A Hybrid Reduction Approach for Enhancing Cancer Classification of Microarray Data

Abeer M.Mahmoud
Dept. of Computer science,
Faculty of computer and information science,
Ain Shams University, Cairo, Egypt

Basma A.Maher
Dept. of Computer science,
Faculty of computer and information science,
Ain Shams University, Cairo, Egypt

***Abstract*—This paper presents a novel hybrid machine learning (ML)reduction approach to enhance cancer classification accuracy of microarray data based on two ML gene ranking techniques (T-test and Class Separability (CS)). The proposed approach is integrated with two ML classifiers; K-nearest neighbor (KNN) and support vector machine (SVM); for mining microarray gene expression profiles. Four public cancer microarray databases are used for evaluating the proposed approach and successfully accomplish the mining process. These are Lymphoma, Leukemia SRBCT, and Lung Cancer. The strategy to select genes only from the training samples and totally excluding the testing samples from the classifier building process is utilized for more accurate and validated results. Also, the computational experiments are illustrated in details and comprehensively presented with literature related results. The results showed that the proposed reduction approach reached promising results of the number of genes supplemented to the classifiers as well as the classification accuracy.**

Keywords—Mining Microarray data; Cancer classification; SVM

I. INTRODUCTION

Creatures consist of organisms and every organism carries the same genetic information. This genetic information is represented in the form of genes, where only a subset of these genes is active or expressed. Simply, Microarray gene expression data refers to such repositories of gene information that made the technology of modern biological research. Its goal is to understand the regulatory mechanism that governs protein synthesis and activity of genes. Furthermore, analyzing the gene with respect to whether and to what degree they are expressed can help characterize and understand their functions. It can further be analyzed how the activation level of genes changes under different conditions such as for specific diseases (e.g. cancers are generally caused by abnormalities in the genetic material of the transformed cells or change in their activation or function) [1].Actually, microarray represents a powerful tool in biomedical discoveries and harnessing the potential of this technology depends on the development of appropriate mining approaches [1-4].

The mining phase in the knowledge discovery process can be defined as the process of discovering interesting and unknown patterns from large amounts of data stored in information repositories [5,6]. The mining task could be one of regression, summarization, clustering and classification [5]. Classification is certainly a helpful research area in cancer diagnosis and drug discovery.

Based on the fact that microarray data is a high dimensional data with small number of samples and huge number of genes; then achieving a successful mining results with target of highly accurate and satisfied classification, the whole mining process must be divided into two main phases 1) finding the prioritized genes subset and2) building the classifier [3-6]. ML approaches achieved powerful results in data mining area. It is a bough of Artificial Intelligence (AI) that uses a variety of statistical, probabilistic and optimization methods that permit computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets [5]. In the literature, there are several ML techniques for both phases. Examples of the most widely applied gene prioritized techniques for microarray data are Mean Difference (MD), Signal to noise ratio (SNR), F(x) score (FS), Fisher discriminant criterion (FC), T-test, Entropy (E), Correlation Coefficient (CC), Euclidean distance (ED), and CS [6]. Also, examples of classification techniques are Support Vector Machine (SVM), K-Nearest neighbor (KNN), Fuzzy Neural Network (FNN), and Linear Discriminate Analysis (LDA) [2-6].

In this paper, we conduct a comprehensive study that focuses on exploring and analyzing the efficiency of applying ML approaches for cancer classification. In addition, the paper mainly proposes a novel hybrid reduction approach for the enhancement of cancer classification of microarray data based on T-test, CS, KNN and SVM. The residue of this paper is primed as follows. Section 2 focuses some of related research work. Section 3 provides our methodology for reaching results of this paper. Section 4 details the four public microarray databases; with samples of their genes and their experimental settings. Computational results, comparisons & discussions are presented in section 5. Section 6 concludes the paper.

II. RELATED WORK

(2005), Wang, et al., [4]; highlighted the challenging task to choose relevant genes involved in different types of cancer. They purposed a feature selection algorithm for microarray data based on Wrappers Filters and CFS (correlation-based feature selector) and the ML algorithms such as decision trees, naïve Bayes and SVM for the classification phase. The data used in this paper was leukemia and lymphoma. F. Chu & L. Wang [7], used a SVM for cancer classification with the microarray gene expression data. The selection of genes has been completed by the use of four effective feature dimensionality reduction methods, for instance, principal

components analysis (PCA), CS measure, FC, and T-test. The data set used here is SRBCT, lymphoma and leukemia data sets. The results showed that genetic selection of T-test performed well than the other three approaches. Also, in all the three data set, the SVMs obtained very good accuracies with very few numbers of genes.

(2006), Jin et al., [8]; proposed a ML techniques and used Serial Analysis of Gene Expression (SAGE) technology to facilitates and concurrently measure the expression levels of tens of thousands of genes in a inhabitants of cells. They used Chi-square is used for tag/gene selection. They investigated both binary and multi-category classification. Their experiments are performed on two human SAGE datasets: brain and breast. The results show that SVM with Chi-square is the outperforming SAGE classifier.

(2007), Wang et al., [9], proposed a new approach of two main steps. First step is gene selection, where the scoring method such as T-test, CS is used. The second one is the classification accuracy of gene combination that has been carried by using a fine classifier. Divide and conquer approach are used to attain good accuracy. Two of the datasets used in this experiment are Lymphoma Data, SRBCT Data. They used a KNN algorithm, for the treatment of missing values in microarray data. Also, they used a FNN and SVM classifier. The top marker genes are passed one by one to the classifier until good accuracy is achieved.

(2009), M. Rangasamy & S. Venketraman [10] developed a new algorithm for ranking the gene based on a classical statistical technique and two various classifiers. The paper used two types of databases, two classes datasets such as Liver and Leukemia and more than two classes database such as Lymphoma. They used a Gene selection like ANOVA, LDA and SVM-OAA RBF Kernel according to suitability of database type. Also, they used SVM-one-against-all (SVM-OAA) and LDA as a classifier for performance evaluation. The classifier is trained using all possible gene combinations; therefore the best gene combination was reported. Manuel et al., [11] presents a Kernel Alignment KNN for cancer classification using gene expression profiles. Kernel alignment KNN performs well when compared with other metric learning strategies and improves the classical KNN.

