MGS-CM: A Multiple Scoring Gene Selection Technique for Cancer Classification using Microarrays

Dina Ahmed Salem
Dept. of Electronics Eng.-
Faculty of Eng.
Misr University for Science and Technology
Al-Motamayez district-6th of October-Giza-Egypt

Rania Ahmed Abul Seoud
Dept. of Electrical Eng.-
Comm. and Electronics Section-Faculty of Eng
El Fayoum University
Fayoum, Egypt

Hesham Arafat Ali
Computer Eng. System Dept-
Faculty of Engineering
Mansoura University
El-Gomhoria st.- Mansoura, Egypt

ABSTRACT

Microarray is a rich topic which gives the opportunity for researchers to classify cancer samples without any previous biological knowledge. Microarrays high dimensionality characteristic motivated the importance of gene selection techniques. In this paper a new filter multiple scoring gene selection technique MGS-CM is proposed. This technique is further combined with three classifiers to introduce three new classification systems (MGS-SVM, MGS-KNN and MGS-LDA) which are validated and evaluated on three microarray datasets. The proposed MGS-CM technique was proven to be an efficient technique as it extracts the highly informative genes reducing the original datasets by at least 99.6%. Also two of the three proposed classification systems guaranteed the perfect classification (100%) of the leukemia microarray samples. The third one classifies the lymphoma microarray samples with only two misclassifications which is the minimum recorded number. The proposed systems achieved very good results and guaranteed reliable classification for new unclassified samples.

General Terms

Bioinformatics, Classification, Data Mining, Gene Selection

Keywords

Cancer Classification, Microarrays, Multiple scoring gene selection

1. INTRODUCTION

Microarrays as an emerging technology offer a great opportunity for researchers to go deep inside its data to get valuable information. DNA microarray is a practical tool to study the gene expressions resulting from the individual gene sequences printed in a high density array on a glass microscope slide [1]. In other words, high-throughput microarray technology is a hybridization procedure that enables the simultaneous measurement of the abundance of tens of thousands of gene-expression levels from many different samples on a small silicon chip [2]. Due to the huge amount of data (gene expression values) that microarrays contain, it's considered to be a rich area for applying data mining. Data Mining is the automated process of analyzing data from different perspectives to extract previously unknown, comprehensible, and actionable information hidden in large data repositories and using it to make crucial decisions. One of the most important data mining methods is classification. Classification is a supervised machine learning technique works on classifying a new data item into a predefined class [3, 4]. Cancer classification is then the process of assigning a new cancer sample to its correct class. Classifying human cancer samples using microarray gene expression data is the main concern of this paper.

In the past few years gene expression data resulting from microarray technology is extensively used in clustering and classification of cancer samples. Given the difficulty of collecting microarray samples, the number of samples is likely to remain small in many interesting cases. This issue accompanied with the large number of involved genes lead to the high dimensionality problem in microarrays [5]. As classifiers are known to be acting poorly on high dimensional data and also accurate cancer classification is essential for its diagnosis and prognosis, a proper gene selection technique is to be combined with the chosen classifier. Thus, gene selection can be considered a must pre-processing step for classification. Gene selection is a special case of feature selection where the feature is renamed to be a gene [6]. Then for proposing a classification system to classify human cancer samples using microarray data, two main steps are to be studied; implementing an effective gene selection technique and adjusting a powerful classifier. An efficient classification system is the one which gives the highest classification accuracy using the smallest number of genes.

In the paper at hand, a new gene selection technique combined with a powerful classifier (Support Vector Machines) forming a classification system will be proposed. Then this technique will be further combined by another two important classifiers (Linear Discriminant Analysis and K-Nearest neighbor) forming two more classification systems. The three resulting classification systems will be validated and evaluated on three different microarray datasets recording different classification accuracies and number of used genes in each case. The proposed systems were able to classify cancer samples in a dependent way using a small number of expression values of highly informative genes. Their accuracy reaches perfect case on one dataset, the highest recorded on another dataset and a good performance on the third dataset.

The remainder of this paper is organized as follows; Section (2) reviews briefly some of the recent work published in the area of classification of cancer using microarray gene expression values. Section (3) introduces and describes the general scheme of the proposed combined systems. Results of the three proposed systems are presented in section (4). Section (5) analyzes these results. Finally, section (6) concludes the paper.

2. RELATED WORK

The topic of classification of cancer samples (especially human samples) using microarray gene expression databases has been a reach research subject in the past decade. As each classification system consists of two main stages (gene selection technique and classifier) and due to the availability of large amount of implementation ideas for each of them, a lot of studies were carried out in an attempt to reach an optimum system.

