International Journal of Intelligent Computing and Information Science

METHODOLOGY FOR SELECTING MICROARRAY BIOMARKER GENES FOR CANCER CLASSIFICATION

E. M-F. El Houby N. I. . Yassin

Engineering Division, Systems & Information Department, National Research Centre, El Buhouth Street, Dokki, Cairo, Egypt

em.fahmy@nrc.sci.eg eng_nesrin@hotmail.com

Abstract: *In the analysis of microarray gene expression data, it is very difficult to obtain a satisfactory classification result by machine learning techniques because of the dimensionality problem. That is the gene expression data are very high dimensional, while datasets usually contain a few tens samples. Microarray data includes many redundant, noisy genes and numerous genes contain inappropriate information for classification.The best combination of gene selection and classification is required to identify biomarker genesfrom thousands of genes. In this research, a methodology has been developed to eliminate noisy, irrelevant and redundant genes and find a small setof significant informative biomarker genes which can classify cancer dataset with high accuracy. The process consists of two phases which are gene selection and classification. In gene selection phase, the genes have been ranked according to their ranking scores; two statistical approaches which are class separability and T-test have been used. Then from the highest ranked genes, different subsets of genes have been used to classify dataset until reach the highest possible accuracy. Two data mining techniques have been used for classifications which are K-Nearest Neighbor and Support Vector Machine. The proposed method has been used to classify 7 benchmarkgene expression cancer datasets. The results showed that the proposed methodology can identifysmall subsetof relevant predictive genes and can achieve high prediction accuracy with this small subset of genes for different datasets.The accuracyand subset of biomarker genes have been identified for different cancer datasets.*

Keywords: *Gene Selection, Support Vector Machines, K-Nearest neighbor, Microarray Gene expression, Class-separability, T-test.*

1. Introduction

Microarray is a technology in the modern biological research to analyze the expression of genes. Microarray techniques provide a platform that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern where one can measure the expression levels of thousands of genes in hundreds of different conditions simultaneously [1]. Therefore microarray studies enable clinicians and biologists to obtain the gene expression profile of a given tissue sample rapidly and compare it with other samples [2].

Microarray studies are used to discover which specific genes are important to the development of a disease. They are used to analyze gene expression data associated with a specific diagnosis. For

example, the study of expression profiles between microarray samples from cancer patients and normal subjects, allowing these genes to be classified based on differences in expression levels [3]. Also, sometimes it is extremely difficult to find clear distinctions between some types of cancers according to their appearances. Hence the microarray technology stands to provide a more quantitative means for cancer diagnosis[4]. Computational analysis and computing can help researchers to collate a group of signature genes for a certain disease [5,2].

However, there are some major technical difficulties or problems that confront researchers in this area. For example, genetic variability affects gene expression. That is, the expression levels of two patients with the same disease may differ significantly [6]. Additionally, there are many noise factors that affect microarray gene expression datasets and how to filter out noise is a thorny problem that must be solved. Actually, there is a high redundancy in microarray data and numerous genes contain inappropriate information for precise classification of diseases or phenotypes. Therefore, the amount of data generated by this technology presents a challenge for the biologists to carry out analysis [1].

It is a challenge to use gene expression data for cancer classification because gene expression data are usually very high dimensional. The dimensionality ranges from several thousands to over ten thousands. Owing to the high price of microarray chips and a lack of tissues from patients, so gene expression data sets usually contain relatively small numbers of samples, e.g., a few tens. These datasets are usually too few in number to use machine learning. In addition, the processing and material used for microarray analysis differ between manufacturers and so it is difficult to identify a unique set of genes that can form an integrated dataset. To obtain good classification accuracy, the genes that benefit the classification most, should be picked out. In addition, gene selection is also a procedure of input dimension reduction, which leads to a much less computation load to the classifier [4, 7]. Therefore, gene selection becomes the most necessary prerequisite for a diagnostic classification system. How to choose a small and discriminative subset of genes from among tens of thousands of genes to solve the dimensionality problem is very difficult. However, the best combination of classification and gene selection is understood poorly, because there is another methodological trouble associated with training microarray data. This is the problem of ‘‘over-fitting”. Over-fitting means that one can obtain good performance using a training set, but when new data is used, a satisfactory result cannot be obtained using the trained model. This occurs often when there are a small number of high-dimension samples [8].

In this research, a methodology for selecting biomarker genes for cancer classification has been developed to reach the least possible number of biomarker genes that can be used to diagnosis different type of cancers with highest possible performance. The remainder of the paper is organized as follows. An overview for the previous work related to our subject is presented in section 2;materials and methods are described in section 3; testing the system and the experimental results are conducted in section 4, before drawing conclusions and future work in section 5.

