A New Gene Selection Technique Based on Hybrid Methods for Cancer Classification Using Microarrays

Dina A. Salem, Rania Ahmed A. A. Abul Seoud, and Hesham A. Ali

***Abstract*—Classification is one of the most important data mining techniques especially in the area of bioinformatics. This can be clearly seen in cancer classification which is recently addressed by many researchers specially after emerging of microarrays. This technology opens the area for computer researchers to classify cancer samples without any previous biological knowledge. Microarrays high dimensionality problem forces scientists to design gene selection techniques as a preceding step to the implemented classifier. Gene selection techniques behavior varies according to the combined classifier. In this paper we are proposing a new gene selection technique which combines F-score and entropy-based methods. The output of the combined gene selection technique is fed into two different classifiers resulting in two hybrid cancer classification systems. The proposed systems achieved reliable classification accuracies when tested on two different microarray datasets. This validates the success of the proposed gene selection technique as it efficiently reduced the original number of genes by 71.29%.**

***Index Terms*—Bioinformatics, Classification, Data mining, Gene selection, Microarrays.**

I. Introduction

Data is tremendously growing in all life aspects resulting in mountains of data. Mining these mountains using powerful data analysis tools is important to obtain the contained valuable information needed to present decision making solutions. 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 [1]–[3]. Data mining owns a variety of methods to achieve its two main data analysis goals named description and prediction. Descriptive methods which focus on finding patterns that describes the data to be interpreted by humans include summarization, change and deviation detection, and clustering, whereas predictive methods which involve using some variables or fields in the data set to predict unknown or future values of other variables of interest include regression and classification.

As classification is an important supervised machine learning technique works on classifying a new data item into a predefined class [4], it will be the main concern of our work. Data mining with all its powerful tools is an important consequence of the natural evolution of information technology and so, it is efficiently applied to almost all computerized fields of life resulting in powerful and reliable solutions for thousands of problems [5].

One of the well known data mining application is in the companies with a strong consumer focus - retail, financial, communication, and marketing organizations - where the principal objective is to reduce cost and increase revenue. Explanation and examples are found in [6]. Homeland security is another important application where data mining is often viewed as a potential means for identifying terrorist activities. Two initiatives that have attracted significant attention include the Terrorism Information Awareness (TIA) project conducted by the Defense Advanced Research Projects Agency (D ARPA), and the Computer-Assisted Passenger Prescreening System II (CAPPS II) that was being developed by the Transportation Security Administration (TSA) and then replaced by a new program called Secure Flight [7]. Lots of other applications are available, but the application which is of a great interest to human beings is the bioinformatics. Bioinformatics can be defined as the application of computer technology to the management of biological information. In bioinformatics, data mining has a primary goal in increasing the understanding of biological processes. Some of the grand areas of research in bioinformatics are analysis of gene expressions and mutations in cancer. Other areas of research are highlighted in [8]. Cancer databases and gene expression values are the data used in this paper and is extracted from the emerging microarray technology.

High-throughput microarray technology is a hybridization procedure that enabled the simultaneous measurement of the abundance of tens of thousands of gene-expression levels from many different samples on a small chip. Microarray data takes the form of a huge m*n matrix, where m (rows) represents the genes, n (columns) represents the samples and each of its cells contains an expression value for a gene in a sample [9]. The large amount of data this matrix holds makes it in a deep need for data mining. Microarray data is mainly used in cancer diagnosis and prognosis where it's well known that the early diagnosis of cancer and determining its type is very helpful in its treatment. Data mining classification techniques are very suitable to address this issue where new samples can be classified into two or more predefined cancer classes. Microarray data is characterized by its high dimensionality which means that the number of samples (always less than 100) is not proportional to the number of genes (always thousands). This explains the need for a feature selection technique before entering the data into the classifier. The feature in the microarray data is the gene and the feature selection is renamed to be a gene selection [10].

Then, the process of classifying microarray data must

Manuscript received November 9, 2011; revised November 22, 2011.

D. A. Salem is with the Department of Computer, Faculty of Engineering, Misr University for Science and Technology, Egypt (e-mail: dena.salem @mail.com).

R. A. Abul Seoud is with the Department of Electrical Eng.-Comm. and Electronics Section-Faculty of Eng - El Fayoum University-Fayoum, Egypt.

