distinguish substances in a similar closeness in an administered way and is characterized as "Given an range of information focuses and an inquiry point in a d-dimensional measurement space, discover the information point that is nearest to the query point"

3.4.2 Support vector machine (SVM)

Sonkamble et.al [21], Zanaty et.al [22]. Support vectors are the focuses which are at the edge and the arrangement depends just on these data points. This is the exceptional element of SVM. At the point when an element space utilizes a bunch of non-straight premise work, the direct SVM can be stretched out to non-direct SVM. The information focuses can be isolated straightly in the component space with higher measurement. A critical component of the SVM is that the change need not be executed to decide the isolating hyper plane in the potentially extremely high dimensional element space. All things considered, a bit demonstration can be utilized to decide the partition of hyper plane where the arrangement assessed at the help vectors is composed as a weighted amount of the estimations of certain kernal capacities.

3.4.3 Optimal Support vector machine

SVM can be used to find out the boundary separating two classes or applying it to identify outliers in a high dimensional data set. To find out the boundary separating the two classes, or the actual number of points which separate them from each other, SVM requires a suitable kernel function which is efficient and robust. The following POLY function performs effectively with nearly all data sets, except high dimension ones, Scho et.al [23].

$$\text{POLY(x,y)}=(x^{T}z+1)^{d} \qquad (1.6)$$

$$\text{RBF(x,z)}= \exp(-\gamma\|x-z\|^{2}) \qquad (1.7)$$

$$\text{PRBF}= ((1+\exp(\omega))/V)^{d} \qquad (1.8)$$

where x = |x z|, V = p * d and p is a prescribed parameter.

$$\text{GRPF(x,z)}=\left(\frac{d+r.\exp(-\|X-z\|^{r}/(r-\sigma^{2}))}{r+d}\right)^{d+1} \qquad (1.9)$$

the values of r(r > 1) and d can be determined by optimising the parameters using the training data, and where is a statistical distribution of the probability density function of the input data.

3.4.4 Optimizing the kernel Parameters

In this work, Scholkopf et al [24] has extended. The iterative procedure for automated regression (also called machine learning) involves the following stages: 1. Initialize to some value, typically a distribution in which both the training data points and their parameters are accurate and close to their true values. 2. Extend SVM to find the maximum of the quadratic form α^0 = arg max w(). 3. Renovate such that T is minimized. This is typically achieved by a gradient step. 4. Repeat until T is minimized. Proposed by Chapelle et al [23].

3.5 Performance Evaluation

To demonstrate the efficiency of the proposed approach in microarray data classification, additional matrices like True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), True Negative Rate (TNR), Precision, Prevalence, Accuracy, F1 score are evaluated. True Positive is the case in which we predicted 1 and the actual output is also 1. True Negative is the case in which we predicted 0 and the actual output is 0. False Positive is the case in which we predicted 1 and the actual output is 0. False Negative is the case in which we predicted 0 and the actual output is 1. The equations for the above metrics are given below.

$$Accuracy = \frac{Number of correctly Classified samples}{Total number of Samples} \tag{1.10}$$

$$True Positive Rate = \frac{True Positive}{False Negative + True Positive} \tag{1.11}$$

$$False Positive Rate = \frac{False Positive}{False Positive + True Negative} \tag{1.12}$$

$$True Negative Rate = \frac{True Negative}{Falde Positive + True Negative} \tag{1.13}$$

$$False Negative Rate = \frac{False Negative}{True Positive + False Negative} \tag{1.14}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{1.15}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{1.16}$$

