Classification of Two Types of Cancer Based on Microarray Data

Basma A. Maher, Abeer M. Mahmoud, El-Sayed M. El-Horbaty, Abdel-Badeeh M. Salem

Computer Science Department, Faculty of Computer & Information Sciences
Ain Shams University, Cairo, Egypt.
eng.basma.ali.cs@gmail.com, abeer_f13@yahoo.com, sayed.horbaty@yahoo.com, abmsalem@yahoo.com

Abstract

Microarray gene expression data has a high dimensionality, e.g. small number of samples with large number of genes. Using machine learning techniques for knowledge discovery in such data become a rich area for researchers. This large number of genes, not all has the useful information that can be used to perform a certain diagnostic test, so feature selections become very important in both research and application communities of data mining. This paper proves the importance of finding the most informative genes in the database by using statistical gene selection technique to achieve a reduction in time, cost and increase the efficiency of the classifier. We applied T-Test statistical feature selection technique and K-Nearest neighbor (KNN) classifier on two public microarray data sets, SRBCT and Leukemia datasets. The feature selection is done on the whole available datasets and the data reduction results are then divided into training and testing and supplemented to the KNN classifier for cancer classification. The results showed that the T-test with KNN reached a test classification accuracy of 100% by the highest ranked 26 genes and 97.06% using the highest ranked 10 genes, for SRBCT and Leukemia respectively.

Keywords*: Microarray Gene expression, Machine learning, Gene selection, Classification, Biomedical informatics*

1. Introduction

Microarray technology has made the modern biological research. The goal of molecular biology is to understand the regulatory mechanism that governs protein synthesis and activity. All the cells in an organism carries equal number of genes yet their protein synthesis can be different due to regulation. Protein synthesis is regulated by control mechanisms at different stages 1) transcription, 2) RNA splicing, 3) translation and 4) post transitional modifications [1]. Microarray techniques provide a plat form where one can measure the expression levels of thousands of genes in hundreds of different conditions. Actually, there is a high redundancy in microarray data and numerous genes contain inappropriate information for precise classification of diseases or phenotypes [2]. Therefore, the amount of data generated by this technology presents a challenge for the biologists to carry out analysis [3].

The main types of microarray data analysis include: gene selection, clustering, and classification. Gene selection is a process of finding the genes most strongly related to a particular class [4]. The benefit of this process is to reduce not only dimensionality but also, the danger of presence of irrelevant genes that affect the classification process. Clustering can be classified into two categories: one-way clustering and two-way clustering. Methods of the first category are used to group either genes with similar behavior or samples with similar gene expressions and the two-way clustering methods are used to simultaneously cluster genes and samples [4, 5]. Classification is applied to discriminate diseases or to predict

outcomes based on gene expression patterns and perhaps even identify the best treatment for given genetic signature [4, 6]. The broad use of machine learning techniques and their applicability in the different areas of bioinformatics reported a success resolving biological problems because it facilitates the process of analyzing such data sets and extracting the important hidden knowledge. Actually, the cancer classification based on the microarray data is one example of this type of analysis. Therefore, the process of classifying this high dimension data increases the need for using an efficient gene selection technique as a very important pre-classification process for reducing time and effort and achieving high classification accuracy.

The rest of the paper is organized as follows. In section 2, a brief literature related work is presented. Section 3, details the implemented gene selection techniques for the discovery of the most informative genes and the machine learning based classifier used in our experiments with the necessary mathematical formulation. Section 4, details the public microarray data sets and, discuss the implementation results relative to comparative results in literature. Section 5 concludes the paper.

2. Related Work

Actually, gene expression analysis and gene selection research area achieved considerable advances, where, a variety of classification techniques and gene selection techniques have been proposed. Table 1 abstract some selected of the earlier related literature work proposed on gene expression cancer classification.