2. Related Work

A variety of gene selection and classification techniques have been proposed in the literature. Li et al. [9] devised a method of combining particle swarm optimization (PSO) with a genetic algorithm (GA)as the classifier for gene selection.Mallika&Saravanan [10] developed a new algorithm called an efficient statistical model based classification algorithm for classifying cancer gene expression data with minimal gene subsets. Classical statistical technique is used for the purpose of ranking the gene and two various classifiers are used for gene selection and prediction. Park et al [11] presented a method for inferring

combinatorial Boolean rules of gene sets for cancer classification.A gene selection scheme called ANOVA was presented by Bharathi&Natarajan [12], which is used to find the minimum number of genes from microarray gene expression for cancer classification. The support vector machine(SVM)was used for the classification process. Zhao et al. [13] presented a novel hybrid framework (NHF) for gene selection and cancer classification of high dimensional microarray data by combining the information gain (IG), F-score, GA, PSO, and SVM. Dina et.al, [14] introduced three hybrid classification systems called (MGS-SVM, MGS-KNN and MGS-LDA), respectively. They also, proposed a gene selection technique named (MGS-CM). Using their methods they achieved reasonable classification accuracy but on limited datasets. The SVM was used as a classifier for microarray genes by Nanni et al [15].Their method combined different feature reduction approaches to improve classification performance of the accuracy and area under the receiver operating characteristic (ROC).Chen et al. [16] used PSO and 1-nearest neighbor (1NN) for feature selection and tested their algorithm against 8 benchmark datasets.Abeer M.Mahmoud, et al. [17] applied machine learning approach to classify two public microarray datasets. The genes were ranked according to their statistical scores using T-test and the highest informative genes are selected for classification using k-nearest neighbor.A novel method utilizing PSO combined with a decision tree as a classifier was developed by Chen et al [18]. They compared the performance of the proposed method with other well-known benchmark classification methods.

3. Material and Method

3.1 Material

Seven public microarray cancer datasets with different characteristics are used for the analysis of the proposed methodology including gene selection and classification techniques.The description of these datasets isshown in table1. These datasets have been obtained from the GEMS website (www.gems-system.org)

Table1: Characteristics of the 7 used public microarray datasets

Dataset name	Diagnostic task	No. of Samples	No. of Genes	Diagnostic categories
SRBCT	Small, round blue cell tumors (SRBCT) of childhood	83	2308	4
DLBCL	Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas	77	5469	2
Leukemia1	Acute myelogenous leukemia (AML), acute lympboblastic leukemia (ALL) B-cell, and ALL T-cell	72	5327	3
Leukemia2	AML, ALL, and mixed-lineage leukemia (MLL)	72	11225	3
Prostate Tumor	Prostate tumor and normal tissues	102	10509	2
Brain Tumor1	5 human brain tumor types	90	5920	5
Lung	4 lung cancer types and normal tissues	203	12600	5

3.2 The Proposed Methodology

In this research, a methodology has been developed to find the least number of biomarker genes that can be used to diagnosis different type of cancers with the highest possible accuracy. The proposed methodologyconsists of two phases which are (1) gene selection phase which uses different statistical approaches to rank genes and select a set of the highest ranked genes which are the most informative genes for classification(2) classification phase to classify different cancer datasets using subset of the selected highest ranked genes by applying different data mining techniques. The methodology aims to study the effect of applying different gene selection approaches prior to classification on the performance.
To fulfill this purpose, genes have been ranked according to their ranking scores which can be measured using gene selection statistical approach. A certain percentage of the highest ranked genes has been selected for classification by dividing dataset by (α), where (α) is a constant specified by the user. This set of selected genes is introduced to the classifier one by one. First, the highest ranked gene is used to classify datasetand the accuracy is measured. Then, the next ranked gene is added to the set of genes which are used to classify dataset, if its effect is positive, i.e. the accuracy of the classifier is improved then this gene is added to the list of biomarker genes.Otherwise it should be ignored,either because it has negative effect by reducing the classification accuracy, or it is redundant by keeping the accuracy constant. The process continues and a list of effective biomarker genes is formed by comparing the measured accuracies until reaching the highest possible accuracy or using the selected set of genes.Figure1 depicts the methodology's steps to find the least number of genes that achieve the best accuracy for microarray classification.In the next subsections gene selection phase and classification phase will be introduced.

3.2.1 Gene Selection

Among the large number of genes, only a small part may benefit the correct classification of cancers. The rest of the genes have little impact on the classification. Even worse, some genes may act as "noise" and undermine the classification accuracy. Hence, to obtain good classification accuracy, the genes that benefit the classification have been picked out. Reducing the number of genes used for classification can help researchers put more attention on these important genes and find the relationship between those genes and the development of the cancers [4].

The gene selection method can be divided into three categories, the wrapper, the filter, and the embedded. Wrappers utilize learning machine to search for the best genes in the datasets of all genes subsets. Wrappers highly depend on the learning model and may suffer from excessive computational complexity. The filter method usually employs statistical methods to collect the intrinsic characteristics of genes in discriminating the targeted phenotype class. Filter approaches are individual feature ranking methods. They are easily implemented, but ignore the complex interaction among genes. Finally, the embedded method is similar to the wrapper method, while multiple algorithms can be combined in the embedded method to perform feature subset selection [19, 20].

Filter method is the adopted gene selection in this research. Filter approaches are characterized 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 [14, 21, 22]. Many of filter gene selection approaches are developed to reduce the

number of genes in the microarray datasets to reach accurate classification accuracy with the smallest number of genes. They also reduce the computational time and the cost of the classification [17]. Class Separability (CS) and T-test (TS) are two gene selection approaches widely applied for microarray data, and they are the selected approaches to be applied in this research.