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1.17}$$

3.6 RESULTS AND DISCUSSION

The implementation of the proposed gene selection and SVM classification are done using MATLAB. The Colon Cancer dataset and Lymphoma dataset have been taken as input. Colon dataset contains 62 samples with each sample having 2000 genes. Out of the 62 samples, 22 samples (“0”) are positive class and 40 samples (“1”) are negative samples. The gene values in the dataset are represented as decimal values. Colon cancer is one of the most severe types of cancer which forms malignant tumors inside the walls of the large intestine. The Colon Cancer occurs mostly for age groups of 30 and above. Study of Colon Cancer is a really important study in the medical field and for the same reason this dataset is chosen. Deeper information related to a patient‟s colon cancer is obtained by doing efficient gene selection and classification. The same information can be used for better treatment. Lymphoma is a cancer that occurs in infecting cells of the immune system. The dataset contains 4026 genes and 66 samples with 3 classes namely DLBCL (Diffuse large B-cell lymphoma (DLBCL), CLL (Chronic Lymphocytic Leukemia) and FL (Follicular lymphoma), the subtype‟s of Lymphoma cancer. In the dataset, there are 46 samples belonging to DLBCL type, 9 samples belonging to CLL type and 11 samples belonging to FL type. Initially the dataset is pre-processed before applying the proposed gene selection algorithm. There is no pre-processing required for the gene selection approach but the data needs to be transposed before applying gene selection algorithm. The input dataset contains the samples as rows and to apply k-means clustering algorithm for genes, gene values should be in the row format. So, the input data is transposed in case of the clustered gene selection approach. During the non-clustered gene selection, the Significance, T-test value and Signal-to- noise ratio are calculated for all 2000 genes in the dataset using the equations given in section 3.2. The top 10 genes in each parameter for colon and lymphoma dataset are given in Table 1 and Table 2.

Table 1: Top 10 Values of Ranking Parameters for Colon Dataset

Rank	Feature | SNR	Feature | t-Test	Feature | Significance Analysis
1	249 | 0.8100	1772 | -5.6443	249 | 163.40
2	765 | 0.7795	1582 | -5.2971	822 | 129.69
3	493 | 0.7309	513 | -5.0784	66 | 111.71
4	1423 | 0.7209	1771 | -5.0588	286 | 110.04
5	245 | 0.7069	780 | -5.0403	415 | 109.12
6	267 | 0.6909	138 | -4.9354	245 | 108.54
7	377 | 0.6373	515 | -4.8644	267 | 107.86
8	822 | 0.6347	625 | -4.7949	111 | 106.25
9	1892 | 0.5786	1325 | -4.7752	201 | 104.01
10	66 | 0.5751	43 | -4.7240	765 | 102.54

Table 2: Top 10 Values of Ranking Parameters for Lymphoma Dataset

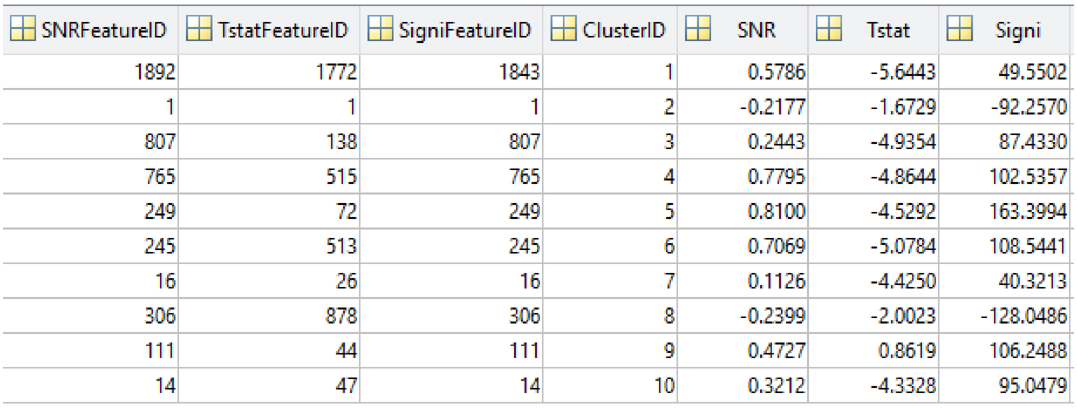
Rank	Feature | SNR	Feature|t-Test	Feature| | Significance Analysis
1	586 | 0.956	125 | -7.625	963 | 165.50
2	658 | 0.9365	562 | -7.8965	456 | 169.71
3	2369 | 0.9258	1258 | -7.0764	824 | 160.75
4	1258 | 0.8756	2586 | -7.0875	26 | 150.04
5	3658 | 0.8236	2365 | -7.0586	365 | 141.12
6	1242 | 0.8056	586 | -6.2365	1286 | 139.84
7	1269 | 0.7856	879 | -6.5693	852 | 131.56
8	878 | 0.7532	3698 | -6.1475	968 | 128.05