Table 1. Comparative study analysis

Author	Objective	Task & Techniques			results (%)
		Data Type	Gene Selection Technique	Classification Technique	
Roberto Ruiz et. al. [7] (2006)	Proposed a new heuristic to choose relevant gene subsets so as to use them for the classification task.	lymphoma, leukemia, colon cancer, GCM	a wrapper method	naive Bayes, an instance-based learner (IB1) and a decision tree learner (C4.5)	95.89
J.Zhang & H.Deng [8] (2007)	Estimating the upper bound of the bayes error to filter out redundant genes from remaining genes derived from gene per selection step.	Colon, DLBCL, Leukemia, Prostate, Lymphoma	Based Bayes Error Filter (BBF)	SVM, KNN	94.86
Chu & Wang [9] (2006)	Proposed a novel radial basis function neural network for cancer classification using expression of very few genes.	Lymphoma, SRBCT, Ovarian cancer	T-test scoring method	Radial Basis Function (RBF) neural network	100
Zhang, et.al, [10] (2007)	Presented a method for multi-category classification in cancer diagnosis with micro array data.	GCM, Lung, Lymphoma	Recursive Feature Elimination (RFE)	Extreme Learning Machine (ELM) algorithm	95

Follow Table 1. Comparative study analysis

Author	Objective	Task & Techniques			results (%)
		Data Type	Gene Selection Technique	Classification Technique	
Wang et.al, [11] (2007)	Proposed the approach for cancer classification using an expression of very few genes.	Lymphoma, SRBCT, Liver, GCM	T-test and Class Separability	FNN, SVM	100
Mallika & Saravanan [12] (2009)	Developed a new algorithm for classifying cancer gene expression data with minimal gene subsets.	Lymphoma, Liver, Leukemia	classical statistical technique	SVM-OAA and LDA	97.22
Bharathi &Natarajan [13] (2010)	Find the minimum number of genes for cancer classification	Lymphoma	analyses of Variance (ANOVA)	SVM	97.91
Osareh et.al, [14] (2010)	Distinguish between the benign and malignant tumors of breast.	Benign, breast malignant tumors	signal-to-noise ratio, sequential forward selection based & principal component analysis	SVM, KNN & probabilistic neural networks	98.80
Zantanu Ghorai et.al, [15] (2011)	Uses mutual information criterion to do minimum gene selection and reduce the computational burden.	ALL-AML, Colon, Lung, Breast, Lymphoma, Liver, Prostate	Minimum redundancy maximum relevance (MRMR) ranking method	nonparallel plane proximal classifier (NPPC)	99
Dina et.al, [16] (2011)	Achieved reasonable classification accuracy but on limited data sets.	leukemia, lymphoma, colon cancer	Mean Difference, F-score, Correlation Coefficient, and Entropy Based	SVM, KNN, LDA	90.97

3. Applied Methodology

3.1 Gene Selection Technique: T-test statistics (TS)

Gene selection techniques fall into three categories; marginal filters, wrappers and embedded methods. Marginal filter approaches are individual feature ranking methods. In a wrapper method, usually a classifier is built and employed as the evaluation criterion. If the criterion is derived from the intrinsic properties of a classifier, the corresponding feature selection method will be categorized as an embedded approach [17]. Filter methods are characterized over the two other types by being powerful, easy to implement and are stand-alone techniques which can be further applied to any classifier. They work on giving each gene a score according to a specific criterion and choosing a subset of genes above or below a specified threshold. Thus, they remove the irrelevant genes according to general characteristics of the data [16, 18]. Filter techniques are further divided into parametric and non-parametric tests. Parametric tests measure a specific property of the gene while non-parametric tests measure a degree of relation between each gene and class. Gene selection techniques can also be divided into univariate and multivariate techniques. Univariate techniques evaluate the importance of each gene individually while multivariate techniques build its evaluation on a subset of genes [16, 19].

Many of gene selection techniques are developed to reduce the number of genes in the microarray datasets to reach accurate classification accuracy with the smallest number of genes. This process reduces the computational time and the cost. Examples of gene selection

techniques most widely applied for microarray data are Mean Difference (MD), Signal to noise ratio (SNR), F(x) score (FS), Fisher discriminant criterion (FC), T-test statistics (TS), Entropy (E), Correlation Coefficient (CC), Euclidean distance (ED), and Class Separability (CS) [20].