In 1999 a generic approach to classifying two types of human acute leukemia using an automatically derived class predictor was first introduced. This paper ends by concluding the possibility of cancer classification using gene expression monitoring without any need to previous biological knowledge [7]. In 2000, Moler *et al.* was able to classify the colon adenocarcinoma tissue specimens labeled as tumor or nontumor microarray dataset by combining a naïve Bayesian model with the support Vector Machine (SVM) classifier [8]. In 2003, another research concentrates its study on introducing a new technique for feature (gene) selection or can be called pre-filtering technique. This technique depends on grouping similar genes by proposing three grouping algorithms. Then it chooses the highly informative genes from each cluster (group) by using a t-statistic technique. Three datasets are used for validation by classifying them with SVM (RBF kernel) and recording the performance of each gene selection technique [9]. Later in 2006, another new feature extraction method based on the discrete wavelet transform (DWT) combined with SVM classifier was proposed. Two standard benchmark data sets are used for evaluating the proposed technique [10].

In 2007, J. Zhang and H. Deng chose their reduced set of genes by first carrying a gene preselection using a univariate criterion function and then estimating the upperbound of the Bayes error to filter out redundant genes from remaining genes derived from gene pre-selection step. To validate their system they used two classifiers; k-nearest neighbor (KNN) and SVM on five datasets [11]. P. Yang and Z. Zhang used two different datasets to validate their two proposed systems using the genetic algorithm (GA) for gene selection. Then, the obtained reduced set of informative genes is applied to two classifiers; Decision Tree and Neural Network forming the two systems (GADT, GANN) [12]. In 2011 two other systems used for classifying the leukemia microarray dataset was by blending of Support Vector Machine as a classifier, once with Locality Preserving Projection technique (LPP) and the other with F-score ranking feature selection technique [13, 14].

The comparative studies are very important in the area of cancer classification using microarray data to evaluate different systems on different datasets. One broad comparative study was carried out by combining seven gene selection techniques with four different classifiers using three combining methods. The result was 42 ensemble classifiers which were evaluated using three public datasets [15]. Another group of researchers compare the most common univariate gene selection techniques with some of the recent multivariate techniques and record their results on seven microarray datasets of different cancer types [16].

2.1 Problem Formulation

Most of the present work in the field of classifying cancer samples using microarray gene expression values concentrates on only one of the two stages of the classification system. Otherwise some of them used only one dataset for evaluating the proposed system. The problem is that there are no rules regulating the design of classification systems and some scientists refer to the solution of this problem by being a trial and error solution. This is because although one gene selection technique can have a perfect performance if combined with a specific classifier and evaluated on one of the valid datasets, it may respond differently when changing the classifier or the dataset. Another problem arises when using only one univariate technique for gene selection. Univariate techniques usually study only one gene criterion in their design. This means that the selected set of highly informative genes may contain redundant genes. The presence of more than one gene in the reduced set carrying the same information leads to increasing the number of used genes. Number of used genes is to be minimized to reduce time, cost and increase performance.

2.2 Plan of Solution

To improve the classification performance and trying to find a unified system which is able to classify different cancer datasets, MGS-CM is proposed. MGS-CM is a new multiple scoring gene selection technique designed after studying the behavior of many univariate techniques. Then it is combined with three classifiers resulting in three classification systems which are used to classify cancer samples from three different microarray datasets. The efficiency of the three systems is measured by recording the classification accuracy for each of them and the number of used genes in every case. Then they are compared with other same task systems proposed in some recent literature. The three proposed classification systems are very powerful systems and achieved very much comparable results. They can integrate any other classifier and can be tested on any other dataset.

3. METHODOLOGY

The main objective of this paper is to introduce three classification systems (MGS-SVM, MGS-KNN and MGS-LDA). The three systems are using the same proposed gene selection technique (MGS-CM) but combined with three different classifiers. Then they are validated using three different microarray datasets (leukemia, lymphoma and colon cancer). The leukemia dataset is first used to design the MGS-CM technique as this dataset had previously shown perfect classification using different set of genes by many researchers [9, 10, 11]. To design the two stages of the proposed classification systems, different parameters are taken into consideration in a trial to reach the highest classification accuracy using the smallest set of genes. The performance of these systems is justified by measuring the classification accuracy (CA) which is equal to the number of correct classified samples divided the total number of samples needed to be classified. This is in addition to recording the number of used genes each time.

3.1 Microarray Gene Expression Datasets

Microarray datasets take the form of expression data matrix where rows represent the genes and columns represent the samples. Each cell in this data matrix is a gene expression value which expresses the gene intensity in the corresponding sample. The expression data matrix will be finally dealt with in the form X_{ij} where; $0 < i \le n_g$, $0 < j \le n_s$ and n_g, n_s are the total number of genes , total number of samples respectively as in figure (1). Each expression data matrix will be further divided into two matrices; training data matrix (Y_{ik}) and test data matrix (Z_{ip}) where k, p are the number of samples used in the training process, test process respectively and p + k = n_s. The training data matrix will be used to train all the used classifiers and their performance will be evaluated using the test data matrix only.