(2010), N. Revathy & R. Amalraj [12] developed a new technique that combines the enrichment score with the SVM classifier for cancer classification in microarray data. The data set is randomly divided into training and testing. The gene ranking is done then the top genes is passed into the classifier one by one if no good accuracy is attained, gene combination can be performed from the ranked data set. The performance accuracy of the SVM with the enrichment score performed well with higher accuracy than the SVM with T-Score.

(2011), Z.Ghorai et.al, [13] offered a nonparallel plane proximal classifier (NPPC) ensemble for cancer classification based on microarray gene expression profiles. A hybrid computer-aided diagnosis (CAD) framework is introduced based on filters and wrapper methods. Minimum redundancy maximum relevance (MRMR) ranking method is used for feature selection. The wrapper method is applied on those gene sets to reduce the computational burden and nonparallel plane proximal classifier (NPPC).

(2013), Abeer M. Mahmoud, et.al [14] highlighted the discovery of differentially expressed genes (DEGs) in microarray data in their way to build an accurate and cost effective classifier. A T-Test feature selection technique and KNN classifier was applied on the Lymphoma data set to reach the DEGs and to analyzing the effect of these genes on the classifier accuracy, respectively.

III. Computational Intelligence Techniques

The main objective of this paper is to successfully mine the high dimensional microarray data using ML techniques and hence propose a better approach for the mining process. The mining process will be divided into two main phases 1) finding the prioritized genes subset and2) building the classifier. Two approaches for gene ranking (T-test and CS) and two classifiers (KNN & SVM) are used. Therefore, the coming subsections presents necessary background and nomenclatures for understanding the applied ML techniques.

A. Finding the Prioritized Genes

Gene expression profiling or microarray analysis has enabled the measurement of thousands of genes. Identifying genes that are differentially expressed under two or more treatment conditions is a primary goal of most microarray studies. Traditionally, the methods for gene selection are broadly divided into three categories: filter, wrapper and embedded methods [15]. A filter method relies on general characteristics of the training data to select genes which show dependences on the class labels without involving any classifier for evaluation [16].They evaluate a gene based on its discriminative power for the target classes without considering its correlations with other genes. Examples are methods based on statistical ranking of individual genes, such as, correlation coefficient, t-statistics, class separability, or Fisher's criterion, etc. [6]. The wrapper methods involve the classifiers as evaluation functions and search for the optimal gene set for classification [16]. Where training sets are used while validation set is kept separated from the training data. Therefore, the wrapper method is very slow as they search several combinations of genes and optimal parameter set and certainly adds excessive computational complexity. The embedded method performs the selection of genes during the training procedure and is specific to the particular learning algorithms [17]. This paper concatenates on the filter method, where the selected two machine learning methods for finding the differentially expressed genes are T-test & CS.

1) 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 Mean Difference and then normalizing it as illustrated in (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 [6].

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

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

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 T-test for gene i can be described from (3) to (7) [6].

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

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 of gene i.

2) Class Separability (CS):CS of gene i is defined as:

$$CS_i = SB_i / Sw_i \quad (8)$$

$$SB_i = \sum_{i=1}^{k} (\bar{x}_{ik} - \bar{x}_i)^2 \quad (9)$$

$$SW_i = \sum_{k=1}^{k} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \quad (10)$$

$$\bar{x}_{ik = \sum_{j \in C_k}} \bar{x}_{ij} / n_k \quad (11)$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n \quad (12)$$

Here SBi is the sum of squares of between-class distances (the distances between samples of different classes). SWi 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 K classes.C_k refers to class k that includes nk samples. $x_{ij}$is the expression value of gene i in sample j. $\bar{x}_{ij}$ 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. CS is calculated for each gene. A larger CS indicates a larger ratio of the distances between different classes to the distances within one specific class. Therefore, CS is used to measure the capability of genes to separate different classes [9].

B. Machine Learning Classifiers

The most important application of microarray in gene expression analysis is to classify the unknown tissue samples according to their gene expression levels with the help of known sample expression levels. The small number of samples and the level of noise make the classification task of a test of challenge. In the following, two machine learning classifiers (KNN and SVM) are presented.

1) K- Nearest Neighbor (KNN): KNN is the simplest machine learning technique for classifying objects based on closest training examples in the feature space [5]. It is instance based learning. It gathers all training data and classifiers often via a majority voting, a new data point with respect to the class of its k-nearest neighbor in the given data set. KNN obtain the neighbors in the given data set. KNN obtain the neighbors for each data by using Euclidian or Mahalanobis distance between pairs of data items. Then, assign a class label to a new sample where the majority of the chosen number of neighbors belongs. Although being a simple technique, KNN shows an outstanding performance in many cases of classifying microarray gene expression. 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 [3].

2) Support Vector Machine (SVM): SVMs are widely used in many machine learning and data mining problems due to the superior performance in data analysis. The SVM algorithm is a supervised learning technique, because they exploit prior knowledge of gene to identify unknown genes. It finds the optimal hyperplane, which maximizes the minimum distance from the hyperplane to the closest training points. This feature makes SVM a powerful tool that has been used in gene expression data analysis [5]. Actually, SVM model is a representation of the samples as points in space, mapped so that the samples of the separate categories are divided by a separate area. New samples are then mapped into that same space and predicted to belong to a category based on which area they fall on. SVMs have many mathematical features that make them attractive for gene expression analysis, including their flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and to identify outliers [6,7].