The T-test statistics is a very famous ranking gene selection technique which is widely used by many researchers. The TS starts by calculating the Mean Difference and then normalizing it as illustrated in equations (1) and (2). Actually, the T-test is used to measure the difference between two Gaussian distributions. Then the P-values which define the difference significance are computed. Therefore, a threshold of P-values is used to determine a set of informative genes [11, 16, 21].

$$\boldsymbol{TS(i)} = \frac{\boldsymbol{\mu_{i1} - \mu_{i2}}}{\boldsymbol{S_w}\sqrt{\frac{1}{n_{s1}} + \frac{1}{n_{s2}}}} \quad (1)$$

$$S_{W^2} = \frac{(n_{s1}-1)\sigma_{i1^2} + (n_{s2}-1)\sigma_{i2^2}}{n_{s1} + n_{s2} - 2} \quad (2)$$

The standard T-test is only applicable to measure the difference between two groups. Therefore, when the number of classes is more than two, we need to modify the standard T-test. In this case, the T-test has been used to calculate the degree of difference between one specific class and the centroid of all the classes. Hence, the definition of TS for gene *i* can be described in equations from (3) to (7) [16,24].

$$TS_i = \max\left\{\left|\frac{\bar{x}_{ik} - \bar{x}_i}{m_k s_i}\right|, k = 1,2, \dots, K\right\} \quad (3)$$

Where

$$\bar{x}_{ik} = \sum_{j \in C_k} \bar{x}_{ij} / n_k \quad (4)$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n \quad (5)$$

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \quad (6)$$

$$m_k = \sqrt{1/n_k + 1/n} \quad (7)$$

Here max {yk; k = 1; 2; . . . K} is the maximum of all yk. C_k refers to class k that includes n_k samples. x_{ij} is the expression value of gene i in sample j. $\bar{x}_{ik}$ is the mean expression value in class k for gene i. n is the total number of samples. $\bar{x}_i$ is the general mean expression value for gene i. s_i is the pooled within-class standard deviation for gene i.

3.2 The Classifier Technique: KNN

Gene expression classification is the process of classifying a new gene expression samples into a predefined class. After reducing the number of the genes, we attempt to classify the data set. Examples of recent classification techniques for microarray data are Support vector Machine (SVM), K-Nearest neighbor (KNN), Fuzzy Neural Network (FNN), and Linear Discriminate Analysis (LDA) [20].

Although being a simple technique, KNN shows an outstanding performance in many cases of classifying microarray gene expression to abstract the data with a gene selection

technique. KNN is known to be a lazy technique as it depends on calculating a distance between a test data and all the training data. Therefore, for using KNN technique three key elements are essential, (1) a set of data for training, (2) a group of labels for the training data (identifying the class of each data entry) and (3) the value of K for deciding the number of nearest neighbors[16,22]. Actually, the main idea of KNN classifier is to assign a class label to a new sample where the majority of the chosen number of neighbors belongs. The KNN calculates its distances by different ways, but Euclidean distance is the most popular. Also, for achieving the highest classification accuracy, it is advised trying different values of k.

4. Experimental Result & Discussion

4.1 The Data Set

We used two public data sets for the analysis of the two selected techniques for gene selection and the classifier. A detailed description of these data sets is in the follow:

4.1.1 The SRBCT Database

We used the public SRBCT data set [http://research.nhgri.nih.gov/microarray/Supplement/]. A sample from the data is shown in Table: 2. The dataset contains of 2308 genes and 88 samples. There are totally 63 training samples and 25 testing samples, five of the testing samples doesn't belongs to SRBCTs and therefore are recognized as a noisy data. The 63 training samples contain 23 Ewing families of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 neuroblastoma (NB), and 8 Burkitt lymphomas (BL). And the 20 SRBCTs testing samples contain 6 EWS, 5 RMS, 6 NB, and 3 BL.

Pre-processing is the process of removing noisy data and filtering the necessary information. The SRBCT dataset downloaded consist of noisy and inconsistent data. Reading the description available for the SRBCT data set, such noisy data, where some additional unnecessary columns exist. After a deep study of the important columns needed to proceed our work, we removed such unnecessary columns (Test 3, Test 5, Test 9, Test 11 and Test 13) [23].