$$X_{ij} \quad = \quad \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots\cdots\cdots & x_{1n_s} \\ x_{21} & x_{22} & x_{23} & \cdots\cdots\cdots & x_{2n_s} \\ x_{31} & x_{32} & x_{33} & . & x_{3n_s} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n_g1} & . & . & . & x_{n_gn_s} \end{pmatrix}$$

Fig (1) : expression data matrix

3.1.1 Leukemia dataset

The Leukemia dataset was first classified by golub et al. in 1999 to two types of acute leukemia; Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). This dataset contains 72 microarray experiments (samples) with 7129 gene expression levels. The complete dataset contains 25 AML samples and 47 ALL samples [7]. 38 out of 72 samples are used as training data (27ALL samples and 11 AML samples) and the remaining 34 samples (20 ALL samples and 14 AML samples) are used as test data.

3.1.2 Lymphoma dataset

Lymphoma is a type of cancer derived from lymphocytes (a type of white blood cell) of the immune system. In this paper, we are about to classify two kinds of lymphoma; Diffuse Large B-Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL). The lymphoma dataset contains 77 samples with 7129 gene expression values; 58 of them are of the first class (DLBCL) and 19 belong to the second class (FL) [17]. The train data matrix will be of size 40*7129 and the size of the test data matrix will be 37*7129.

3.1.3 Colon dataset

The original dataset contains the expression of 6000 genes with 62 cell samples taken from colon cancer patients, but only 2000 genes were selected based on the confidence in the measured expression levels [18]. The proposed classification systems are supposed to classify between the 40 tumor and 22 non-tumor samples. The gene expression data matrix is divided into 32*2000 train data matrix and 30*2000 test data matrix.

3.2 Gene Selection Techniques

Cancer microarray data usually consists of a few hundred samples with thousands of genes as features. Classification of data in such a high dimensional space is impossible as this may lead to over fitting, in addition to the ultimate increase in the processing power and time [19]. This gives rise to the need of the gene selection techniques which aim to find a subset of highly informative and relevant genes by searching through the space of features. These 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 [20].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 [21]. 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]. To design the proposed gene selection technique the following eight univariate filter gene selection techniques (six of them are parametric and two are non-parametric) are implemented and evaluated using a support vector machines classifier on leukemia dataset.

Means difference (MD): This is the simplest idea of the parametric gene selection techniques ever but it is very important to be introduced because some of the following parametric techniques depend on its operation. This technique will be abbreviated after by MD. It depends mainly on the two classes (which is required to discriminate between them) distinction. First the data set is split into two sets; one for the first class and another one for the second class. Then calculate the mean of the expression values for each of the n_g genes (μ_{i1} for the first class and μ_{i2} for the second class) and obtain the absolute differences between the calculated means ($|\mu_{i1} - \mu_{i2}|$). At last rank the genes in a descending order.

Signal to noise ratio (SNR): The signal to noise ratio (SNR) test was first proposed by Golub *et* al. in1999 [7]. It gives each gene a value according to the maximal difference in mean expression between two groups and minimal variation of expression within each group [22]. In this method genes are first ranked according to their expression levels using SNR test Statistic. The SNR is defined by the following equation where μ_{i1} and μ_{i2} are the mean differences for the sample class 1 and class 2 respectively and σ_{i1} and σ_{i2} are the standard deviations for the samples in each class and i = 1 to n_g.

$$\mathbf{SNR\ (\ i\ ) = (\mu_{i1} - \mu_{i2})\ /\ (\sigma_{i1} + \sigma_{i2})} \qquad (1)$$

F(x) score (FS): F-score ranks the genes twice; one time according to the two classes mean difference for each gene and another time according to the Signal-to-Noise ratio (SNR) criterion. So it can identify the genes whose expression shows great change in both classes. As shown it can be considered a combination of the two previous techniques. But it first chooses the highest 250 genes according to the MD technique and then it gives a score to these genes only according to their SNR value and ranks them descendingly [14].

Fisher discriminant criterion (FC): This test was first introduced in 1973 by Duda *et* al. [23]. It gives higher values to features whose means differ greatly between the two classes, relative to their variances and it is expressed by equation (2). Then the genes are arranged descendingly where the first genes are considered the most informative genes from FC point of view [24].

$$\mathbf{FC\ (\ i\ ) = (\mu_{i1} - \mu_{i2})^2 \,/\, (\sigma_{i1}^{\ 2} + \sigma_{i2}^{\ 2})} \qquad \textbf{(2)}$$

T-test statistics (TS): 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 MD and then normalizing it by an expression of variances as illustrated in equations (3) and (4). 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. Then, we can use the significance level, which is a threshold of *p*-values, to determine a set of informative genes [25].