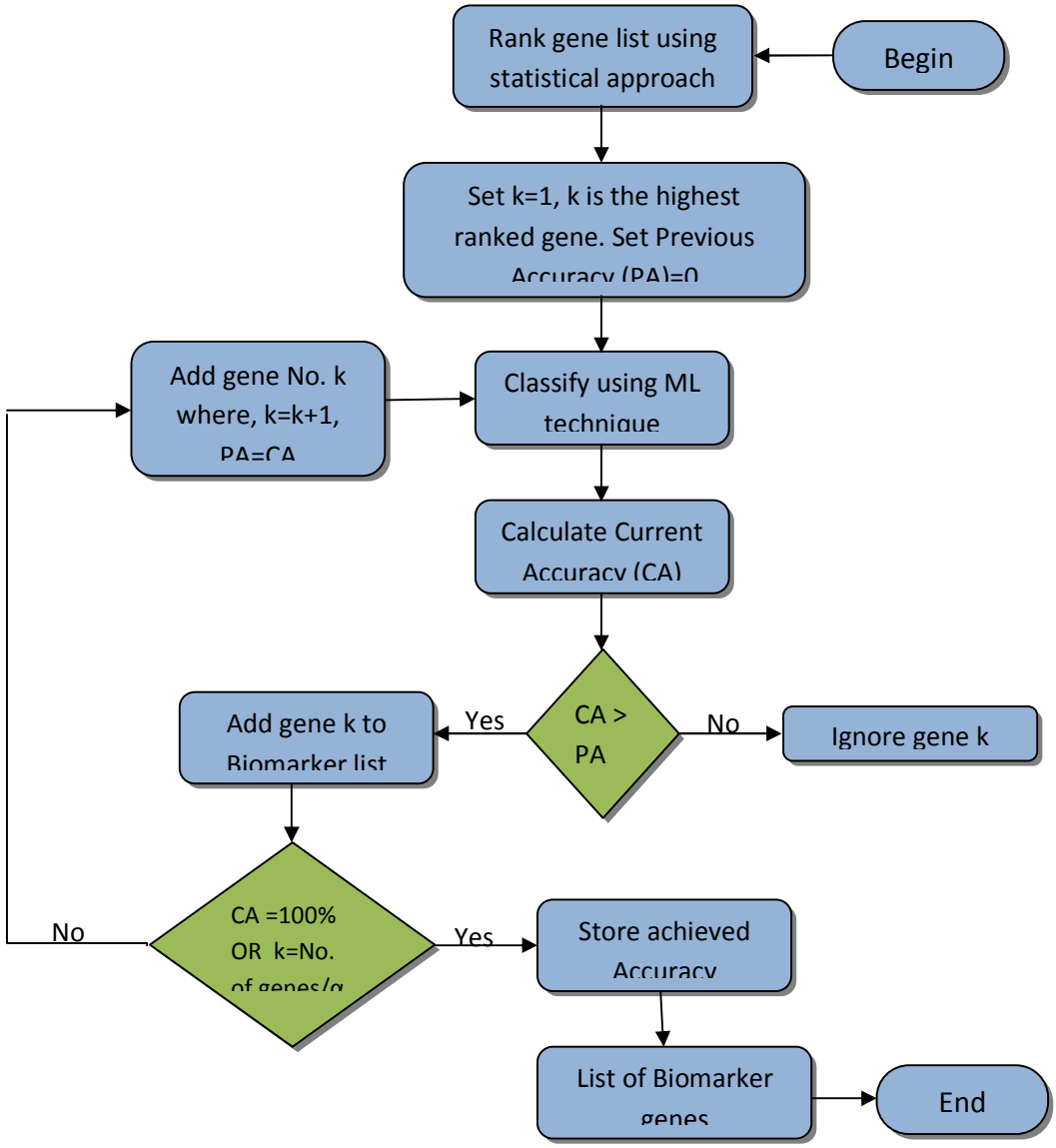

Figure1: Different steps to find the least number of genes for microarray classification

Class-SeparabilityApproach

Class-separability(*CS*) [23] is anapproach used for gene selection. *CS* of gene *i* is defined as:

$$CS_i = SB_i / SW_i (1)$$

Where

$$SB_i = \sum_{i=1}^{K} (\bar{x}_{ik} - \bar{x}_i)^2 (2)$$

$$SW_i = \sum_{k=1}^{K} \sum_{j \in C_k} \left(x_{ij} - \bar{x}_{ik}\right)^2 (3)$$

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

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

$SB_i$is the sum of squares of betweenclass distances (the distances between samples of different classes). $SW_i$is the sum of squares of with-in class distances (the distances of samples within the same class). In the whole data set, there are Kclasses. $C_k$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. A $CS$is calculated for each gene. A larger CSindicates a larger ratio of the distances between different classes to the distances within one specific class. Therefore, CSscan be used to measure the capability of genes to separate different classes [4].

T-TestBased Approach

T-test is a statistical approach proposed by Welch [24]. It is used to measure how large the difference is between the distributions of two groups of samples. For a specific gene, if it shows larger distinctions between 2 groups, it is more important for the classification of the two groups.
To select important genes using T-test a score based on T-test (named T-score or TS) is calculated for each gene. Then, all the genes are rearranged according to their TSs. The gene with the largest TS is put in the first place of the ranking list, followed by the gene with the second largest TS, and so on. In multi-class problems, T-test is 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 like this:

$$TS_i = max\{\left| \frac{\bar{x}_{ik} - \bar{x}_i}{m_k s_i} \right|, k = 1,2,\dots,K\} (6) \qquad s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} \left(x_{ij} - \bar{x}_{ik}\right)^2 (7)$$

$$m_k = \sqrt{1/n_k + 1/n} (8)$$

HereTS_itakes the maximum of all calculated values for $\{k = 1,2,\dots,K\}$. $s_i$is the pooled within class standard deviation for gene i [4].