Table 2. A Sample data from SRBCT Dataset

No.	Gene ID	Name	Values (EWS)	Values (BL)	Values (NB)	Values (RMS)
11	24145	adenylyl cyclase-associated protein	1.2607	1.4646	0.5277	0.8178
12	25584	ubiquinol-cytochrome c reductase core protein II	2.9001	2.0438	1.899	2.1544
19	29054	ARP1 homolog A	1.4482	0.8015	1.3726	1.103
20	34945	Tu translation elongation factor, mitochondrial	3.3214	1.4196	2.4937	3.0199
36	39993	superoxide dismutase1, soluble	2.1497	2.5377	1.9207	3.5434

Table 3. A Sample data from Leukemia Dataset

No.	Gene ID	Name	Values (ALL)	Values (ALL)	Values (AML)	Values (AML)
63	AB000114_at	Osteomodulin	72	21	39	1
64	AB000115_at	mRNA	281	250	214	103
65	AB000220_at	Semaphorin E	36	43	71	-61
66	AB000409_at	MNK1	-299	-103	-52	39
67	AB000449_at	VRK1	57	169	178	181

4.1.2 The Leukemia Database

We used the public Leukemia data set [http://www.broadinstitute.org/cgi-bin/cancer/publications/ pub_paper.cgi?mode%20=%20view&paper_id=43]. A sample from the data is shown in Table: 3. The Leukemia dataset contains of 7129 genes and 72 samples (47 the acute lymphoblastic leukemia (ALL) samples and 25 the acute myeloid leukemia (AML) samples). There are a total of 38 training samples and 34 testing samples. The 38 training samples contain 27 ALL and 11 AML, and the 34 Leukemia testing samples contain 20 ALL and 14 AML. The Leukemia dataset downloaded doesn't consist of noisy and inconsistent data. In addition, the description available for the Leukemia data set showed that, the only preprocessing task is normalization for its values to reduce the systemic bias introduced during experiments. Some authors normalized it [24] and some other research papers neglected the normalization step and proceed with the original data [25]. In our experiments, we followed the second trend.

4.2 Gene Ranking

Gene ranking simplifies gene expression tests to include only a very small number of genes rather than thousands of genes. We rank the genes by using the T-test based on their statistical score. Finding the informative genes greatly reduces the computational burden and noise arising from irrelevant genes [26]. The T-score of the genes are sorted and the genes with the highest T-scores are ranked.

4.2.1 The SRBCT Database

Table: 4 show gene ranking sample of the first most informative 20 genes. The genes with the highest scores are retained as informative genes. A comparison of our SRBCT ranked list of genes with a recent scientific paper[24], is shown in Table: 5. Where Genes (8), (12), (13), (15), (21), (23), (24), (25), (27) and (28) in our result are reported by Feng & Lipo [24] but in other order (Gene (4), (8), (2), (29), (21), (10), (13), (30), (6)and (3) respectively). We believe that the difference in genes order return to different probability of calculation.

Table 4. Informative genes based on their T-test for SRBCT dataset

No.	Gene ID	T-test Value	No.	Gene ID	T-test Value
1	812105	14.18978	11	814526	9.18717
2	236282	13.99938	12	325182	8.78664
3	183337	11.97544	13	784224	8.61800
4	745019	11.86478	14	283315	8.61102
5	767183	11.51372	15	241412	8.49859
6	624360	11.34247	16	383188	8.38719
7	1469292	10.95408	17	297392	8.35941
8	770394	10.0311	18	740604	8.35607
9	812105	9.48801	19	80109	8.31463
10	236282	9.43792	20	609663	8.02943

Table 5. A Comparison of our T-test result with another one for the SRBCT dataset

No.	Our T-test for all data	F. Chu and L. Wang [24]	No.	Our T-test for all data	F. Chu and L. Wang [24]
1	812105	810057	16	383188	44563
2	236282	**784224**	17	297392	866702
3	183337	**296448**	18	740604	21652
4	745019	**770394**	19	80109	814260
5	767183	207274	20	609663	298062
6	624360	**244618**	21	**629896**	**629896**
7	1469292	234468	22	786084	43733
8	**770394**	**325182**	23	**377461**	504791
9	812105	212542	24	**796258**	365826
10	236282	**377461**	25	**1435862**	1409509
11	814526	41591	26	68977	1456900
12	**325182**	898073	27	**244618**	1435003
13	**784224**	**796258**	28	**296448**	308231
14	283315	204545	29	193913	**241412**
15	**241412**	563673	30	395708	**1435862**