The structure of SVM depends on kernel functions, where the most commonly used are liner and polynomial. If there are more than two classes in the data set, binary SVMs are not sufficient to solve the whole problem. To solve multi-class classification problems, the whole problem should be converted into a number of binary classification problems. Usually, there are two approaches [7]. One is the "one against all" scheme and the other is the "one against one" scheme. In "one against all", if there are N classes in the entire data set, then N independent binary classifiers are built. Each binary classifier is in charge of picking out one specific class from all the other classes. For one specific pattern, all the N classifiers are used to make a prediction. The pattern is categorized to the class that receives the strongest prediction. The prediction strength is measured by the result of the decision function.

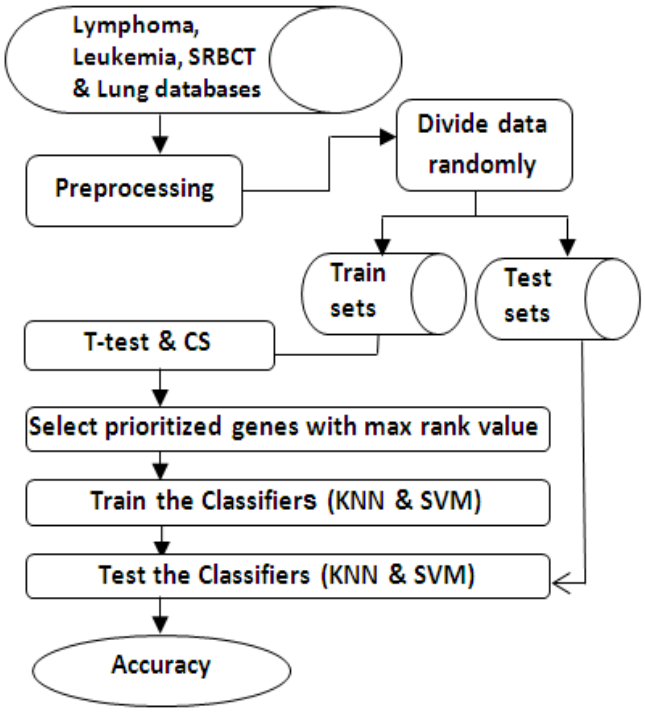

Fig. 1. Proposed scenario for gene expression data mining workflow

For the "one against one" scheme, there must be one (and only one) classifier taking charge of the classification between any two classes. Therefore, for a data set with K classes, K(K−1)/2 binary classifiers are used. To get the ultimate result, a voting scheme is used. For every input vector, all the classifiers give their votes so there will be K(K−1)/2 votes, when all the classification (voting) finished, the vector is mapped to the class getting the highest votes. If a vector gets highest votes for more than one class, it is randomly designated to one of them [7,10]. In our practice, we choose the "one against one" scheme for database with multiclass.

C. Divcovery Workflow

Actually, to demonstrate the computational results detailed in next section, we divided the effort of this study into two main consequent lines.

In the first line,we ran all combination of applying the gene filter techniques (T-test & CS) and two ML classifiers (KNN & SVM) on four gene expression databases. Fig. 1, shows the mining workflow. First, a T-test was applied then evaluated using a KNN classifier on the three gene expression databases. Second, the prioritized genes by T-test were evaluated one more time using SVM to analyze the effect of the classifier technique on the classification accuracy. Third, CS was applied on the identical databases and evaluated by the(KNN & SVM) classifiers to analyze the effect of using different prioritized genes by different filter techniques on classification accuracy. Finally, concluding the key results.

```
Input :X←full featured set of training data {x1,..,xn}
Y←test samples
N←Highest obtained classification accuracy of
T-test and CS separately
        M←control parameter of no of algorithm trails
Output : Subset1←T-test ranked features
Subset2←CS ranked features
        CommonSet←Subset1∩Subset2
T-test(X, Subset1);
CS(X, Subset2);
Confirm_Prioritized_Set (Subset1, Subset2, CommonSet)
While (C<=N) or (Count<=M)
{       Optimal(CommonSet, OptGenSet);
        Test(OptGenSet, C);
       Count ++;
}
```

Fig. 2. The Proposed hybrid reduction approach of microarray data

In the second line, as an improvement of the obtained results, we proposes a novel hybrid reduction approach for enhancing classification accuracy. A pseudo code of the proposed approach is show in Fig.2. From the figure, our proposed methodology starts by applying the T-test on the microarray data, where a ranked genes subset1 is obtained (ex: the first 100 prioritized genes). In the other hand a ranked genes subset2 is obtained from CS. A third reduction step is done by intersecting Subset1 and Subset2, where this step confirms the most important genes as a CommonSet. Then the confirmed CommonSetis searhed for the optimal genes set that enhance classification accuracy.

IV. Data Sets & Experimental Setting

For knowledge discovery in gene expression microarray data, an essential understanding of the nature of the data sets must be reached before the rest mining-workflow could proceed successfully. Hence, introductions of these databases with the experimental settings are presented in this section.

A. Microarray Genes Profile Data

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 Xij where; $0<i \leq ng$, $0<j \leq ns$ and ng, ns are the total number of genes, total number of samples respectively as in Fig. 3. [2].

Microarray data could be one of two types, paired and unpaired. Paired Data, is collected where two measurements from each patient, one before treatment and one after treatment. Then the difference between the two measurements (the log ratio) shows whether a gene has been up-regulated or down-regulated following that treatment.

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots \cdots \cdots & x1_{ns} \\ x_{21} & x_{22} & x_{23} & \vdots & x2_{ns} \\ x_{31} & x_{32} & x_{33} & \vdots & x3_{ns} \\ \cdots & \cdots & \cdots & \vdots & .. \\ \cdots & \cdots & \cdots & \vdots & .. \\ x_{ng1} & \cdots & \cdots & \vdots & x_{ngns} \end{bmatrix}$$

Fig. 3. Expression Data Matrix

Unpaired Data, is collected where two groups of patients with two or more classes exists. To identify the genes that is up- or down-regulated in unpaired data relative to the targeted classes (i.e., differentially expressed between the two classes are selected ex: based on their statistical p-value). Therefore, the smaller p-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.