3.2.2 Classification

After ranking genes using CS and T-test, theset ofgenes which represent the highest ranked genes have been classified. Gene expression classification is the process of classifying gene expression sample into a predefined class. Support vector Machine (SVM) and K-Nearest Neighbor (KNN)are two important

classification techniques for microarray data. In this research these two techniques have been used for classifying the selected gene sets and reducing them as possible.

Support Vector Machine (SVM)

SVM classification technique is one of the most powerful machine learning classifiers which is based on the statistical learning theory [25]. SVM is used widely to classify gene expression data. This approach uses the kernel trick to deal with nonlinearly separable data. SVM maps the initial data to a higher dimensional space, using a proper kernel function, in which the data are linearly separable. The kernel function that has been used is a polynomial:

$$K(X, X_i) = (X^T X_i + 1)^p \quad (9)$$

Wherep is a constant specified by users.

K-Nearest Neighbor (KNN)

KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It is a lazy algorithm that it depends on calculating a distance between a test data and all the training data. It decides in which class the instance belongs to by using a majority of the chosen K of neighbors. Although being a simple technique, KNN shows a high performance in classifying microarray gene expression. The KNN calculates its distances by different ways, but Euclidean distance is the most popular[26].

As mentioned, the Euclidean distance is used in the k-nearest-neighbor to calculate the distance between a test sample and the specified training samples. Let x_i be an input sample with *p*gene expression values for different genes (x_{i1},x_{i2},…,x_{ip}) ,*n* be the total number of input samples (*i*=1,2,…,*n*) and *p* the total number of genes (*j*=1,2,…,*p*), $x_{ij}$is the expression value in sample*i*for gene*j* . The Euclidean distance between sample x_i and x_l(*l*=1,2,…,*n*) is defined as

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + .. + (x_{ip} - x_{lp})^2} \quad (10)$$

The pseudo code depicted in algorithm1 sums up the steps for selecting the least number of biomarker genesfor classifying different cancer datasets with the highest possible accuracy. As it is shown in algorithm1, in the first step the genes are ranked using gene selection approach. In step2,the first highest ranked gene datais added to the gene data list and its ranking order is added to the list of biomarker genes in step3. In step 4, this gene is used to build classifier, and its accuracy is calculated in step 5. In step 6, the second ranked gene is selected and its data is added to the gene data list (first ranked gene in this case) in step 8, the classifier is build using the list of added genes data in step 9 and the accuracy of the classifier is calculated in step 10.The achieved accuracy is compared with the current accuracy in step 11, if the achieved accuracy is greater than the current accuracy, the ranking number of the added gene is added to the list of biomarker genes in step 12, and the achieved accuracy is set as current accuracy in step 13. Otherwise the added gene is neglected in step 15. Then the next ranked gene is tried and so on the processcontinues while achieved accuracy<100% and the tried genes are less than the selected no of genes (step 7-18). The biomarker genes list has got in step 19and the reached accuracy has got in step 20.

```
1. Rankgenes using gene selection approach
2. Gene_array_data_list[1] = the top ranked gene data
3. Gene_no_list[] = [1]
4. Build a classifier using Gene_array_data_list[1]
5. Calculate accuracy, set its value as current_accuracy
6. Gene_no = 2
7. Do While ((accuracy<100%) and (Gene_no<= TotalGene_no/α))
8.        Add Gene_data[Gene_no] to Gene_array_data_list[]
9.        Build classifier using Gene_array_data_list[]
10.       Calculate accuracy, set its value as accuracy [i]
11.       If accuracy [i] >current_accuracy
12.       Add Gene_no to Gene_no_list[]
13.       current_accuracy = accuracy [i]
14.       Else
15.               RemoveGene_data[Gene_no] fromGene_array_data_list[]
```

Algorithm1: selecting the least number of genes for classifying different cancer datasets with the highest accuracy

4. Experimental Result

This section shows an empirical performance evaluation of the proposed methodology. Extensive experimental studies had been tried in order to test the methodology. Cross-validation has been used to evaluate and compare different results; 10-fold cross validation has been used for estimating the accuracy.Seven public microarray cancer datasets which have been mentioned before have been used for the analysis of the proposed methodology.Among the seven used public datasets, SRBCT is a common dataset used in previously published literatures that contain the results including the required number of genes so it will be easy to verify the proposed algorithm. So SRBCT will be tackled in some details to clarify the different phases of the used methodology. Then the results of the remaining datasets will be illustrated.

The two statistical approaches T-test and class-separability have been applied to SRBCT to rank genes.Table 2 shows genes ranking sample of the first most informative 30 genes of SRBCT dataset using the T-test and the corresponding ranking ordersof the same genes using class-separability.The table containsgene ID, the ranking values and the ranking orders using T-test and class-separability.Then sets of the highest ranked genes have been selected from the 2 different ranked lists for classification, supposing that $\alpha = 70$.