H. A. Ali is with the Computer Engineering System Department, Faculty of Engineering-Mansoura University, Egypt.

enclose two main steps; implementing an effective gene selection technique and choosing a powerful classifier. These two steps will form the workflow of this paper trying to go through each step details and validating its outputs. Then, two hybrid classification systems are proposed, both of them employ the same proposed gene selection technique but each one uses a different classifier. One of the proposed systems achieved the highest classification accuracies on the two used datasets with a considerable number of genes. Although the other proposed system couldn't reach the same results, it contributes in validating the proposed gene selection technique.

The remainder of this paper is organized as follows; Section II reviews briefly some of the recent work published in the area of classification of cancer using microarray gene expression values. Section III introduces and describes the general scheme of our proposed combined data mining technique. Results of the proposed technique are presented in Section IV. Section V analyzes these results. Finally, Section VI concludes the paper.

II. Related Work

A lot of research has addressed the topic of the classification of the microarray data by using different gene selection methods with different classifiers. A generic approach to classifying two types of acute leukemias was introduced in Golub et al. [11]. Two other systems used for classifying the same 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. Both systems result in effectual and powerful classification of gene expression data [12, 13]. SVM is used again by Moler *et al.* but this time combined with a naive Bayesian model for classifying the colon adenocarcinoma tissue specimens labeled as tumor or nontumor dataset for the first time [14]. The two previous datasets were used by P. Yang and Z. Zhang 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) [15]. 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 preselection step. To validate their system they used two classifiers; k-nearest neighbor (KNN) and SVM on five datasets [16].

Another group of researchers concentrate on the comparative studies. S. Deegalla and H. Bostrom used KNN classifier to compare four different high dimensionality reduction methods on eight public microarray datasets [17]. S. Cho and H. Won carried out a broad research using seven gene selection techniques, four classifiers and three combining methods resulting in 42 ensemble classifiers and applied them to three public datasets [18]. A similar study using fourteen gene selection techniques on simulated data was carried out in [19]. More and more research is needed in this area to improve the classification accuracy of different microarray datasets and hopefully, to reach system biologists can use to classify any new sample with minimum number of gene expression values.

The expression value of thousands of genes is the key to specify the class of a sample. Using this big number of genes to classify a new sample is time consuming and may reduce classification accuracy [20]. Most of the previous work uses only one gene selection technique to reduce the original number of genes. This means that they take a single gene criterion into their consideration according to the operation of the chosen technique. Sometimes this is not sufficient to obtain the highly informative genes. Some other work proposed a combined gene selection technique but connected to only one classifier. Here the problem is that some gene selection techniques work perfectly if combined with a specific classifier but very poor if combined with a different one. Another problem arises when a researcher decides to evaluate his proposed classification system on only one dataset. These researchers usually use the leukemia dataset which is known to be an easily classified microarray dataset.

Thus, it's a challenge to design a classification system which is capable of classifying new samples using a smaller highly informative gene subset of the original set of genes. At the same time it's a demand to result in high classification accuracy when tested on more than one microarray dataset. For this purpose, we designed a new gene selection technique which combines two univariate gene selection techniques for reducing the number of involved genes. Each one of them selects the informative genes according to a different criterion. Then this proposed gene selection technique is attached to two different classifiers working with two very different ideas (SVM and KNN). We end up with two classification systems which are evaluated on two cancer microarray datasets by recording the classification accuracy (CA) studying different parameters for each. One of the proposed systems achieved excellent and comparable results while the other one had a fair performance.

III. Methodology

A. System Description

The two proposed classification systems (First CS and Second CS) receive preprocessed high dimensionality microarray dataset as its input. The system first step is reducing the total number of genes in the input dataset to a smaller subset using F-score and entropy ranking techniques as a combined gene selection technique. Then the reduced data will be the data used by the chosen classifiers to assign new samples into their correct classes instead of using the original full data. At this point we can measure and record the test classification accuracy which is equal to the number of correct classified test samples divided by the total number of introduced test samples. The workflow of the proposed systems is shown in Fig. 1.

B. Microarray Gene Expression Datasets

When working with any classification system, any used dataset must be split into two sub-datasets; a training dataset which the classifier uses to learn and form its learned

structure and, a test dataset to see the effectiveness of the proposed system. The two proposed classification systems work on two public datasets and can be extended to classify other datasets. Table I contains the details of the two datasets.

One dataset is the leukemia dataset which was first classified by Golub et al. in 1999 into two classes; Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) [11]. The other dataset is the lymphoma dataset which was classified by Shipp et al. in 2001 into two classes; Diffuse Large B-Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL) [21]. Each sample in both datasets has expression patterns of 7129 genes measured by the Affymetrix oligonucleotide microarray. The two datasets are available and downloaded from Broad Institute of MIT and Harvard website (www.broadinstitute.org).