B. Lymphoma dataset

The lymphoma dataset is downloaded from Lymphoma Molecular Profiling Project (LLMPP) webpage [http:// llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt].This dataset contains 4026 genes and 62 samples, 42 samples derived from diffuse large B-cell lymphoma (DLBCL), 9 samples from follicular lymphoma (FL), and 11 samples from chronic lymphocytic lymphoma (CLL). The lymphoma dataset downloaded consist of noisy and inconsistent data based on its available description, ex: some additional unnecessary columns exist, and after a deep study of its important columns needed to precede our work, which are (Gene ID, Name, Class Label (DLCL, FL, CLL)), we removed such unnecessary data. Also, we found many cells values equal zero, and although we concerned reflect of such values on the classifier, but many references of our related work kept these zeros values without concern [5,6,8]. Finally, the treatment of missing attribute values (empty string), where we imputed these missing values using KNN impute technique (Matlab), where this technique replaces such data with the corresponding nearest neighbours columns and if that value is also missing, it go further to the next nearest column and so on until the treatment is achieved. Table: 1 show a sample of the Lymphoma data, where the cells in bold are the ones that were missing and then their values are imputed after pre-processing.

C. Leukemia dataset

This dataset is downloaded from the web site [http://www.broadinstitute.org/cgi-bin/cancer/publications/ pub_paper.cgi?mode%20=%20view&paper_id=43].It contains of 7129 genes and 72 samples (47 the acute lymphoblastic leukemia (ALL) samples and 25 the acute myeloid leukemia (AML) samples). The original Leukemia data was already divided into training and testing sets. There are totally 38 training samples and 34 testing samples. The 38 training samples contain 27 ALL and 11 AML. Also, the 34 testing samples contain 20 ALL and 14 AML. Actually, the downloaded leukemia dataset is already partially preprocessed where no noisy or inconsistent data exists. The available description of the leukemia dataset showed that, the only preprocessing task needed is normalization for its values to reduce the systemic bias introduced during experiments. A sample from the data is shown in Table 2.

D. The SRBCT dataset

This dataset is downloaded from [http://research.nhgri.nih .gov/microarray/Supplement/]. The SRBCT dataset is pre-divided into training and testing sets on their web site. It contains 2308 genes and 88 samples. There are totally 63 training samples and 25 testing samples. Based on its formal description, five of the testing samples doesn't belong to SRBCTs and therefore are recognized as a noisy data. These unnecessary columns are (Test 3, Test 5, Test 9, Test 11 and Test 13) [18]. The 63 training samples contain 23 Ewing families of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 neuroblastoma (NB), and 8 Burkitt lymphomas (BL). The 20 SRBCTs testing samples contain 6 EWS, 5 RMS, 6 NB, and 3 BL. A sample from the data is shown in Table 3.

TABLE I. SAMPLE OF PRE-PROCESSED LYMPHOMA DATASET

No	Gene ID	Name	Value DLCL	Value DLCL	Value DLCL	Value DLCL	Value FL	Value FL
22	GENE3069X	Clone=1340681	-1.88	**-3.3**	-2.39	**-3.3**	-0.66	**-0.66**
23	GENE2584X	Clone=1317515	-0.32	0.08	-0.08	-024	0.34	**0.34**
24	GENE3070X	Clone=1355987	-0.18	**-0.5**	-0.47	0.48	0.06	-0.14
25	GENE1843X	Clone=1268758	-0.22	0.23	**0.18**	0.52	-0.18	0.24
26	GENE3166X	Clone=1317098	-0.65	-0.26	**-0.05**	0.07	0.53	0.19
27	GENE3165X	Clone=1339226	-0.25	-0.08	**-0.32**	**0.23**	-0.12	0.1

TABLE II. 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

TABLE III. A SAMPLE DATA FORM 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 IV. A SAMPLE DATA FROM LUNG CANCER DATASET

No	Gene ID	Values (MPM)	Values (MPM)	Values (ADCA)	Values (ADCA)
2	1000_at	214.9	249.6	60.3	202.3
3	1001_at	116.7	32.2	54	61.5
4	1002_f_at	8.4	15.2	32.6	-19.6
5	1003_s_at	-79.8	-40	-222.7	-172.4
6	1004_at	-0.3	15.3	64	18.1

E. Lung Cancer Dataset

This dataset is downloaded from the web site [http://datam.i2r.astar.edu.sg/datasets/krbd/LungCancer/LungCancer-Harvard2.html]. It contains of 12533 genes and 181 samples (31 the malignant pleural mesothelioma (MPM) samples and 150 an adenocarcinoma (ADCA) samples). The original Lung Cancer data was pre-divided into training and testing. There are totally 32 training samples and 149 testing samples. The 32 training samples contain 16 MPM and 16 ADCA. Also, the 149 testing samples contain 15 MPM and 114 ADCA. A sample from the data is shown in Table 4.

V. RESULTS AND DISCUSSION

In this section, the experimental results are presented to establish the contribution of each factor used during the mining task (Phase 1: gene ranking &Phase 2: classifiers). We have conducted numerous assessments of the proposed mining workflow on four public microarray databases (Lymphoma, Leukemia, SRBCT and Lung).We have implemented it in MATLAB 7.11 (R2010b) in Windows 7 running on a PC with system configuration Intel Core 2 Duo processor (2.40 GHz) with 3 GB of RAM. Actually, to more accurately compare the performances of our applied machine learning approaches for mining microarray data, we have utilized the strategy to select genes only from the training samples. The testing samples are totally excluded from the classifier building process.

A. Phase 1: Gene Ranking and Dimensionality Reduction

1) Lymphoma dataset:We divide the lymphoma data set randomly into a three cases of subsets to study the effect of different scenarios for selecting the training and testing samples side by side with different numbers of DEGs supplemented to the classifiers on the classification accuracy. In the following, the three cases are described separately in table 5.We rank the genes by using the T-test based on their statistical score and the first 15 DEGs and their corresponding t-test values for case1, case2 and case3 is shown in table 6.