9	258 | 0.7502	2725 | -6.2365	2589 | 125.01
10	2568 | 0.7326	456 | -6.7240	3658 | 120.58

In the next step genes are ranked based on these metrics separately and then the ranks are added together using LRS scheme. By adding ranks from all metrics together, it is possible to select most promising genes that will be more suitable in classification. Ranking for significance and SNR ratio are made in decreasing order as these two should have a high value; and ranking for t-Test value is done in increasing order as this value has to be low to be a better gene. Number of genes to be selected is reserved as 10 and so the top 10 genes from the LRS scheme have been selected. Finally the SVM based classifier is trained by using only the selected genes. The SVM classification is implemented using the in-built MATLAB function that makes use of a linear kernel function for generating the hyperplanes. Each sample in input dataset is classified based on training data obtained using only the selected genes. Next, Clustered gene selection approach uses k-means algorithm for clustering genes in Colon cancer dataset. Value of *k* is taken as 10 similar to Non-Clustered approach. After pre-processing dataset to put genes in row-wise order, the k-means clustering algorithm is applied in MATLAB and the genes are split into 10 clusters. By using clustering mechanism to select only one gene from within a cluster, the duplicate genes can be ignored as a cluster contains all similar type of genes. The number of genes within the different cluster also varies. From the clustering results, it was evident that genes that have a different significance, t-Test value and SNR ratio are also clustered together. That is, not all similar genes are clustered together and there is a need for an enhanced and dynamic clustering mechanism especially for the process of gene selection in microarray data. This will be handled in future where an efficient clustering mechanism will be used to enhance performance of Clustered gene selection. After the clustering process, the Significance, t-Test value and SNR ratio of all genes in all clusters are calculated and the genes that have the best value of each of these metrics are selected from each cluster. For the current dataset the selected genes from the 10 clusters are shown in Table 3.

From the above Table 3, it can be seen that feature number 1 is within cluster 2 and it has best values of all three metrics within that cluster. In cluster 5, feature number 249 is selected as it has best values in two metrics. Only exception is in cluster 1, where three different features have top values for the three metrics. So in this case only 9 features have been selected and the 10th feature can be selected using the LRS scheme as in Non-Clustered approach. Table 4 shows the top 10 selected genes from Colon and Lymphoma datasets.

Table 3 Features Selected from Each Cluster



SNRFeatureID	TstatFeatureID	SigniFeatureID	ClusterID	SNR	Tstat	Signi
1892	1772	1843	1	0.5786	-5.6443	49.5502
1	1	1	2	-0.2177	-1.6729	-92.2570
807	138	807	3	0.2443	-4.9354	87.4330
765	515	765	4	0.7795	-4.8644	102.5357
249	72	249	5	0.8100	-4.5292	163.3994
245	513	245	6	0.7069	-5.0784	108.5441
16	26	16	7	0.1126	-4.4250	40.3213
306	878	306	8	-0.2399	-2.0023	-128.0486
111	44	111	9	0.4727	0.8619	106.2488
14	47	14	10	0.3212	-4.3328	95.0479

Table 4 Selected Genes from Colon and Lymphoma Dataset

S.No.	**Gene No.**	
	Colon Dataset	**Lymphoma Dataset**
1	1772	3878
2	1	3739
3	807	3670
4	765	2939
5	249	2913
6	245	2874
7	16	2863
8	306	2858
9	111	2841
10	14	2762

Finally after selecting all features KNN and SVM classifiers are applied for gene classification. Optimal SVM is trained by algorithm. Trial and error procedure was followed to identify the optimal number. The performance of Optimal SVM are given in Table 5.

Table 5: Performance of the Network

Dataset	**Number of inputs**	**No.of Epoch**	**Traing Error (Mse)**	**Training time(s)**
Colon	32	184	9.0235×10^{-4}	17.856
Lymphoma	36	255	9.6954×10^{-4}	20.7188

The performance of the network during training for Lymphoma dataset is shown in Figure 6 with number of epochs in X-axis and Mean Square Error (mse) in Y-axis.