$$\boldsymbol{TS\ (i) = \frac{\boxed{?}_{i1} - \boxed{?}_{i2}}{S_w\sqrt{\frac{1}{n_{s1}}+\frac{1}{n_{s2}}}}} \qquad \textbf{(3)}$$

$$\boldsymbol{S_W{}^2 = \frac{(n_{s1}-1)\sigma_{i1}{}^2+(n_{s2}-1)\sigma_{i2}{}^2}{n_{s1}+n_{s2}-2}} \qquad \textbf{(4)}$$

Entropy (E): According to Shannon's information theory [26], entropy can be expressed by equation (5). Some researchers measure the entropy as the mutual information which expresses the dependency relationship between two probabilistic variables of events. The mutual information given in equation (6) is equal to zero if the two variables are completely independent and its value becomes closer to one when the dependency increases. Other researchers define entropy as the uncertainty of a random variable [15]. Here we are using entropy as the information gain described by Shannon' information theory. This means that the highest value of entropy for a gene, the more informative this gene is.

$$\boldsymbol{E(x_i) = \sum P(x_i)\log_2 P(x_i)} \qquad \textbf{(5)}$$

$$\boldsymbol{MI(x_i, c_j) = \log\frac{P(x_i,c_j)}{P(x_i)P(c_j)}} \qquad \textbf{(6)}$$

Correlation coefficient (CC): The correlation coefficient test measures how much each gene is correlated to all the samples which means it is a non-parametric test. This is achieved by assuming an ideal gene named Y which is interpreted as the most informative gene that predicts the two classes of all the samples perfectly. Then a CC value is calculated according to equation (7) where n_s is the total number of samples, Y is the ideal gene and X is the gene expression value. The highest the CC value the more informative the gene as it will be more correlated to ideality [15].

$$\boldsymbol{CC(i) = \frac{n_s\sum XY-\sum X\sum Y}{\sqrt{(n_s\sum X^2-(\sum X)^2)(n_s\sum Y^2-(\sum Y)^2)}}} \qquad \textbf{(7)}$$

Euclidean distance (ED): This is a non-parametric test where the distance between each gene and the ideal gene Y is calculated according to the Euclidean distance and expressed by equation (8). Each gene expression value in one sample and the corresponding value in the ideal gene are treated as two points in space. The distance between each gene and the ideal gene is the summation of the distances in all samples [15].

$$\boldsymbol{ED\ (i) = \sqrt{\sum(X-Y)^2}} \qquad \textbf{(8)}$$

3.3 Classifiers

As mentioned before, classification is the process of classifying a new data item (cancer sample in microarray data) into a predefined class. Adjusting powerful classifiers are a very important stage in this process. Classifiers act in different ways on different datasets resulting in a variety of classification accuracies. To study this problem three classifiers are to be used in this paper which are; Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN) and Support Vector Machines (SVM). Following is a brief description of each of employed classifiers.

3.3.1 Linear Discriminant Analysis (LDA)

LDA approaches the classification problem by finding the transformation matrix which helps to preserve most of the information that can be used to discriminate between the different classes. This is achieved after reshaping the data by projecting high-dimensional data onto a low-dimensional space. To reach the optimum transformation matrix two matrices S_B (between-class scatter) and S_W (within-class scatter) are to be calculated according to equations (9),(10) where; n_k is the number of training samples from class k, c_k is the set of indices of the training examples belonging to class *k*, x_i is the gene expression value of gene i, μ_k is the mean value of class k and μ is the result mean value of the two classes. LDA becomes ready to classify new samples after finding an optimum value for vector *w* such that w^tS_Bw is maximized while w^tS_Ww is minimized as shown in equation (11) [27].

$$\mathbf{S_w = \sum_k \sum_{x_i \in\, c_k} (x_i - \mu_k)\,(x_i - \mu_k)^t} \qquad \textbf{(9)}$$

$$\mathbf{S_B = \sum_k n_k\,(\mu_k - \mu)\,(\mu_k - \mu)^t} \qquad \textbf{(10)}$$

$$\boldsymbol{F(w) = \frac{w^tS_Bw}{w^tS_ww}} \qquad \textbf{(11)}$$

3.3.2 K- Nearest Neighbor (KNN)

KNN is known to be a lazy technique as it depends on calculating a distance between a test data and all the train data. So for using KNN three key elements must be present; a set of data for training (train data), a group of labels for the train data (identifying the class of each data entry) and the value of K to decide the number of nearest neighbors. KNN main idea is to assign a new data item (sample) to the class to which the majority of the chosen number of neighbors belongs. Distances for KNN can be calculated by different ways such as Euclidean distance which is the most used one. Other examples are cosine measure, cityblock and correlation measure. Then to guarantee the highest classification accuracy, it's better to try different values of k accompanied with different measures of the distance. Although being a simple technique and easy to implement, KNN shows an outstanding performance in many cases such as cancer classification using microarray gene expression values. This is because microarray data is characterized by having a small number of samples and after using a gene selection technique it also have few number of genes [28].