TABLE V. THE PERCENTAGE OF TRAINING AND TESTING SAMPLE FOR CASE 1, CASE 2 AND CASE3

Case 1		Case 2		Case 3	
Training	**Testing**	**Training**	**testing**	**training**	**Testing**
50% (31) samples	50% (31) Samples	60% (37) samples	40% (25) samples	75% (47) samples	25% (15) samples

2) The SRBCT Database:Based on T-Test ranking, Table 7, shows the20 top informative 20 with their corresponding statistical values.

3) The Leukemia Database: both T-test and CS were applied o Leukemia to study the effect of different gene ranking techniques on classification accuracy. The 20 top informative genes using T-test and CS are show in table 8 with their corresponding statistical, respectively.

B. Phase 2: Classifiers

1) The SRBCT Database:it is a multiclass datase.SVM classifier deals with multiclass database in two ways, "One against one" and "one against all", we applied "one against one" on SRBCT. The SEBCT dataset originally has four classes which are EWS, BL, NB and RMS. Therefore, applying the formula of finding the number of binary SVM classifiers K(K-1)/2, result in 6 binary classifiers. These classifiers are shown in Fig. 4 in comparison with results of applying KNN classifier on the identical testing set. Actually, during the voting scheme for obtaining the single SVM classifier, we discovered that from the six implemented and tested binary SVM classifiers of SRBCT in Fig. 4, only two classifiers (EWS & NB and BL & RMS) cover all the testing samples and hence are combined to get the average classification accuracy.

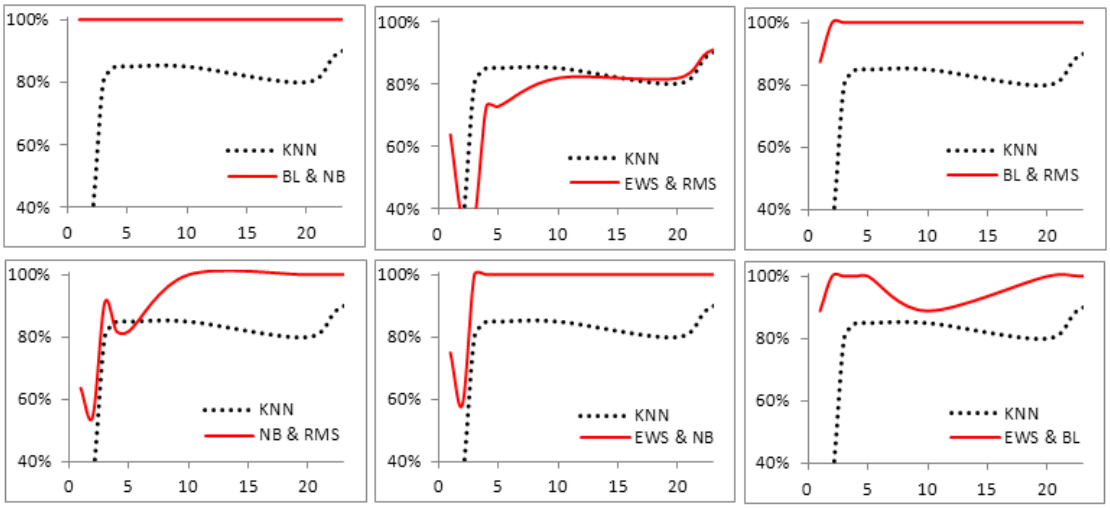

Fig. 4. six binary SVM classifiers versus KNN classifier on SRBCT database

TABLE VI. TOP 15 PRIORITIZED GENES AND THEIR CORRESPONDING T-TEST FOR CASE1,CASE2 & CASE3.

No	Results of Case 1		Results of Case 2		Results of Case 3	
	Gene-ID	***T-test***	***Gene-ID***	***T-test***	***Gene-ID***	***T-test***
1	Gene 639X	11.18238	Gene 653X	10.74554	Gene 653X	12.80793
2	Gene 653X	10.89338	Gene 639X	10.08582	Gene 708X	11.63526
3	Gene 769X	9.592778	Gene 563X	10.07219	Gene 699X	10.8992
4	Gene 642X	9.566121	Gene 708X	9.779857	Gene 704X	10.56475
5	Gene 2374X	8.826267	Gene 537X	9.249515	Gene 563X	10.40575
6	Gene 708X	8.775379	Gene 769X	9.091355	Gene 675X	10.33056
7	Gene 563X	8.751787	Gene 699X	8.963873	Gene 709X	10.2826
8	Gene 652X	8.667277	Gene 2203X	8.876297	Gene 706X	10.13764
9	Gene 709X	8.598637	Gene 675X	8.812909	Gene 537X	10.01828
10	Gene 704X	8.489373	Gene 704X	8.801415	Gene 639X	9.981593
11	Gene 705X	8.445308	Gene 705X	8.66119	Gene 700X	9.953322
12	Gene 2395X	8.434764	Gene 2374X	8.654025	Gene 771X	9.862814
13	Gene 2391X	8.287241	Gene 1646X	8.554017	Gene 651X	9.640458
14	Gene 721X	8.228957	Gene 700X	8.530127	Gene 2391X	9.639435
15	Gene 711X	8.193166	Gene 2395X	8.484026	Gene 540X	9.610315

TABLE VII. SRBCT 20 TOP INFORMATIVE GENES BASED ON (T-TEST)

No	Gene ID	T-test Value	No	Gene ID	T-test Value
1	236282	13.72563	11	745019	9.340171
2	183337	11.4937	12	609663	9.016319
3	812105	10.97844	13	325182	8.677164
4	770394	10.4634	14	784224	8.443444
5	814526	10.26562	15	68977	8.306498
6	624360	10.25416	16	769657	8.236335
7	1469292	10.02259	17	740604	8.184522
8	47475	9.939925	18	344134	8.023031
9	241412	9.790006	19	283315	7.99332
10	767183	9.471839	20	383188	7.989644

Fig. 5: shows the testing accuracy of the obtained SVM classifier in comparison with applying the KNN classifier on the same SRBCT testing samples. The figure showed that SVM recorded better classification accuracy than the KNN classifier.