Two machine learning techniques which are SVM and KNN have been applied to the selected sets of the highest ranked genes using the proposed methodology to classify SRBCT dataset by least possible number of informative genes and highest possible accuracy. By applying the proposed methodology, the positive effect genes are considered and added to the biomarker genes list while the negative effect and redundant genes have been neglected. The process continues until reaching 100% accuracy or trying the selected number of genes.The results showed that 10 biomarker genes using T-test and 9 genes using CS have been required to classify SRBCT dataset using SVM to reach 100% accuracy. Whileto classify the SRBCT dataset using KNN, 14 genes are required from the list ordered by T-test and the reached accuracy is 98.7952 %, and it needs 12 genes from the list ranked by CS to achieve 100% accuracy. Table 3 shows the list of biomarker genes' ids and orders for SRBCT which is required to classify it by SVM and KNN using the two used ranking approaches.AlsoFig. 2 shows the accuracy versus number of genes for SBRCT using SVM& KNN with T-test& CS. It is shown that highest accuracy with the least number of genes can be achieved using SVM with CS it reaches 100% accuracy by 9 genes. Since the results of SRBCT including the required number of genes are available in the literatures, it will be used in the comparison of the proposed algorithm as shown in table 4, which presents a comparison of the

proposed methodology resultsfor the SRBCT datasets with 2scientific papers [17] that used KNN and [27] that used SVM to classify the same dataset. Also, the table shows the necessary number of genes required for achieving the reported accuracy.

Table2: A comparison between T-test and class-separability ranking order for most informative genes of SRBCTdataset

No.	Gene ID	Gene Description	T-TestOrder	T-Test value	CS Order	CS value
1	812105	transmembrane protein	1	14.18978	2	0.22540
2	236282	Wiskott-Aldrich syndrome (ecezema-thrombocytopenia)	2	13.99938	1	0.27998
3	183337	major histocompatibility complex, class II, DM alpha	3	11.97544	3	0.20323
4	745019	EH domain containing	4	11.86478	4	0.19662
5	767183	hematopoietic cell-specific Lyn substrate 1	5	11.51372	5	0.18770
6	624360	proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional protease 7)	6	11.34247	6	0.18348
7	146922	pim-2 oncogene	7	10.95408	7	0.17078
8	770394	Fc fragment of IgG, receptor, transporter, alpha	8	10.03110	9	0.11961
9	814526	ESTs	9	9.488010	8	0.14192
10	325182	cadherin 2, N-cadherin (neuronal)	10	9.437927	17	0.09655
11	784224	fibroblast growth factor receptor 4	11	9.187176	18	0.08654
12	283315	phosphoglyceratemutase 2 (muscle)	12	8.786647	10	0.11268
13	241412	E74-like factor 1 (ets domain transcription factor)	13	8.618004	11	0.10518
14	383188	recoverin	14	8.611020	13	0.10307
15	297392	Metallothionein 1L	15	8.498598	14	0.10221
16	740604	interferon stimulated gene (20kD)	16	8.387190	16	0.09853
17	80109	major histocompatibility complex, class II, DQ alpha 1	17	8.359410	15	0.09856
18	609663	protein kinase, cAMP-dependent, regulatory, type II, beta	18	8.356070	12	0.10418
19	629896	microtubule-associated protein 1B	19	8.314639	20	0.08065
20	786084	chromobox homolog 1 (Drosophila HP1 beta)	20	8.029430	29	0.06889
21	377461	caveolin 1, caveolae protein, 22kD	21	7.959241	27	0.07212
22	796258	sarcoglycan, alpha (50kD dystrophin-associated glycoprotein)	22	7.889858	32	0.063296
23	1435862	antigen identified by monoclonal antibodies 12E7, F21 and O13	23	7.887555	22	0.07854
24	68977	proteasome (prosome, macropain) subunit, beta type, 10	24	7.717375	19	0.08341
25	244618	ESTs	25	7.502730	39	0.059212
26	296448	insulin-like growth factor 2 (somatomedin A)	26	7.357450	40	0.057568
27	193913	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog	27	7.225395	23	0.077996
28	395708		28	7.210156	35	0.062181
29	626502	actin related protein 2/3 complex, subunit 1B (41 kD)	29	7.194638	26	0.07272
30	782811	high-mobility group (nonhistone chromosomal) protein isoforms I and Y	30	7.099770	21	0.07965

The same algorithm has been applied for the remaining 6 datasets. Table 5 shows the achieved accuracy for different datasets and the required number of genes using T-test &CS with SVM & KNN. Table 6 shows list of biomarker genes for all used dataset using TS & CS with SVM. Table7 shows list of biomarker genes for all used dataset using T-test& CS with KNN. Fig. 3 shows the comparison between accuracy using SVM with T-test and CS for different datasets. Fig. 4 shows the comparison between the required number of genes using SVM with Ts and CS for different datasets. Fig. 5 shows the comparison between accuracy using KNN with Ts and CS for different datasets. Fig. 6 shows the comparison between the required number of genes using KNN with Ts and CS for different datasets.

As shown from the results, to achieve almost the same accuracy using SVM or KNN with CS and T-test, the required numbers of genes may be varied for the same dataset, and so the set of biomarker genes. Also it is noticeable that the ranked list order for either DLBCL or prostate tumor are identical using CS and T-test and so the biomarker genes using the same data mining technique (SVM or KNN),

that may be because eachof these datasets have 2 diagnostic categories and so the separation between these 2 categories is clear, so there is no difference between gene selection approaches.

Table3: List of biomarker genes ID& orders to classify SRBCT by SVM and KNN using T-test and CS

SVM				KNN			
TS		CS		TS		CS	
Gene order	Gene ID	Gene order	Gene ID	Gene order	Gene ID	Gene order	Gene ID
1	812105	1	236282	1	812105	1	236282
2	236282	2	812105	2	236282	2	812105
3	183337	3	183337	3	183337	3	183337
8	770394	9	770394	4	745019	4	745019
10	325182	10	283315	6	624360	6	624360
11	784224	17	325182	7	146922	7	146922
21	377461	22	1435862	8	770394	9	770394
22	796258	27	377461	9	814526	10	283315
23	1435862	32	796258	10	325182	12	609663
25	244618			11	784224	13	383188
				12	283315	32	796258
				14	383188	39	244618
				22	796258		
				23	1435862		

Table 4: Comparison of the proposed methodology results for SRBCT dataset with others.