3.3.3 Support Vector Machines (SVM)

SVM is a very powerful data mining technique which can be used for classification and regression. It is considered also a machine learning technique as it learns by examples. SVM is widely used by many researchers in classification of cancer samples using microarray gene expression profiling and shows promising results. To classify cancer samples SVM first construct a hyperplane which separates the two classes using equation (12). Then it uses an optimization solution to find the maximum margin hyperplane which have the largest distance to the nearest data points from both classes as expressed by equation (13)

$$\boldsymbol{y_i(w \cdot x_i - b) \geq 1 \quad \forall \quad i = 1 : n} \qquad \textbf{(12)}$$

$$\boldsymbol{min_{w,b} max_{\propto} \left\{\left\{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \propto_i [y_i(w \cdot x_i - b) - 1]\right\}} \qquad \textbf{(13)}$$

Where w is the normal vector, $\frac{b}{\|w\|}$ is the offset of the hyperplane from the origin along the normal vector w, x_i is the sample with i ranges from 1 to total number of samples and α is the lagrange multiplier factor used for the optimization problem. SVM solves two more problems for classifying the samples. The first problem is that some data points may lie in the region of the other class which is solved by introducing the soft margin using the slack variable ζ in equation (14). Another problem arises if the samples are not linear separable which is solved by using different kernel functions which map the non-linear separable samples into the feature space. Different kernel functions include; Gaussian, polynomial, and RBF [29], [30].

$$\boldsymbol{min_{w,\zeta,b} max_{\propto,\beta} \left\{\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \zeta_i - \sum_{i=1}^{n} \propto_i \left[y_i(w \cdot \right.\right.\right.}$$
$$\boldsymbol{xi – b – 1 + \zeta i – i = 1 n \beta i \zeta i} \qquad \textbf{(14)}$$

For perfect training of SVM classifier, it runs usually accompanied with a cross-validation technique. Cross-Validation is a statistical technique which divides the training data into two sets according to a k value to be named K-fold cross-validation. One set is used to train the classifier and the other set is used to validate the training process. The simplest form of the cross-validation is when using k=1 which is called Leave-One-Out-cross-Validation (LOOCV). In LOOCV all the training dataset is used to train the classifier except one which is left for validation [31].

3.4 Proposed System Workflow

To reach the final form of the proposed classification systems, three stages are implemented. The first stage is shown in figure (2) and it is the stage of choosing the univariate filter techniques which will be further used to implement the MGS-CM technique. This stage is carried out for each one of the eight previously explained univariate gene selection techniques alone. Each technique is evaluated using the SVM classifier (linear kernel) and recording the CA on the leukemia dataset. This stage takes the following steps:

- First the leukemia dataset is divided into train dataset and test dataset.
- Then the chosen gene selection technique is carried out on the train dataset to rank the genes according to its criterion.
- Five reduced train datasets are chosen where the number of the highest informative genes is different in each one. This means that the five datasets contains the highest 200, 100, 50, 20 and 10 relevant genes.
- Another five same size test datasets are extracted from the original test dataset where each one contains the expression values of the genes present in the five reduced train datasets.
- Each one of the reduced train datasets is used to train the SVM and then the corresponding test dataset is introduced to the SVM classifier and the resulting five values of the CA is measured and recorded.
- The last step in this stage is to calculate the average CA (ACA) for each one of the eight gene selection techniques.

The gene selection techniques with the highest ACA (ACA ≥ 0.9) are chosen to be introduced to the second stage. They are the MD, FS, CC and E techniques.

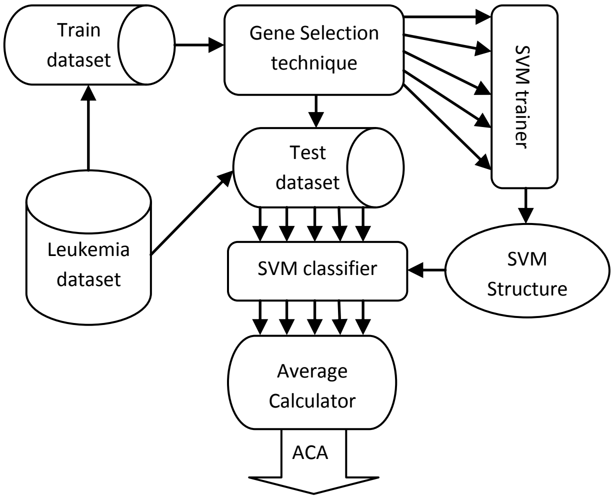

Fig 2: Univariate Gene Selection Technique

The second stage is the responsible stage for giving the MGS-CM technique its final form. The four chosen filter univariate gene selection techniques from the first stage are combined together by an intersector as shown in figure (3). This intersector works as follows:

- First specifying a value g which is the number of genes considered highly informative according to each one of the four chosen gene selection techniques.
- A matrix (n_g*4) is formed where the rows represent the original total number of genes and each column represent one of the four chosen techniques.
- A gene is assigned a value 1 if it is considered highly informative from the chosen technique point of view.
- Only the genes which have value 1 in all the columns are chosen to build the final reduced train and test datasets.
- Different values of g are chosen in the design.