Table 6. Informative genes based on their T-test for Leukemia dataset

No	Gene ID	T-test Value	No	Gene ID	T-test Value
1	X95735_at	7.6034	11	M55150_at	5.4784
2	X17042_at	6.2784	12	M62762_at	5.3676
3	M23197_at	6.2510	13	U50136_rna1_at	5.2267
4	M84526_at	5.9593	14	X61587_at	5.1651
5	L09209_s_at	5.8696	15	X16546_at	5.1366
6	U46499_at	5.8176	16	M11147_at	5.0894
7	M27891_at	5.7960	17	M32304_s_at	5.0370
8	M16038_at	5.7672	18	X52056_at	4.8925
9	M22960_at	5.6135	19	D49950_at	4.8778
10	M63138_at	5.5949	20	M19507_at	4.6818

4.2.2 The Leukemia Database

Table: 6 show gene ranking sample of the first most informative 20 genes. The genes with the highest scores are retained as informative genes.

4.3 Applying the KNN classifier

4.3.1 The SRBCT Database

The original SRBCT data was already divided into training and testing sets [25] and this data selection for training and testing have been used by many authors without modifications [27], [28], [29]. Therefore, we accepted the same selection for ease of scientific comparison. Table: 7 and Figure: 1 shows the exact result for the testing classification accuracy. From Figure: 1 the classification accuracy is 100% using at least 26 highest ranked genes, but it starts to decrease as the number of the ranked genes decreases. The classification accuracy reaches 70% with only 5 genes. Table: 8 present a comparison of applying the KNN classifier on the SRBC datasets with other five classification techniques on the same data. Also, the table shows the necessary number of genes required for achieving the reported accuracy. From these results, we can conclude that [27, 29] techniques are advanced in reducing the number of genes and achieving the same classification accuracy. Therefore, other classification techniques will be applied in future to improve our results.

Table 7. The SRBCT testing classification accuracy

Top Gene Selected No.	Classification Accuracy for 20 Sample	Top Gene Selected No.	Classification Accuracy for 20 Sample
30	100%	10	80%
26	100%	9	85%
25	95%	8	85%
23	90%	7	70%
20	80%	6	70%
15	85%	5	70%

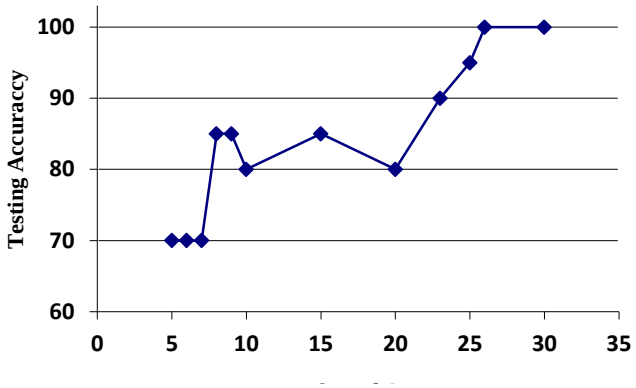

Figure 1. The Testing Classification Accuracy for the SRBCT Dataset

Table 8. Comparisons of results for the SRBCT dataset obtained by different approaches

Method	Accuracy	Number of genes required
MLP neural network [23]	100%	96
Nearest shrunken centroids [28]	100%	43
Evolutionary algorithm [27]	100%	12
SVM [29]	100%	20
FNN [11]	96%	3
Our KNN	100%	26

4.3.2 The Leukemia Database

The original Leukemia data was already divided into training and testing sets [30] and also, this data selection for training and testing have been used by many authors without modifications [25], [16], [31]. Table: 9 and Figure: 2 show the exact result for the testing classification accuracy. Table: 10 present a comparison of applying the KNN classifier on the Leukemia datasets with other three classification techniques on the same data. Also, the table shows the necessary number of genes required for achieving the reported accuracy. From these results, we can conclude that [25, 31] techniques are advanced in achieving higher classification accuracy but with greater number of genes. Also, we believe that applying the KNN classifier only achieved reasonable results in terms of cost and time and less number of genes. Figure 2 shows that the classification accuracy is best using 10 genes and start to decrease with decreasing the number of genes. Also, it is not a valuable in terms of classification accuracy if we increase the number of selected genes more than 30 ranked genes.