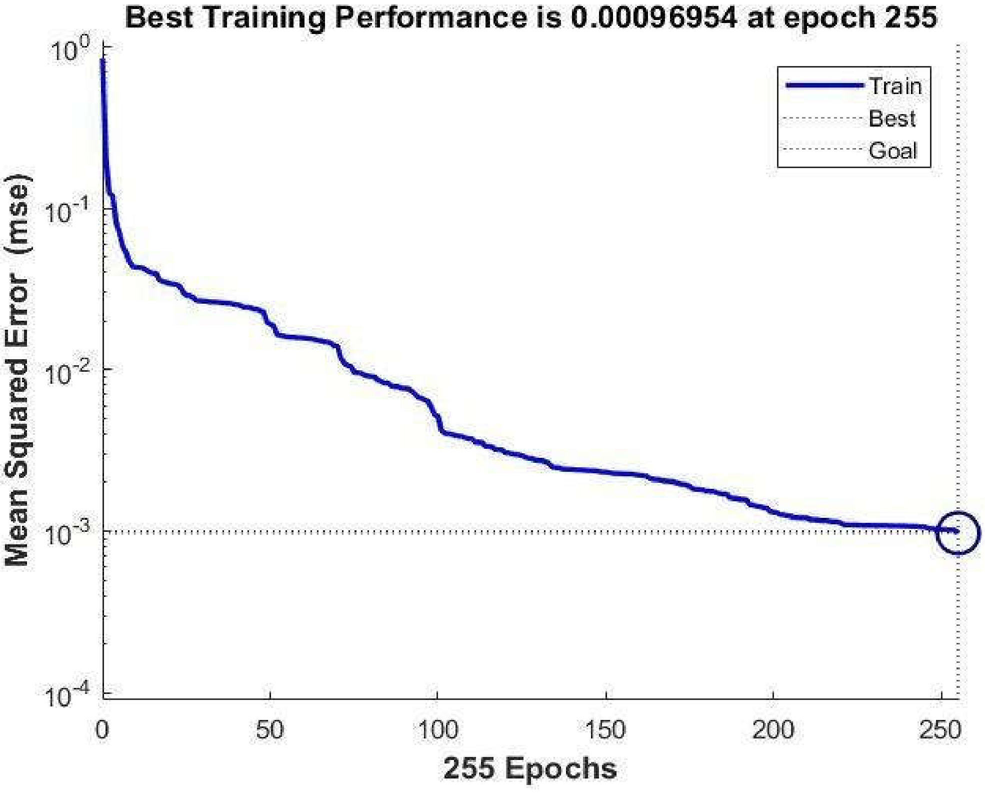


Figure 6 Training Performance Graph for Lymphoma Dataset

It is observed that the mse value decreases gradually and finally it reaches 0.00096954 at 255 epochs. As long as the mse decreases, training will get continue. Whenever the mse begins to increase, the net is starting to memorize the training pattern. At this point training is terminated. After training the network, the testing data was fed as input to the network and the output is captured. Next, the SVM classification is implemented using linear kernel function for generating the hyperplanes. Each sample in input dataset is classified based on training data obtained using only the selected genes. Then the K Nearst Nieghopur, Optimal KNN and Support Vector Machine classification accuracy are compared. In Table 3.6 the classification accuracy during testing for each class of both Colon and Lymphoma dataset is given.

Table 6: Result Comparison

Datasets	**Classifier**	**Approach**	**Classification Accuracy**
Colon	SVM	Non-Clustered Gene Selection	90%
		Clustered Gene Selection	94%
Colon	KNN	Non-Clustered Gene Selection	85%
		Clustered Gene Selection	90%
Colon	Optimal SVM	Non-Clustered Gene Selection	91%
		Clustered Gene Selection	96%
Lymphoma	SVM	Non-Clustered Gene Selection	90%
		Clustered Gene Selection	95%
Lymphoma	KNN	Non-Clustered Gene Selection	86%
		Clustered Gene Selection	88%
Lymphoma	Optimal SVM	Non-Clustered Gene Selection	90%
		Clustered Gene Selection	92%

From Table 6, we can infer that the SVM classifier produces high classification accuracy when compared with KNN and Optimal KNN in both Clustered and Non-Clustered gene selection methods. Clustered gene selection obtained good classification accuracy when compared with Non-Clustered gene selection.