TABLE I: USED DATASETS DETAILS

Dataset	*Classes*	*Genes*	*Total samples*	*Train samples*	*Test samples*
Leukemia	AML,ALL(2)	7129	72	38	34
Lymphoma	DLBCL,FL(2)	7129	77	40	37

C. Gene Selection

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 overfitting (meaning that one can easily find a decision function that correctly classifies the training data but this function may behave very poorly on the test data), in addition to the ultimate increase in the processing power and time [22]. 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 [23].

Filter methods are characterized over the two other types by being powerful, easy to implement and is a stand-alone technique which can be further applied to any classifier. They are simply feature ranking methods; work on giving each gene a score according to a specific criterion and choosing a subset of genes above or below a specific threshold. Thus, they remove the irrelevant genes according to general characteristics of the data. However, it is not clear how to determine the optimal threshold for the data. One heuristic approach (the so called $n - 1$ rule) in microarray cancer analysis chooses the top $n - 1$ genes to start the analysis [24]. This is exactly how we chose the reduced gene subset from the two datasets to be classified to reach high classification accuracy. But instead of using only one filter technique, we use a combination of two efficient techniques; the F-score and the entropy-based. The 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 the classes [13]. The entropy-based technique ranks the subset of genes resulting from the F-score technique according to their entropy value. The combined technique is implemented in MATLAB 7.10.0 (R2010a) going through the following steps:

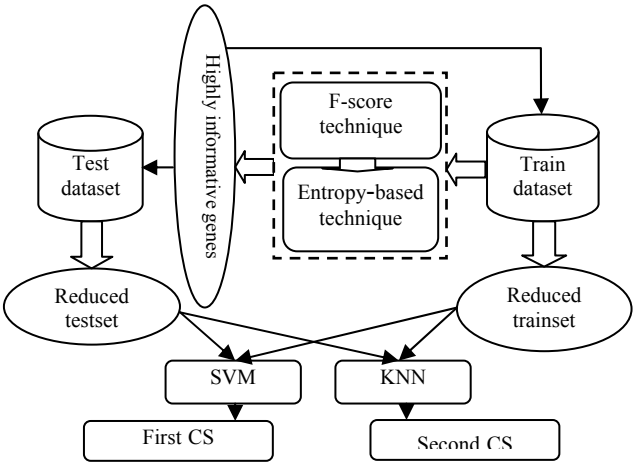

Fig. 1. Proposed systems workflow.

- Calculating the mean of the expression values for each of the n genes (μ_{n1} for the first class and μ_{n2} for the second class).
- Obtaining the absolute differences between the calculated means ($\left| \mu_{n1} - \mu_{n2} \right|$).
- Arranging the result in descending order.
- Selecting the top 250 genes.
- Calculating the SNR for each of the selected 250 genes ($F = (\mu_{n1} - \mu_{n2})/(\sigma_{n1} + \sigma_{n2})$).
- Selecting 200 genes with highest F-score.
- Ranking the selected 200 genes according to their entropy.
- Selecting the first 100 genes.

D. Classifiers

Support Vector Machine: Support Vector Machines (SVMs) have been widely used in the recent years in the field of computational biology due to their high accuracy and their flexibility in modeling diverse sources of data. They are mainly used in binary classification and regression. They are very suitable for classifying microarray gene expression data [25]. The SVM is partitioned into two parts; the SVM-trainer and the SVM-classifier. The SVM trainer must be preceded by a cross-validation analysis. Fig. 2 shows the general workflow of the implemented SVM classifier.

Cross-validation: Cross-Validation (CV) is very helpful in evaluating and comparing learning algorithms. It is a statistical technique used during the training process of the classifier where its task is to divide the train dataset into two segments; one is used for training and the other is used for validation. The training and validation sets must cross-over in successive rounds such that each sample has a chance of being validated against. CV is carried out in different forms where the most general form is the k-fold cross-validation. K-fold CV splits the data into k equal sized segments and then it carries out k iterations of training and validation. During every iteration, it holds out a different fold (segment) of the data for validation while the remaining k-1 folds are used for learning. One special form of the k-fold CV is the

leave-one-out-cross-validation (LOOCV) where it uses one sample for test and all other samples for training [26]. No agreement exists on a common value of k for CV in microarray data classification. As the number of samples used in the training of the classifier usually doesn't exceed several tens, we consider four values of k in our study (1, 2, 5, 10). Cross-validation is available as a function in the bioinformatics toolbox in MATLAB 7.10.0 (R2010a).