Table 9. The Leukemia testing classification accuracy

Top Gene Selected No.	Classification Accuracy for 34 Sample	Top Gene Selected No.	Classification Accuracy for 34 Sample
100	94.12%	15	97.06%
90	94.12%	10	97.06%
60	94.12%	7	94.12%
40	94.12%	5	94.12%
30	94.12%	3	94.12%
25	97.06%	2	94.12%
20	97.06%	1	91.18%

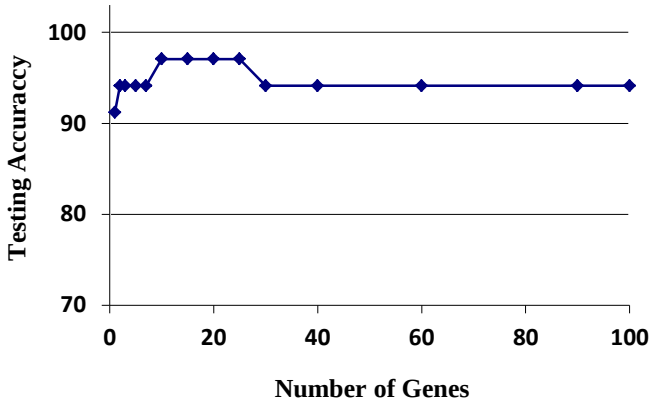

Figure 2. The Testing Classification Accuracy for the Leukemia Dataset

Table 10. Comparison of Leukemia classifiers

Authors	Accuracy	Number of genes required
Dina A. et, al. [25]	100%	100
Dina A. et, al. [16]	100%	5
D Mishra, B Sahu, [31]	98.1%	20
Our KNN	97.06%	10

4. Conclusion & Future Work

The recent technological advances, have led to an exponential increase in the biological data. This large amount of the biological data increases the using of machine learning techniques to generate a hidden knowledge from this biological data. Classifying the cancer datasets (microarray expression datasets) into a predefined class is divided into two main steps to reach accurate classification accuracy. The first step is implementing an effective gene selection technique to reduce the number of genes involved in the classification process to reduce time and effort and hence increase classifier efficiency and accuracy. The second step is adjusting a powerful classifier to achieve accurate classification accuracy for new unclassified samples.

The high dimensionality input and the small sample data size are the main two problems that have triggers the discovery process in microarray data. This paper presented the discovery of ranked list of genes using T-test statistical machine learning technique and applying a KNN classifier on two public cancer databases. The results of our ongoing gene expression research showed that the T-test reported a successful ranked list of genes compared to recent related work and reduced the data size, therefore improved the complexity of the analysis of data. Also, KNN achieved promising classification accuracy proportion to the supplemented ranked genes numbers. informative genes. Also, different classifier will be used. In addition to, other gene expression data sets.

References:

[1]Our future work intends to apply other machine learning techniques for the discovery of the T.D.Pham and D.I.Crane, " Analysis of microarray gene expression data" , Bentham science Ltd, 2006.

[2]A.Osareh and B.Shadgar, "Classification and diagnostic prediction of cancers using gene microarray data analysis", J.of applied sciences, vol.9, no.3, 2009, pp.459-468.

[3]S.Kim, Y.Tak and L.tari, "Mining gene expression profiles with bioogical prior knowledge ", Proc. of IEEE life science systems & applications workshope, 2006, pp.1-2.

[4]G.Tzanis, C.Berberidis, &I.Vlahavas, "Biological data mining", Encyclopedia of Database Technologies and Applications, 2006.

[5]R.Tibshirani, et.al, "Clustering methods for the analysis of DNA microarray data", Dep. of Statistics, Stanford University, 1999.

[6]G.Shapiro and P.Tamayo, "Microarray data mining: facing the challenges". SIGKDD Explorations, vol.5, no.2, 2003, pp.1-5.

[7]Robert Ruiz, Jose C.Riquelme and Jesus S.Aguilar Ruiz, "Incremental wrapper based gene selection from microarray data from cancer classification", pattern recognition, vol.39, no.12, 2006, pp. 2383-2392.