2) Lymphoma dataset: For the lymphoma dataset and for each of the previously explained three cases, we applied the KNN classifier. Fig. 6: shows the testing classification accuracy for the three cases. From Fig. 6, if the highest classification accuracy is a target, then dividing training and testing subsets such case 3 is the way out, where for this dataset it reached 100% using 52 first most informative genes. But, when concerning the number of submitted genes side by side with the classification accuracy, then dividing training and testing subsets such case1 is recommended, where it reached around 85% with very few genes (less than 5 genes). For overall average classification accuracy, case 2 recorded more stability relative to changing the number of DEGs submitted to the KNN classifier.

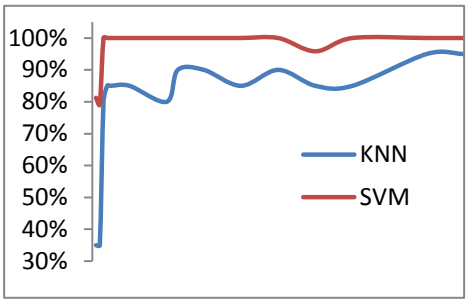

Fig. 5. Classification accuracy of KNN versus SVM classifiers on SRBCT

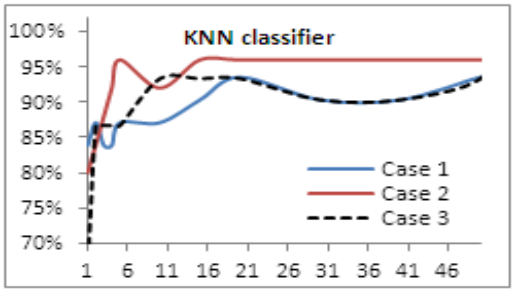

Fig. 6. KNN Testing Classification Accuracy of Cases 1,2,3 of lymphoma

TABLE VIII. LEUKEMIA 20 TOP PRIORITIZED GENES BASED ON (T-TEST & CS)

Ranked 20 genes by T-Test			Ranked 20 genes by CS	
No	**Gene-ID**	**T-test**	**Gene-ID**	**CS-value**
1	U50136_rna1_at	6.584	M55150_at	8.091
2	X95735_at	6.435	U22376_cds2_s_at	7.904
3	M55150_at	6.177	X59417_at	6.803
4	M16038_at	5.493	U50136_rna1_at	6.435
5	Y12670_at	5.488	M31211_s_at	6.293
6	M23197_at	5.387	L13278_at	6.281
7	D49950_at	5.172	U82759_at	6.268
8	X17042_at	5.042	M92287_at	6.217
9	U82759_at	5.005	U05259_rna1_at	6.181
10	M84526_at	4.952	U12471_cds1_at	6.146
11	L08246_at	4.789	U09087_s_at	6.120
12	Y00787_s_at	4.787	D26156_s_at	6.097
13	M80254_at	4.7617	X74262_at	6.016
14	U46751_at	4.7423	M81933_at	5.933
15	M27891_at	4.643	X95735_at	5.805
16	M62762_at	4.608	M28170_at	5.794
17	M63138_at	4.498	L47738_at	5.733
18	M28130_rna1_s_at	4.480	AF009426_at	5.693
19	M81695_s_at	4.414	M31523_at	5.677
20	X85116_rna1_s_at	4.338	S50223_at	5.676

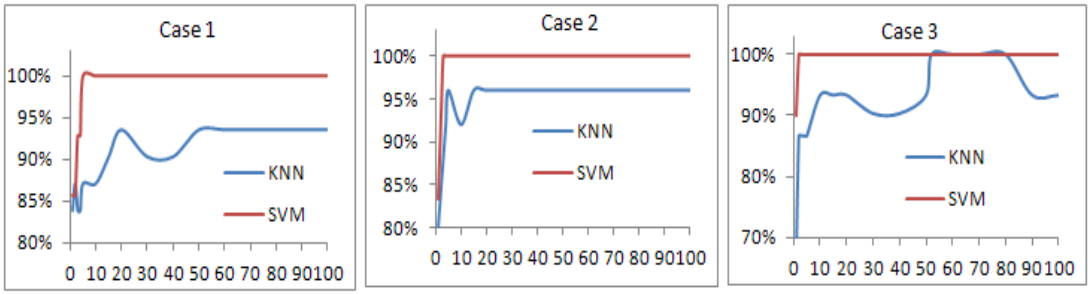

Fig. 7. Classification accuracy of KNN versus SVM on Lymphoma three cases

Actually, lymphoma is also a multiclass database with three classes (DLCL, FL and CLL), where three binary SVM classifiers (DLCL & FL, DLCL & CLL, FL &CLL)have been implemented. In a similar manner, one SVM classifier was chosen for each case of the three cases. Fig. 7: shows the classification accuracy of KNN versus SVM classifiers of every case. The figure shows that although obtaining the SVM classifier for multiclass data set took more computational effort, the SVM recorded better classification accuracy than the KNN classifier for the three cases.

3) The Leukemia Database: Fig. 8: shows the classification accuracy of KNN versus SVM classifiers for both gene ranking T-test and CS on leukemia database. The figure shows that, with few numbers of DEGs (less than 15 submitted to KNN and less than 30 genes submitted to SVM), the T-test reported better classification accuracy than CS as a ranking technique. But with greater than mentioned number of DEGs for each classifier, the CS recorded better classification accuracy.