SVM-trainer: In this section SVM uses the train dataset to construct a hyperplane to separate two sets of data points (samples). It solves an optimization problem to reach the maximum margin. The maximum margin is the largest distance from the hyperplane to the nearest data points. Data points fall on this margin are called support vectors. This can be easily achieved for linear separable data points. Otherwise, SVM uses the kernel functions to map the non-linear separable samples into the feature space. Different kernel functions include; Gaussian, polynomial, and RBF. The output of the SVM-trainer is the SVM-structure.

SVM-classifier: In this section SVM uses the SVM-structure to classify the test data into the predefined classes. As the CV and SVM parameters are accurately chosen, as the classification accuracy of the test samples increases. The SVM-trainer and SVM-classifier are available in the bioinformatics toolbox in MATLAB 7.10.0 (R2010a).

The cross-validation coupled with the SVM-trainer runs several times in a continuous loop until reaching maximum train classification accuracy (correct rate=1). If the correct rate does not reach the value 1, the loop is manually stopped recording the highest value of the correct rate. Achieving highest train classification accuracy forms the optimum SVM structure which in turn leads to minimum misclassifications for the test samples. The SVM structure is the output of the SVM trainer. It contains all the information needed for the SVM classifier to classify the new test samples.

K-Nearest Neighbor: 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 neighbours. KNN main idea is to assign a new data item (sample) to the class to which the majority of the chosen number of neighbours belongs. Neighbors are determined by measuring the distances for KNN which 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 we must use 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 [27]. Different number of neighbors (k) will be used during the implementation process. The smallest k value is one which means that the new sample will be assigned to the same class of the first nearest neighbor. This is done after measuring the Euclidean distance between this test sample and all the samples in the reduced train dataset. Note that odd values of k are preferred so we use three small odd values (1, 3, 5) and one big even value (10).

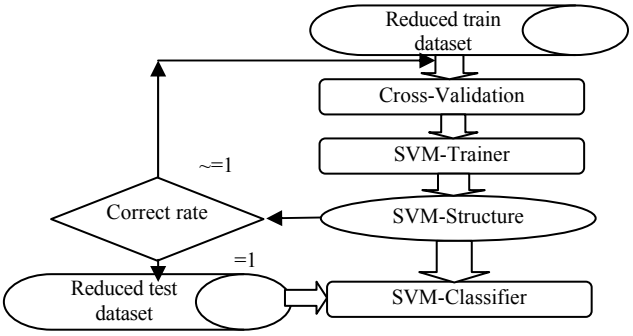

Fig. 2. SVM general scheme.

IV. Results

The accurate classification of microarray samples using the gene expression values in the proposed systems is an exhausting process. This happens due to the large number of attributes involved in each of its stages and the variety of values each attribute can take. But to fully evaluate our system in a comprehensive way, we take into our consideration the effect of every change and record the classification accuracy (CA) every time. For SVM classifier the CA is calculated once for the training stage, giving it the name train classification accuracy (train CA) and again for the test stage with the name test classification accuracy (test CA). For KNN classifier the CA (taking the name KNN-CA) is measured four times assuming four different values of k (K=1, 3, 5, 10). The CA is not the only important issue to evaluate the system, the number of genes used for training and test has the same importance. Thus, we recorded the train CA, test CA and KNN-CA for a subset of 200 genes (Table II, III, IV) and again for a subset of 100 genes (Table V, VI, VII). The 200 genes subset is the result of the f-score gene selection technique only without combining the entropy- based technique. While, the 100 genes subset is the result of our combined gene selection technique.

Tables II, II, V and VI are dedicated to the SVM classifier and they emphasize the effect of two cross-validation methods, the LOOCV and the K-fold with three k values (2, 5, 10), on the train CA and test CA for both datasets used. We run the program several times for each attribute variation and record the least CA. The SVM used here is the linear SVM with linear kernel function. Applying polynomial and Gaussian kernel functions decrease the CA. So, we ignore recording the results when using kernel functions for SVM.