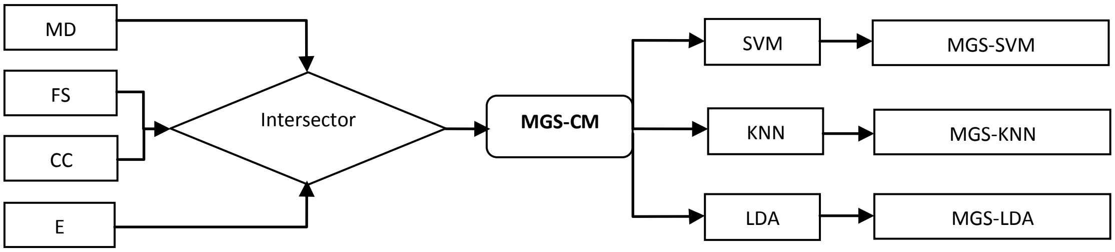

Fig 3: Proposed System Workflow

The last stage is combining the proposed MGS-CM technique with the three previously explained classifiers to form the three classification systems (MGS-SVM, MGS-KNN and MGS-LDA). For SVM classifier different values for k in the cross validation (1, 2, 5, 10) and different kernel functions (linear, Gaussian and RBF) are used. For KNN classifier different numbers of nearest neighbors are chosen with different types of distance measurements. The second and third stages are repeated for the other two datasets. All the previous work was implemented in MATLAB 7.10.0 (R2010a).

4. RESULTS

Recording the results of the work done in this paper is carried out on three stages. The first stage lists the results of the eight filter univariate techniques used in the design of the proposed MGS-CM technique. Table (1) records the results of this stage represented in the Average Classification Accuracy (ACA) of each of the eight techniques when combined with linear SVM and applied to the leukemia dataset. The highest ACA are highlighted in bold font. Then table (2) shows the importance of the proposed MGS-CM technique by recording the used number of ranked genes and number of genes after reduction. This is carried out on the three datasets.

Table 1. ACA of the univariate gene selection techniques

Tech.	**MD**	**SNR**	**FS**	**FC**	**TS**	**E**	**CC**	**ED**
ACA	**0.95884**	0.8652	**0.92944**	0.85882	0.88236	**0.97648**	0.91178	0.81856

Table 2. MGS-CM technique evaluation

	500	**400**	**300**	**200**	**100**	**50**
Leukemia	54	15	9	5	4	0
Lymphoma	67	37	26	11	3	0
Colon	158	95	54	51	16	8

The second stage records the results of the three proposed classification systems (MGS-SVM, MGS-KNN and MGS-LDA). For each classification system many attributes for each classifier are used as discussed in the methods section. This results in a huge amount of results. So it's more practical to list only the highest result of each system for every reduced dataset as this is the aim of this paper. This is represented by three graphs for the three datasets (see Figures 4,5,6). Each graph contains the highest CAs of each of the proposed systems when tested on the gene subsets of table (2). In each graph the X-axis represents the number of genes after reduction by the proposed MGS-CM technique while the Y-axis represents the CA.