Method	**Accuracy**	**Number of required genes**
SVM[27]	100%	20
KNN [17]	100%	26
The proposed TS-SVM	100%	10
The proposed CS-SVM	100%	9
The proposed TS-KNN	98.7952 %	14
The proposed CS-KNN	100%	12

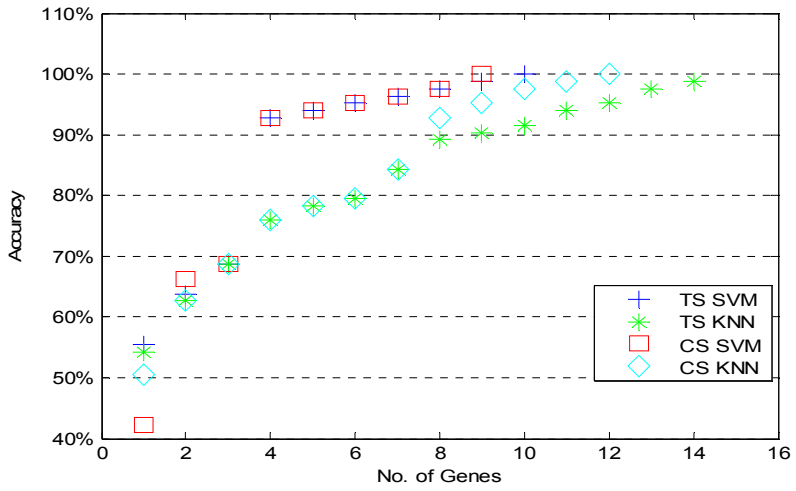

Figure 2: The comparison of accuracy versus number of genes for SBRCT usingcombinations of T-test& CS with SVM& KNN

Table5: The achieved accuracy for different datasets and the required number of genes using T-test &CS with SVM&KNN

Data Set Name	SVM				KNN			
	Accuracy		No. of Genes		Accuracy		No. of Genes	
	TS	CS	TS	CS	TS	CS	TS	CS
SRBCT	100 %	100 %	10	9	98.7952 %	100 %	14	12
DLBCL	97.4026 %	97.4026 %	6	6	98.7013 %	98.7013 %	8	8
Leukemia1	95.8333 %	95.8333 %	8	6	97.2222 %	97.2222 %	6	9
Leukemia2	95.8333 %	95.8333 %	6	6	95.8333 %	95.8333 %	8	6
Prostate Tumor	95.098 %	95.098 %	6	6	95.098 %	95.098 %	6	6
Brain Tumor1	91.1111 %	86.6667 %	9	6	88.8889 %	87.7778 %	10	11
Lung Cancer	87.6847 %	86.6995 %	8	7	92.1182 %	90.1478 %	11	13

Table6: List of biomarker genes for all used dataset using TS & CS with SVM

Dataset	**T-test**	**CS**
DLBCL	1+2+3+5+8+50	1+2+3+5+8+50
Leukemia1	1+7+13+19+27+28+31+43	1+7+22+25+28+30
Leukemia2	1+2+3+4+5+7	1+2+3+5+6+7
Prostate Tumor	1+2+4+5+6+12	1+2+4+5+6+12
Brain Tumor1	1+2+15+18+20+24+25+31+34	1+2+24+47+65+95
Lung Cancer	1+2+20+30+36+44+46+47	1+2+3+4+7+9+37

Table 7: List of biomarker genes for all used dataset using T-test& CS with KNN

Dataset	**T-test**	**CS**
DLBCL	1+2+3+5+8+10+13+39	1+2+3+5+8+10+13+39
Leukemia1	1+2+3+10+13+14	1+2+3+10+16+22+28+29+30
Leukemia2	1+2+3+5+6+13+29+30	1+3+6+9+10+17
Prostate Tumor	1+2+3+6+12+14	1+2+3+6+12+14
Brain Tumor1	1+2+5+13+16+18+19+20+56+91	1+2+5+13+15+24+26+28+93+95+96
Lung Cancer	1+2+3+4+6+9+30+31+43+47+48	1+2+3+4+15+23+26+30+33+37+38+43+57

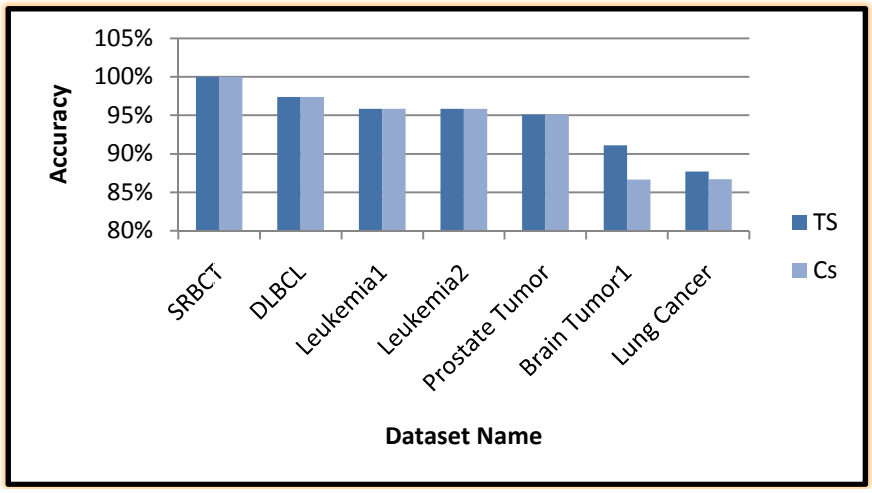

Figure 3: The comparison betweenaccuraciesusingSVM with T-test and CS for different datasets