Table 7: Performance Evaluation of Gene Selection using Confusion Matrix

Dataset	Gene Selection and Classifier	TPR	FPR	FNR	TNR	Precision	Prevalence	Accuracy	F1 Score
Colon	Clustered Gene selection-SVM	0.95	0.0909	0.05	0.9091	0.95	0.6452	0.9315	0.95
	Non-Clustered Gene selection-SVM	0.92	0.0902	0.03	0.9026	0.93	0.6423	0.8953	0.90
	Clustered Gene selection-KNN	0.94	0.0435	0.05	0.9565	0.9759	0.6378	0.8951	0.9878
	Non-Clustered Gene selection-KNN	0.92	0.0463	0.06	0.945	0.968	0.6789	0.8402	0.9878
Lymphoma	Clustered Gene selection-SVM	0.8286	0.0188	0.1714	0.7813	0.8923	0.6863	0.9446	0.8593
	Non-Clustered Gene selection-SVM	0.8286	0.0688	01714	0.5313	0.894	0.6863	0.8986	0.8112
	Clustered Gene selection-KNN	0.81	0.082	0.012	0.188	0.8501	0.7561	0.8802	0.8543
	Non-Clustered Gene selection-KNN	0.95	0.0909	0.05	0.9091	0.95	0.6452	0.8591	0.95

Dataset	Gene Selection and Classifier	TPR	FPR	FNR	TNR	Precision	Prevalence	Accuracy	F1 Score
Colon	Clustered Gene selection-SVM	0.97	0.1109	0.07	0.9291	0.97	0.6652	0.9515	0.97
	Non-Clustered Gene selection-SVM	0.94	0.1102	0.05	0.9226	0.95	0.6623	0.9153	0.92
	Clustered Gene selection-Optimal SVM	0.96	0.0635	0.07	0.9765	0.9959	0.6578	0.9151	1.0078
	Non-Clustered Gene selection- Optimal SVM	0.94	0.0663	0.08	0.965	0.988	0.6989	0.8602	1.0078
Lymphoma	Clustered Gene selection-SVM	0.8486	0.0388	0.1914	0.8013	0.9123	0.7063	0.9646	0.8793
	Non-Clustered Gene selection-SVM	0.8486	0.0888	1714.02	0.5513	0.914	0.7063	0.9186	0.8312
	Clustered Gene selection- Optimal SVM	0.83	0.102	0.032	0.208	0.8701	0.7761	0.9002	0.8743
	Non-Clustered Gene selection- Optimal SVM	0.97	0.1109	0.07	0.9291	0.97	0.6652	0.8791	0.97

From the Table 7 it is clear that the accuracy obtained for all datasets is more than 80% and for Colon Cancer and Lymphoma cancer with only 10 genes. The precision value for them is also above 90% in colon and above 85% in Lymphoma. This demonstrates the efficiency of the proposed gene selection algorithm for microarray data classification.

ROC Analysis

The ROC (Receiver Operating Characteristics) curve is a graphic representation of the relationship between sensitivity and specificity. The curve is plotted by using True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, starting coordinate (0,0) and ending coordinate (1,1). This curve helps to visualize the performance of the classifier.