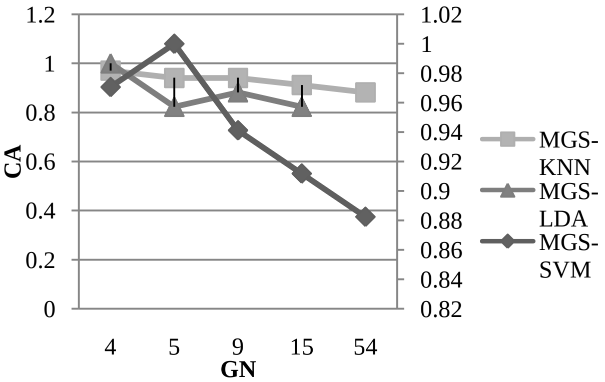

Fig 4: Leukemia dataset

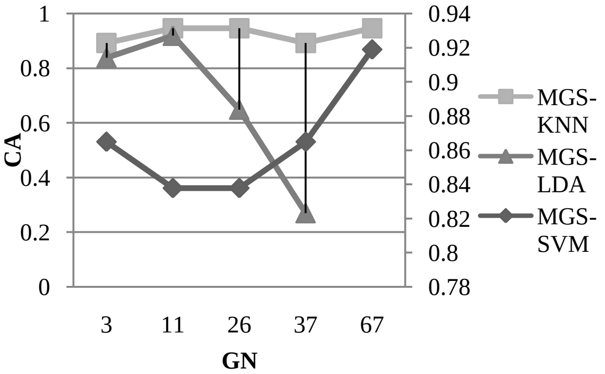

Fig 5: Lymphoma dataset

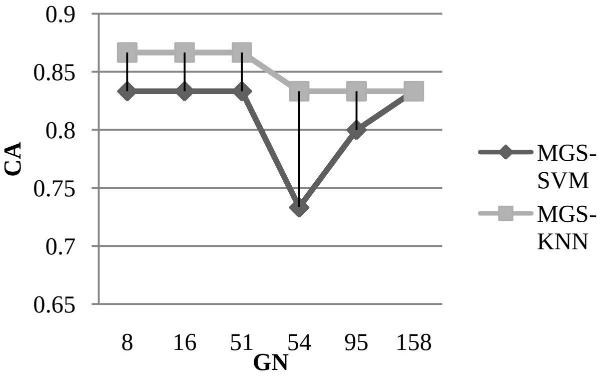

Fig 6: Colon dataset

The third stage is to establish a search process for the best results. This process has two main steps. The first step gives the priority to the CA. This means that it starts with locating the highest CA for each system on each dataset. The second step comes if this CA appears more than once in the same system. This step searches for minimum number of genes resulted in the highest CA (GN). Table (3) lists the results of this search process, in addition to the classifier features (CF) corresponds to the recorded values.

Table 3. Proposed systems evaluation

		Leukemia	Lymphoma	Colon
MGS-SVM	CA	1	0.9189	0.8333
	GN	5	67	8
	CF	Linear SVM all k values	Linear SVM K=2,5,10	Linear SVM all k values
MGS-KNN	CA	0.9706	0.9459	0.8667
	GN	4	11	8
	CF	K=1,2,5	K=10	K=1
MGS-LDA	CA	1	0.9189	0.7333
	GN	4	11	16
	CF	Quadratic	Linear	Linear

To fully evaluate the work done it is compared with some very recent published papers. Table (4) shows the results of these research papers on the used datasets. Each cell in the table contains two main assessment values; the CA and the used number of genes (between parentheses). The table cells with a dash enclosed means that the author of the corresponding paper didn't test his system on the specified datasets.

Table 4. Related Work Results

	Leukemia	Lymphoma	Colon
Li et al.[10]	100 (100)	-	93.55 (250)
Zhang et Deng [11]	100 (3)	92.21(6)	90.32 (12)
Yang et Zhang [12]	98.05 (5)	-	94.92 (15)
Salome et Suresh [13]	97.2973 (38)	-	-
Seeja et Shweta [14]	94.1176 (200)	-	-

5. ANALYSIS

For the leukemia dataset two of the three proposed systems (MGS-SVM and MGS-LDA) succeeded to reach the perfect classification (i.e. all samples are correct classified) using 5 and 4 genes respectively. This result is better than all the papers listed in table (4) except for [11] who records the same CA with only one gene less. With the same 4 genes MGS-KNN results in only one misclassification. Then the three proposed systems are very much effective for this dataset. They employ the same gene selection technique but they are very different in their operation. This means that the proposed MGS-CM technique can efficiently extract the minimum number of highly informative genes from the leukemia dataset.

Opposing to the leukemia dataset which is almost used in all the related published papers, the Lymphoma dataset is rarely found. In the proposed MGS-KNN the lymphoma samples were classified with an error rate equals 0.0541 which is better than the minimum error previously recorded. This is achieved using only 11 genes. Also the MGS-SVM and MGS-LDA have high performance on this dataset as they misclassified only one more sample.

None of the previous work could reach perfect classification of the colon dataset samples. In the proposed classification systems the maximum CA is 0.8667 with only 4 misclassifications using only 8 genes. Although this is a good result but it's not the best as achieved on the two previous datasets. This may be a consequence of using a partial reduced dataset (containing 2000 genes) instead of the original dataset (6000 genes). The problem is that only the partial reduced colon dataset is available for download.

6. CONCLUSION

In the paper at hand three classification systems are proposed (MGS-SVM, MGS-KNN and MGS-LDA). These classification systems have only one common part which is the new designed multiple scoring gene selection technique (MGS-CM). Otherwise the three of them are totally different as each has a different theory. The proposed technique idea is to select the genes which are ranked as highly informative from different perspectives. Three microarray datasets are used to evaluate the proposed systems. The main objective is to correct classify the human cancer samples stored in these datasets. The proposed gene selection technique is proven to be highly effective as it is capable of fetching the minimum highly informative genes from very large number of genes. Also the proposed classification systems could achieve perfect classification of the leukemia dataset samples with only 4 genes. On the Lymphoma samples the proposed systems works great and results in classifying more correct samples that have ever been recorded using only 11 genes.

7. REFERENCES

[1] J. Derisi, V. Iyer, and P. Brosn, "Exploring the metabolic and genetic control of gene expression on a genomic scale", Science 278:680-686,1997.