The proposed hybrid reduction approach was applied on leukemia and resulted with CommonSet that contains 24 confirmed prioritized genes in Table 9. These common 24 genes were then searched for the optimal subset that intended to be submitted to the classifiers by applying combination and permutation. Actually, after building the first 24 SVM classifiers, where a single gene was tried at a time, the hybrid approach could reach the highest classification accuracy (94.12%) and (100%), by submitting only one gene (ranked No 6 in our CommonSet and named (M23197-at)) and integrated with SVM and KNN, respectively. Please note that (94.12%) classification accuracy was recorded before without using the proposed reduction approach and instead using the integration of (T-test+SVM) but with Subset of top ranked 30 genes instead of only one gene in our proposed reduction approach.

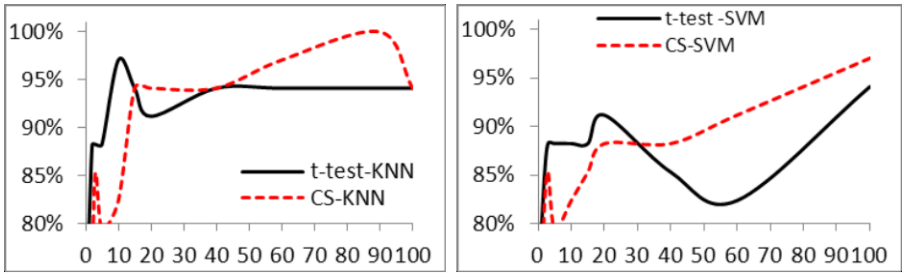

Fig. 8. T-test versus CS for both KNN & SVM classifiers on Leukemia

TABLE IX. THE24 COMMONLIST OF LEUKEMIA CANCER AND LUNG CANCER

No	Gene Name		No	Gene Name	
	Leukemia	*Lung*		*Leukemia*	*Lung Cancer*
1	U50136_rna1_at	37205_at	13	M11147_at	1030_s_at
2	X95735_at	32046_at	14	X04085_rna1_at	709_at
3	M55150_at	2047_s_at	15	M81933_at	33327_at
4	M16038_at	38482_at	16	U22376_cds2_s_at	36369_at
5	Y12670_at	37716_at	17	M86406_at	32551_at
6	M23197_at	41286_at	18	M21551_rna1_at	35822_at
7	D49950_at	40936_at	19	X15414_at	40496_at
8	X17042_at	34320_at	20	X52142_at	33328_at
9	U82759_at	33245_at	21	X59417_at	39756_g_at
10	M80254_at	39409_at	22	M31211_s_at	291_s_at
11	M62762_at	33833_at	23	D26156_s_at	34329_at
12	U12471_cds1_at	41755_at	24	L13278_at	37027_at

TABLE X. SVM & KNN CLASSIFICATION ACCURACY USING T-TEST

Data Set	Results with SVM			Results with KNN	
	(Linear SNM)	*No. of Genes*		*Accuracy (K = 1)*	*No. of Genes*
Leukemia	94.12%	6 (T-test)	70 (CS)	97.06%	10 (T-test)
				94.12%	15(CS)
Lymphoma	100% (Cases)	2 (T-test)		96% (Case2)	15 (T-test)
SRBCT	100%	3 (T-test)		90%	23 (T-test)

TABLE XI. COMPARISON OF LEUKEMIA CLASSIFIERS VERSUS GENES NUMBER

Authors	Accuracy	No. of genes
D Mishra, B Sahu, [19]	98.1%	20
Our T-tes + KNN	97.06%	10
Our T-test + SVM	94.12%	6
Our hybrid reduction approach +SVM	94.12%	1
Our hybrid reduction approach +KNN	100%	1

TABLE XII. RESULTS FOR THE SRBCT DATA SET OBTAINED BY DIFFERENT APPROACHES

Method	Accuracy	No of genes	Method	Accuracy	No of genes
MLP neural network [18]	100%	96	FNN [9]	95%	3
Nearest shrunken centroids [20]	100%	43	SVM (polynomial p=2) [7]	100%	6
Evolutionary algorithm [21]	100%	12	Our KNN	90%	23
SVM [22]	100%	20	Our SVM	100%	3

4) The Lung Cancer Database: the proposed hybrid approach reached (98.65%) with SVM using also one gene only (the ranked No 20 in the CommonSet in table 9 and named (33328_at)). Also the same classification accuracy was reached using (T-test & SVM) and without the proposed reduction approach but with subset of 16 genes. In addition, the proposed approach with KNN classifier, reached 97.31% using only one gene with rank No= 5, named (37716_at).

TABLE XIII. COMPARISON OF LYMPHOMA CLASSIFIERS

Authors	Accuracy	Number of genes
Dina A. et, al. [23]	94.59%	11
RBF SVM [7]	100%	5
Our KNN	96%	5
Our SVM	100%	2

C. Comparisons & Discussion

From table 10, 11 and Fig. 8, it is obvious that for the ***leukemia*** dataset, SVM reached 94.12% by 6 and 70 top prioritized genes by using T-test and CS ranking techniques, respectively. In addition integrating (T-test & KNN) recorded a test classification accuracy of 97.06% using 10 genes. But integrating (T-test and KNN) recorded 94.12% with gene Set=15. Therefore, based on the results (KNN+T-test) recorded higher classification accuracy. Also from table 10, the proposed approach recorded remarkable classification accuracy relative to the number of genes in comparison with almost D Mishra, B Sahu, [19].

For the ***SRBCT*** dataset, integrating (T-test& SVM), recorded 100% accuracy with 3 genes. Also, integrating (T-test& KNN) recorded 90% with 23 genes. Based on that, it is clear that SVM classifier achieved higher results on SRBCT. Actually, from table 12: and in comparison with some very related work on the same SRBCT dataset, it can be concluded that among many applied mining methods, integrating (T-test and SVM) classifier recorded best results in both classification accuracy and number of selected genes (in this case=3).