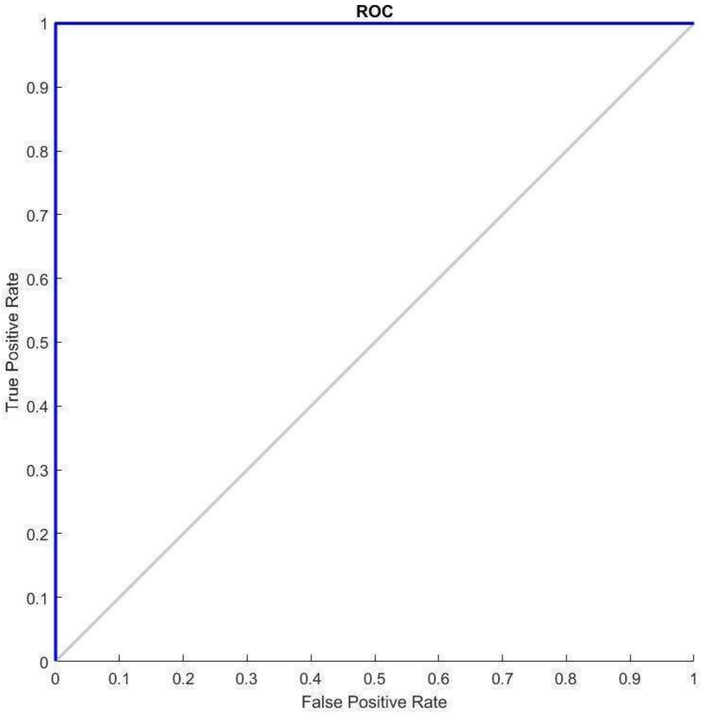


Figure 7 ROC Curve for Lymphoma

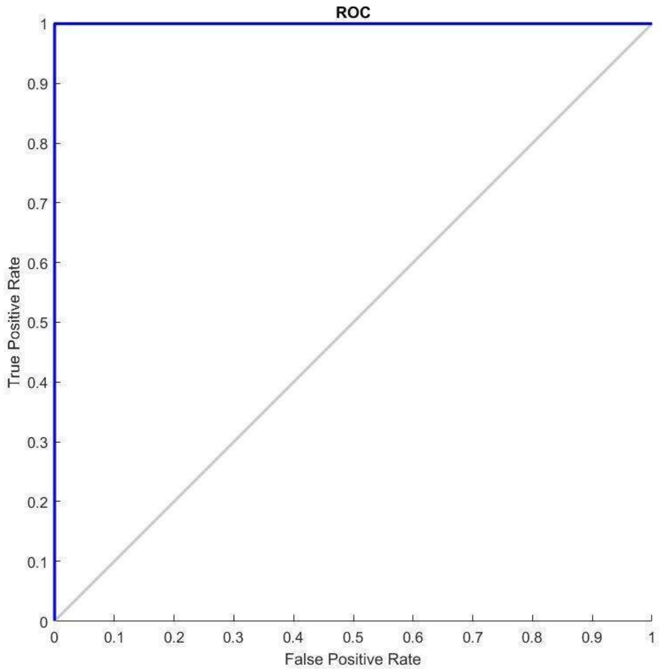


Figure 8: ROC Curve for Colon

The curve lies on the upper left corner that confirms the higher specificity, sensitivity and accuracy rate. The Figures 7 and 8 show that the classifier can able to perfectively distinguish between positive and negative classes.

3.7 Conclusion

This work mainly focused on the gene selection process in order to improve the classification accuracy. The proposed gene selection uses Significance analysis, T-test statistics and Signal-to-noise ratio parameters for gene selection. Gene selection process is implemented by ranking genes based on these parameters in Non-clustered gene selection algorithm and by using clustering approach in Clustered gene selection algorithm. The KNN/SVM/optimal SVM classifiers are used for classification. Classification results imply that proposed gene selection improves classification accuracy of microarray cancer datasets.

References

1. H. M. Alshamlan, G. H. Badr and Y. A. Alohali, Genetic Bee Colony (GBC) Algorithm: A New Gene Selection Method for Microarray Cancer Classification, Computational Biology and Chemistry, 56 (2015), 49-60.