[2] C. Kong, J. Yu, F. Minion, K. Rajan, "Identification of Biologically Significant Genes from Combinatorial Microarray Data", ACS Combinatorial Science, 2011.

[3] Larose, D. T. 2005 Discovering knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc.

[4] E. Simoudis, "Reality check for data mining", IEEE Expert, 26-33, 1996.

[5] G. Piatetsky-Shapiro and P. Tamayo, "Microarray Data Mining: Facing the Challenges", SIGKDD Explorations, 5: 1-5, 2004.

[6] H. Ong, N. Mustapha, M. Sulaiman, Integrative Gene Selection for Classification of Microarray Data", 4(2):55-63, 2011.

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

[8] E. Moler, M. Chow, I. Mian, "Analysis of molecular profile data using generative and discriminative methods", Physiological Genomics , 4(2):109-126,2000.

[9] Jaeger, J., Sengupta R., and Ruzzo,W. 2003. Improved gene selection for classification of microarrays. Pacific Symposium on Biocomputing. pp. 53-64.

[10] Li, S., Liao, C., and Kwok, J. T. 2006.Wavelet-Based Feature Extraction for Microarray Data Classification. Presented at the International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada.

[11] J. Zhang, H. Deng, "Gene selection for classification of microarray data based on the Bayes error", BMC Bioinformatics, 8(1):370, 2007.

[12] Yang, P., and Zhang, Z. 2007. Hybrid Methods to Select Informative Gene Sets in Microarray Data Classification. In Proceedings of the Australian Conference on Artificial Intelligence. Verlag Berlin Heidelberg: pp.810-814.

[13] J. Salome, R. Suresh, "An Effective Classification Technique for Microarray Gene Expression by Blending of LPP and SVM", Medwell Journals : Asian Journal of Information Technology, 10(4):142-148, 2011.

[14] K. Seeja, and Shweta, "Microarray Data Classification Using Support Vector Machine", International Journal of Biometrics and Bioinformatics (IJBB), 5(1):10-15, 2011.

[15] S.-B. CHO, H.-H. WON, "Data Mining for Gene Expression Profiles from DNA Microarray", International Journal of Software Engineering and Knowledge Engineering, 13(6):593-608, 2003.

[16] C. Lai, M. J. Reinders, L. J. van't Veer, and L. F. Wessels, "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets", BMC Bioinformatics, vol. 7:235, 2006.

[17] M. A. Shipp et al., "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expresssion Profiling and Supervised Machine Learning", Nature Medicine, 8(1):68-74, 2001.

[18] U. Alon et al., "Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays," Proceedings of the National Academy of Sciences of the United States of America,1999, vol. 96, pp. 6745-6750.

[19] Huerta, E. B., Duval, B., and Hao, J.-k. 2006. A hybrid GA/SVM approach for gene selection and classification of microarray data. In Proceedings of the EvoWorkshops, LNCS 3907.pp. 34-44.

[20] I. Guyon, A. e. Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, 3:1157-1182, 2003.

[21] Y. Wanga, I. V. Tetkoa, M. A. Hallb, E. Frankb, A. Faciusa, K. F. X. Mayera, and H. W. Mewesa, "Gene selection from microarray data for cancer classification", Computational Biology and Chemistry, 29(1):37-46,2005.

[22] D. Mishra and B. Sahu, "Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach", International Journal of Scientific & Engineering Research, vol. 2, 2011.

[23] Duda, R. O. and Hart, P. E. 1973 Pattern Classification and scene analysis. Wiley.

[24] Hernandez, J. C., Duval, B. and Hao, J.-K. 2007. A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data. In proceedings of EvoBIO, LNCS 4447, pp. 90-101.

[25] Deng, L., Peiz, Ma, J. J. and Lee, D. L. 2004. A Rank Sum Test Method for Informative Gene Discovery. In Proceedings of KDD'04, Seattle, Washington, USA.

[26] Shannon, C. E. and Weaver, W. 1949 The Mathematical Theory of Communication. University of Illinois Press.

[27] G. Balakrishnama, "Linear discriminant analysis - a brief tutorial," 1998. [Online]. Available: http://citeseer.ist.psu.edu/contextsummary/1048862/0

[28] X. Wu et al., "Top 10 algorithms in data mining", *Knowl Inf Syst*, vol. 14, pp. 1-37, 2008.

[29] Noble, W. S., "What is a support vector machine?", NATURE BIOTECHNOLOGY, vol. 24, pp. 1565-1567, 2006.

[30] Lessmann, S., Stahlbock, R. and Crone, S. F. 2006. Genetic Algorithms for Support Vector Machine Model Selection. In Proceedings of the International Joint Conference on Neural Networks, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, pp. 3063-3069.

[31] Saeedmanesh, M., Izadi, T. and Ahvar, E. 2010. HDM: A Hybrid Data Mining Technique for Stock Exchange Prediction. InProceedings International MultiConference of Engineers and Computr Scientists (IMECS).