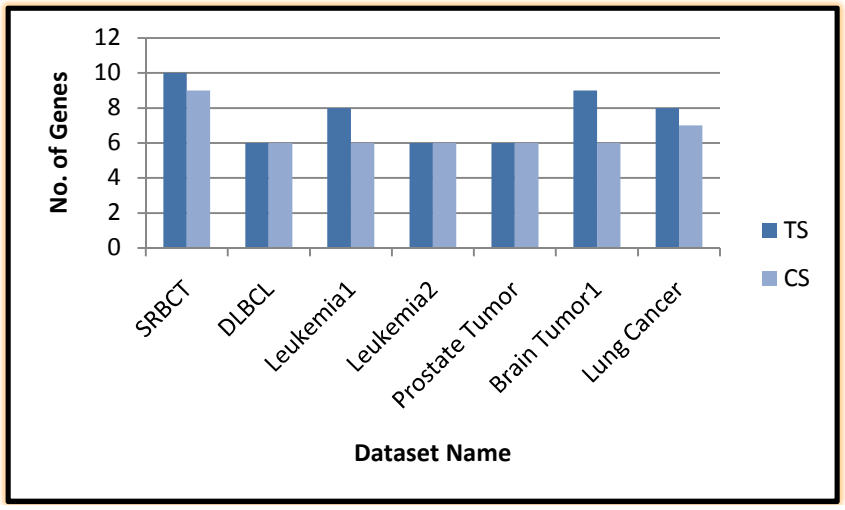

Figure4: The comparison between the required no. of genes using SVM with Ts and CS for different datasets

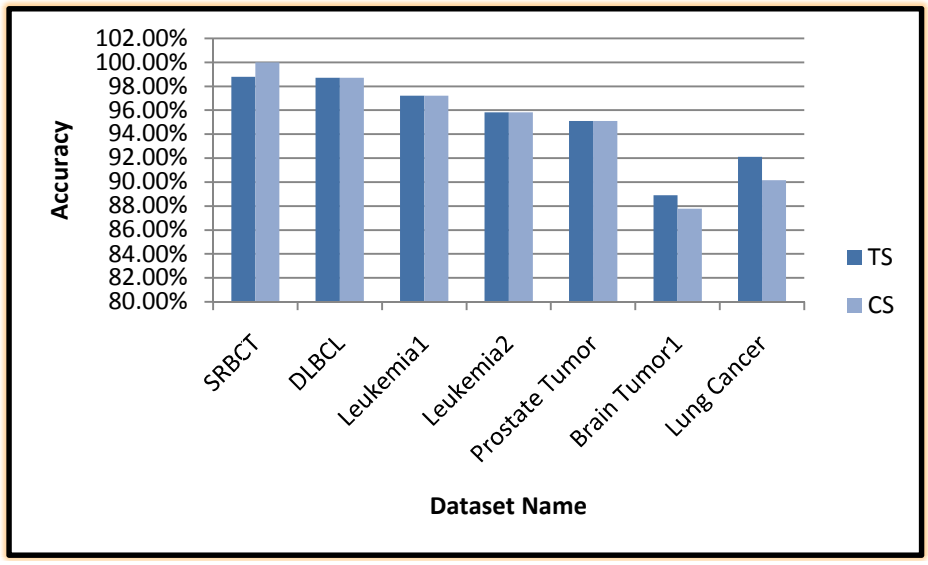

Figure 5: The comparison betweenaccuracies using KNNwith Ts and CS for different datasets

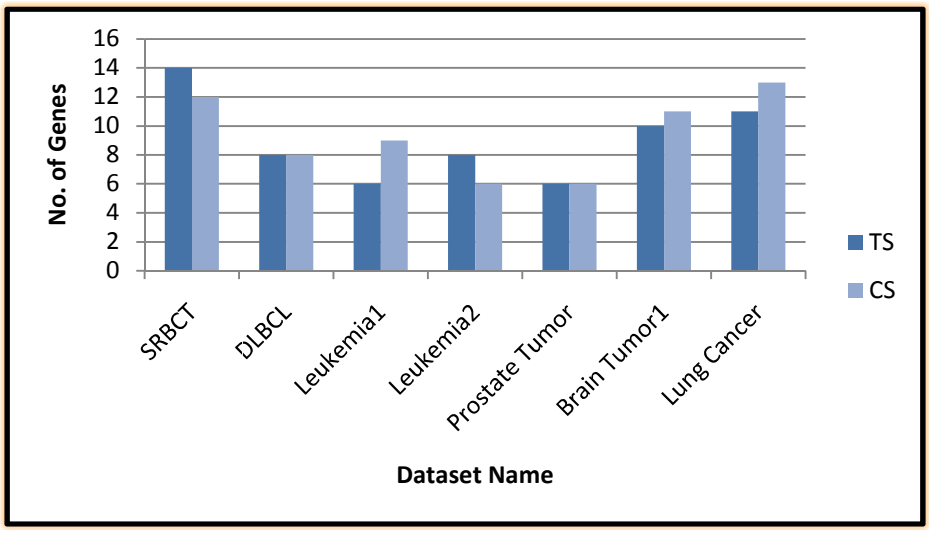

Figure6:The comparison between the required no. of genes for different datasets using KNN with Ts and CS

5. Conclusions

In this research a methodology have been developed to find the least number of the most informative genes for classifying different cancer microarrays. Different gene selection approaches which are class separability and T-test have been used to rank genes. Then from the highest ranked genes, the most informative biomarker genes have been used to classify various datasets using KNN and SVM. Different cancer microarray datasets have been classified using the proposed methodology. The accuracyfor different datasets has been measured and the numbers of biomarker genes which used to achieve these accuracies have been identified.