2. F. Han, C. Yang, Y. Q. Wu, J. S. Zhu, Q. H. Ling, Y. Q. Song and D. S. Huang, A Gene Selection Method for Microarray Data based on Binary PSO Encoding Gene-to-Class Sensitivity Information, IEEE Transactions on Computational Biology and Bioinformatics, 14 (1) (2017), 85-96.
3. C. S. Yang, L. Y. Chuang, C. H. Ke, and C. H. Yang, A Hybrid Feature Selection Method for Microarray Classification, IAENG, International Journal of Computer Science, 35 (3) (2008), 1-3.
4. S. Datta and S. Datta, Evaluation of Clustering Algorithms for Gene Expression Data, BMC Bioinformatics, 7 (Suppl 4) : S17 (2006), 45-53. 110
5. W. Fang-Xiang, W.J. Zhang and A. J. Kusalik, A Genetic K-means Clustering Algorithm Applied to Gene Expression Data, Lecture Notes in Computer Science, 2671 (2003),520-526.
6. M. Kalyani, S. Hanuman, S. C. Satapathy, V.Chaganti and V. Babu, A Software Tool for Data Clustering Using Particle Swarm Optimization, Lecture Notes in Computer Science, 64 (6) (2010), 279–295.
7. T. Chandrasekhar, K. Thangavel and E. Elayaraja, Effective Clustering Algorithms for Gene Expression Data, International Journal of Computer Applications, 32 (4) (2011), 25-29. [65] F.X. Wu, Genetic Weighted K-means Algorithm for Clustering LargeScale Gene Expression Data, BMC Bioinformatics, 9 (6) (2008), 1-15.
8. A. Thalamuthu, I. Mukhopadhyay, X. Zheng and G. C. Tseng, Evaluation And Comparison of Gene Clustering Methods in Microarray Analysis, Bioinformatics, 22 (19) (2006), 2405-2415.
9. V. Ravi, N. Aggarwal and N. Chauhan, Differential Evolution Based Fuzzy Clustering, Lecture Notes in Computer Science, 6466, (2010), 38 – 45.
10. D. Dembele and P. Kastner, Fuzzy C-Means Method for Clustering Microarray Data, Bioinformatics, 19 (8) : 973 (2003), 973-980.
11. Dae-Won Kim, Kwang H Lee and Doheon Lee, Detecting Clusters of Different Geometrical Shapes in Microarray Gene Expression Data, Bioinformatics, 21 (9) (2005), 1927–1934.
12. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, Cluster Analysis And Display of Genome-Wide Expression Patterns, Proceedings of the National Academy of Sciences, 95 (25) (1998), 14863-14868.
13. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra and D. Mack, Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed By Oligo Nucleotide Arrays. Proceeding of National. Academic Science, 96 (1999), 6745–6750. 111
14. D. Jiang, C. Tang, and A. Zhang, Cluster Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, (2004), 370–1386.
15. C. Liao, S. Li and Z. Luo, Gene Selection for Cancer Classification using Wilcoxon Rank Sum Test and Support Vector Machine, IEEE International Conference on Computational Intelligence and Security, Guangzhou, China, 3 rd -6 thNovember 2006, 88-93.
16. W. Zhong, G. Altun, R. Harrison, P. C. Tai and Y. Pan, Improved k-means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property, IEEE Transactions on NanoBioscience, 4(3) (2005), 255-265.
17. D. Jiang, J. Pei and A. Zhang, DHC: A Density-Based Hierarchical Clustering Method for Time Series Gene Expression Data, Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering, (2003), 393.
18. K. Y. Yeung, and R. E. Bumgarner, Multiclass Classification of Microarraydata with Repeated Measurements: Application To Cancer, Genomebiology, 4 (12) (2003), 83–83.
19. D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Pearson Education, 2011.
20. R. Díaz-uriarte and S. A. De Andrés, Gene Selection and Classification of Microarray Data Using Random Forest, BMC Bioinformatics, 13 (2006), 1– 13.
21. https://biit.cs.ut.ee/gprofiler/gost

22. C. Devi ArockiaVanitha, D. Devaraj and M. Venkatesulu, Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection, Procedia Computer Science, 47 (2015), 13-21.
23. Naik, N., Sharath Kumar, Y.H. (2022). Efficient Feature Selection Algorithm for Gene Classification. In: Guru, D.S., Y. H., S.K., K., B., Agrawal, R.K., Ichino, M. (eds) Cognition and Recognition. ICCR 2021. Communications in Computer and Information Science, vol 1697. Springer, Cham. https://doi.org/10.1007/978-3-031-22405-8_14
24. Munkácsy, G.; Santarpia, L.; Győrffy, B. Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features. Biomedicines 2022, 10, 248.
25. Oliveira, L.J.C.; Amorim, L.C.; Megid, T.B.C.; De Resende, C.A.A.; Mano, M.S. Gene expression signatures in early Breast Cancer: Better together with clinicopathological features. Crit. Rev. Oncol. Hematol. 2022, 175, 103708.
26. Schettini, F.; Chic, N.; Brasó-Maristany, F.; Paré, L.; Pascual, T.; Conte, B.; Martínez-Sáez, O.; Adamo, B.; Vidal, M.; Barnadas, E.; et al. Clinical, pathological, and PAM50 gene expression features of HER2-low breast cancer. NPJ Breast Cancer 2021, 7, 1.
27. Zhong, Y.; Chalise, P.; He, J. Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. Commun. Stat. Simul. Comput. 2023, 52, 110–125.