[8]J. Zhang, andH. Deng, "Gene selection for classification of microarray data based on the Bayes error", BMC Bioinformatics, vol.8, no.1, 2007, pp.370-378.

[9]F.Chu and L.Wang, "Applying Rbf neural networks to cancer classification based on gene expressions", Int. Joint Conf. on Neural Networks, 2006, pp.1930-1934.

[10] Zhang, et.al, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no.3, 2007, pp. 485 – 495.

[11] L.Wang, F.Chu and W.Xie, "Accurate cancer classification using expressions of very few genes", IEEE Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, 2007, pp.40-53.

[12] M.Rangasamy and S.Venketraman, "An efficient statistical model based classification algorithm for classifying cancer gene expression data with minimal gene subsets", Int. J. of Cyber Society & Education, vol. 2, no. 2, 2009, pp.51-66.

[13] A.Bharathi and A.M.Natarajan, "Cancer classification of bioinformatics data using ANOVA", Int. J. of Computer Theory & Engineering, vol. 2, no. 3, 2010, pp.1793-8201.

[14] A.Osareh and B.Shadgar, "Machine learning techniques to diagnose breast cancer", 5th Int. Symposium on Health Informatics and Bioinformatics (HIBIT) , 2010, pp.114 – 120.

[15] Z.Ghorai, et.al, "Cancer classification from gene expression data by nppc ensemble", IEEE Transactions on Computational Biology &Bioinformatics, vol.8, no.3, 2011, pp.659-671.

[16] Dina A. Salem, Rania Ahmed and Hesham A. Ali, "MGS-CM: A multiple scoring gene selection technique for cancer classification using microarrays", International J. of Computer Applications, vol.36, no.6, 2011, pp.0975 – 8887.

[17] L.Guyon and A.Elisseeff, "An introduction to variable and feature selection", J. of Machine Learning Research, vol.3, 2003, pp.1157-1182.

[18] Y.Wanga, et.al, "Gene selection from microarray data for cancer classification", Computational Biology and Chemistry, vol.29, no.1, 2005, pp.37-46.

[19] C.Lai, et.al, "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets", BMC Bioinformatics, vol. 7, 2006, pp.235-244.

[20] Abeer M. Mohamed, et. al, "Analysis of machine learning techniques for gene selection and classification of microarray data", Proceeding of 6th IEEE int. conf. on Information Technology, Cloud Computing, 2013.

[21] L.Deng, J.Ma and D.Lee, "A rank sum test method for informative gene discovery", In Proceedings of 10th int conf. of ACM SIGKDD'04, 2004, pp.410-419.

[22] X. Wu et al., "Top 10 algorithms in data mining", KnowlInfSyst, vol. 14, 2008, pp. 1-37.

[23] J.M. Khan et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, vol. 7, 2001, pp. 673-679.

[24] F. Chu and L. Wang,"Applications of support vector machines to cancer classification with microarray data", International Journal of Neural Systems, Vol. 15, No. 6, 2005, pp. 475–484.

[25] Dina A. Salem, Rania Ahmed &Hesham A. Ali, "DMCA: A combined data mining technique for improving the microarray data classification accuracy", Int. Conf. on Environment and BioScience, IPCBEE vol.21, 2011, pp. 36-41.

[26] V.Bhuvaneswari and .Vanitha,"Classification of microarray gene expression data by gene combinations using fuzzy logic (MGC-FL)", International Journal of Computer Science, Engineering and Applications (IJCSEA), Vol.2, No.4, 2012.

[27] J. Deutsch, "Evolutionary algorithms for finding optimal gene sets in microarray prediction," Bioinformatics, vol. 19, 2003, pp. 45-52.

[28] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," Proc. Nat'l Academy of Sciences USA, vol. 99, 2002, pp. 6567-6572.

[29] Y. Lee and C.K. Lee, "Classification of multiple cancer types by mulitcategory support vector machines using gene expression data," Bioinformatics, vol. 19, 2003, pp. 1132-1139.

[30] T. Golub et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, vol.286, 1999, pp. 531–537.

[31] D Mishra, B Sahu, "Feature selection for cancer classification: a signal-to-noise ratio approach", International Journal of Scientific & Engineering Research, Vol.2, 2011.