Form table 10, for the ***Lymphoma*** dataset, integrating (T-test & SVM) reached 100% in case3 with only 2 genes. Also, integrating (T-test and KNN) records 96% using 15 genes in case2. Therefore, integrating (T-test and SVM) recorded better

mining results in comparison with other listed mining approaches in table 13 in terms of both classification accuracy and number of genes.

VI. Conclusion& Future Work

Biological data is known to be with a huge size; therefore mining this data is a very important research area as it deeply reflects the drug discovery, diseases diagnosis and treatment. Classifying the cancer into a predefined class based on microarray expression datasets is divided into two main phases. Phase 1 is implementing an effective gene ranking technique to reduce the number of genes involved in the classification process. Phase 2 is adjusting a powerful classifier to achieve accurate classification accuracy for new unclassified samples.

This paper presented a novel hybrid machine learning (ML) reduction approach to enhance cancer classification accuracy of microarray data based on two ML gene ranking techniques (T-test and CS). The proposed approach was integrated with two ML classifiers; KNN and SVM; for mining microarray gene expression profiles. Four public cancer microarray databases were used for evaluating the proposed approach and successfully accomplish the mining process. These were Lymphoma, Leukemia SRBCT, and Lung Cancer. The strategy to select genes only from the training samples and totally excluding the testing samples from the classifier building process was utilized for more accurate and validated results. Also, the computational experiments were illustrated in details and comprehensively presented with literature related results. Actually, integrating (T-test+SVM) recorded higher classification accuracy than the mining integrated approaches (T-test+KNN, CS+SVM, CS+KNN), where it recorded a test classification accuracy of 100% using the highest ranked 2 and 3 genes for Lymphoma and SRBCT, respectively. It also recorded 94.12% using the highest ranked 6 genes for Leukemia. The results showed that the proposed reduction approach reached promising results of the number of genes supplemented to the classifiers as well as the classification accuracy in comparison with literature similar mining approaches for microarray data.

Our future work intends to apply the proposed hybrid reduction approach on more microarray data for confirmation and verification of it performance. Also, more classifiers and ranking techniques will be studied.

Acknowledgment

The authors would like to thank Prof.Abdel Badeeh M.Salem and Prof.El-Sayed M.El-horbaty for their help, care and advices during this research.

References

[1] John N. Weinstein, et.al. The Bioinformatics of Microarray Gene Expression Profiling. Wiley-Liss, Inc,pp. 2001:46-49.

[2] Wolfgang Huber, Anja Von Hey debreck, Martin Vingron. Analysis of microarray gene expression data. Hand book of statistics genetics. 2nd edition, Wiley. 2003.

[3] Joseph S. Verducci, et.al. a Microarray analysis of gene expression: considerations in data mining and statistical treatment. Physiol. Genomics. 2006; 25(3):pp.355-363.

[4] Wang, Y., Tetko, I. -V., Hall, M. -A., Frank, E., Facius, A., Mayer, K. -F., andMewes H. -W. Gene Selection from Microarray Data for Cancer Classification—A Machine Learning Approach.ComputBiolChem.2005; 29(1):37-46.

[5] Jiawei Han, MichelineKamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, An Imprint of Elsevier, First Indian Reprint. 2001.

[6] Abeer M. Mohamed, BasmaA.Maher, El-SayedM.El-horbaty& Abdel-Badeeh M. Salem. 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.

[7] F. Chu & L. Wang.Applications of Support Vector Machines To Cancer Classification With Microarray Data. International Journal of Neural Systems. 2005; 15(6):475–484.

[8] Xin Jin, A.Xu, B.Rongfang&P.Guo. Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles.Springerlink, Data Mining for Biomedical Applications. 2006; 3916:106-115.

[9] L.Wang, F.Chu, &W.Xie. Accurate Cancer Classification Using Expressions Of Very Few Genes. IEEE/ ACM Transactions on Computational Biology and Bioinformatics. 2007; 4(1):40-53.

[10] M.Rangasamy&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.2009; 2(2):51-66.

[11] M. Martin-Merino &J.d.l.Rivas. Kernel Alignment k-NN for Human Cancer Classification Using the Gene Expression Profiles. Springer link, Artificial Neural Networks – ICANN. 2009; 5769:195-204.

[12] N. Revathy& R. Amalraj, Accurate Cancer Classification Using Expressions Of Very Few Genes.Int. J. of Computer Applications. 2010;14(4):19-22.

[13] Z.Ghorai, et.al. Cancer Classification From Gene Expression Data By NPPC Ensemble. IEEE Transactions on Computational Biology & Bioinformatics. 2011; 8(3):659-671.

[14] Abeer M. Mohamed, BasmaA.Maher, El-SayedM.El-horbaty& Abdel-Badeeh M. Salem. Applying a Statistical Technique for the Discovery of Differentially Expressed Genes in Microarray Data. Proc. Of recent advances in circuits systems, telecommunications and control, France. 2013 :220-227.

[15] Lai C, Reinders MJ, van'tVeer LJ, Wessels LF. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics. 2006; 7:235-253.

[16] Blanco R, Larranaga P, Inza I, Sierra B. Gene selection for cancer classification using wrapper approaches. International Journal of Pattern Recognition and Artificial Intelligence. 2004; 18(8):1373-1390.

[17] H. Peng, F. Long, and C. Ding. Feature Selection on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. Pattern Analysis and MachineIntelligence. 2005; 27(8):1226-1238.

[18] J.M. Khan et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine. 2001; 7:673-679.

[19] D Mishra, B Sahu. Feature selection for cancer classification: a signal-to-noise ratio approach. International Journal of Scientific & Engineering Research. 2011; 2(4).

[20] 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. 2002; 99(10):6567-6572.

[21] J. Deutsch. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics. 2003; 19(1):45-52.

[22] Y. Lee and C.K. Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics. 2003; 19(9):1132-1139.

[23] 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. 2011; 21:36-41.