In the future we hope that real data will be available to try the proposed method. More gene selection approaches can be tried and more data mining techniques or combinations for more than one technique can be tried to achieve the highest possible accuracy with least number of informative biomarker genes.

References

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

2.Yuhang Wang, Fillia S. Makedon, James C. Ford and Justin Pearlman, "HykGene: A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data", Bioinformatics, 21(8), pp.1530–1537, 2005.

3.Qiu, P., Wang, Z. J.& Liu, K. J., "Ensemble dependence model for classification and prediction of cancer and normal gene expression data. Bioinformatics, 21(14), 3114–3121, 2005.

4.Fengchu&lipowang, "Applications of support vector machines to cancer classification with microarray data", International Journal of Neural Systems, Vol. 15, No. 6 (2005) 475–484

5.Buturovic, L. J. PCP: A program for supervised classification of gene expression profiles, Bioinformatics, 22(2), 245–247, 2006.

6.Li, T., Zhang, C., &Ogihara, M.,"A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression". Bioinformatics, 20(15), 2429–2437, 2004.

7.LiatEin-Dor, ItaiKela, Gad Getz, David Givol and EytanDomany, "Outcome signature genes in breast cancer: Is there a unique set?", Bioinformatics, 21(2), pp.171–178, 2005.

8.Alexander Statnikov, Constantin F. Aliferis, IoannisTsamardinos, Douglas Hardin and Shawn Levy, "A comprehensive evaluation of multi category classification methods for microarray gene expression cancer diagnosis", Bioinformatics, 21(5), pp. 631–643, 2005.

9.Li S, Wu X, Tan M,"Gene selection using hybrid particle swarm optimization and genetic algorithm", Soft Compute 2008, 12:1039–1048.

10.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, pp.51-66, 2009.

11. Park I, Lee KH, Lee D: Inference of combinatorial Boolean rules of synergistic gene sets from cancer microarray datasets. Bioinformatics 2010, 26:1506–1512.

12.A.Bharathi and A.M.Natarajan, "Cancer classification of bioinformatics data using ANOVA", Int. J. of Computer Theory & Engineering, vol. 2, no. 3, pp. 369-373, 2010.

13. Wei Zhao, Gang Wang, Hong-bin Wang, Hui-ling Chen, Hao Dong, Zheng-dong Zhao, "A Novel Framework for Gene Selection", International Journal of Advancements in Computing Technology, Volume 3, Number 3, pp. 184–191, April 2011

14. 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, pp.0975 – 8887, 2011.

15.Nanni L, Brahnam S, Lumini A,"Combining multiple approaches for gene microarray classification", Bioinformatics, 12Vol. 28 no. 8, 2012, pages 1151–1157.

16. Li-Fei Chen, Chao-Ton Su, Kun-Huang Chen, Pa-Chun Wang,"Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis", Neural Computing & Applications, 2012, 21(8):2087–2096.

17.AbeerM.Mahmoud, BasmaA.Maher, EL-Sayed M.EL-Horbaty and Abd EL-Badeeh M. Salem , "Machine Learning Approaches for Cancer Classification of Microarray Data", The Sixth

International Conference on Intelligent Computing and Information Systems (ICICIS 2013), Dec. 14-16, 2013, Ain Shams University, Cairo, Egypt.
18. Chen et al, “Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm”, BMC Bioinformatics 2014, 15:49
19. Ahmad A, Dey L “A feature selection technique for classificatory analysis”, PatternRecognition Letters, 26(1), pp. 43–56, 2005.
20. Su Y, Murali TM, et al “RankGene: identification of diagnostic genes based on expression data”, Bioinformatics 2003, 19:1578–1579.
21. Yu.Wanga, et.al, "Gene selection from microarray data for cancer classification-a machine learning approach ", Computational Biology and Chemistry, vol.29, no.1, pp.37-46, 2005.
22.Carmen Lai, Marcel JT Reinders, Laura J van't Veer and Lodewyk FA Wessels, "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets", BMC Bioinformatics, vol. 7, pp.235-244, 2006.
23.S. Dudoit, J. Fridlyandand T. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data”, J. Am. Stat. Assoc. 97 (2002) 77–87.
24. B. L. Welch, “The generalization of student’s problem when several different population are involved”, Biomethika 34 (1947) 28–35.
25.Cortes C, Vapnik V,“Support-vector networks”, Mach Learn 1995, 20:273–297.
26.X. Wu et al., "Top 10 algorithms in data mining", KnowlInfSyst, vol. 14, pp. 1-37, 2008.
27.Y. Lee and C.K. Lee, “Classification of multiple cancer types by mulitcategory support vector machines using gene expression data,” Bioinformatics, vol. 19, pp. 1132-1139, 2003.