TABLE II: Train CA on 200 Genes Subset

Dataset	LOOCV	k-fold CV		
		K=2	K=5	K=10
Leukemia	1	1	1	0.9737
lymphoma	1	0.975	0.975	0.9250

TABLE III: Test CA on 200 Genes Subset

Dataset	LOOCV	k-fold CV		
		K=2	K=5	K=10
Leukemia	0.97	1	0.94	0.94
lymphoma	0.9459	0.9459	0.9459	0.9459

TABLE IV: KNN CA ON 200 GENES SUBSET

Dataset	K=1	K=3	K=5	K=10
Leukemia	0.8235	0.7353	0.8529	0.7647
lymphoma	0.7568	0.8108	0.8108	0.8378

TABLE V: TRAIN CA ON 100 GENES SUBSET

Dataset	LOOCV	k-fold CV		
		K=2	K=5	K=10
Leukemia	1	1	1	1
lymphoma	1	0.95	0.975	0.95

TABLE VI: TEST CA ON 100 GENES SUBSET

Dataset	LOOCV	k-fold CV		
		K=2	K=5	K=10
Leukemia	0.9706	1	1	0.9706
lymphoma	0.9459	0.9459	0.9189	0.9730

TABLE VII: KNN CA ON 100 GENES SUBSET

Dataset	K=1	K=3	K=5	K=10
Leukemia	0.7941	0.8235	0.8824	0.9706
lymphoma	0.7279	0.7027	0.8378	0.7838

For further evaluation of the proposed systems, we compared the highest obtained results with some published papers which are previously mentioned in the related work section. In this comparison we take two attributes into our consideration; the CA and the reduced number of genes used in the classification process. As these systems ware applied to two different datasets, we compared the results for each dataset with the results from published papers dealing with the same dataset, separately from the other. For the leukemia dataset, results are compared with [12], [13], [15]. For the lymphoma datasets, the comparison was with [16], [21]. One of the two proposed systems shows the highest CA for the two datasets over the other systems with a considerable number of genes. Fig. 3, 4 show the compared results where name of the system used and number of involved genes (between parentheses) are written on the horizontal axis, and the CA on the vertical axis. Our two proposed systems are represented by on the last two lines in each chart.

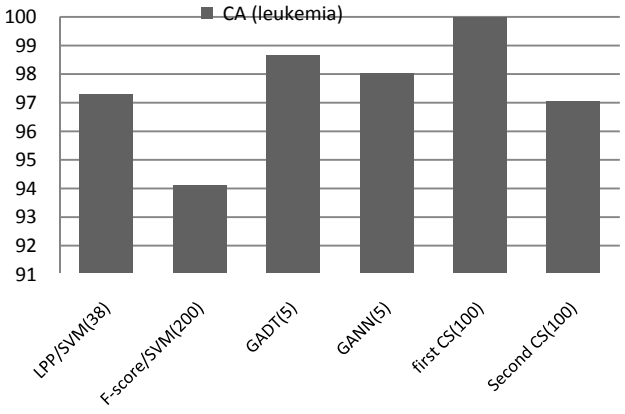

Fig. 3. Compared results of leukemia dataset.

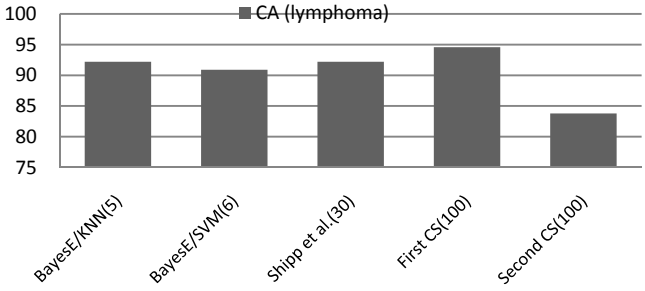

Fig. 4. Compared results of lymphoma dataset.

V. ANALYSIS

For analyzing the obtained results, the two proposed classification systems (First CS and Second CS) must be studied each alone at first. For the First CS, we can observe from the previous results in Tables (II, III, V, VI) that the highest train and test classification accuracies for leukemia dataset is 1(means all train and test samples are classified correctly) when using k-fold Cross Validation with k=2. Also the highest train and test classification accuracies of lymphoma dataset are 1 and 0.9459 which means all the training samples are classified correctly and only two test samples are misclassified. This is achieved when applying leave-one-out-cross-validation. These great results occur in both reduced datasets. This means that the chosen entropy-based gene selection technique is of a great value as it reduced the number of genes needed to classify a microarray sample to 50% of the number results from applying f-score technique only. These results are very promising compared to [12]. Also we noticed that applying different kernel functions didn't enhance the classification accuracy but sometimes reduced it. However, when applying the First CS on other microarray gene expression datasets, a need to kernel functions may arise to increase the CA.

For the Second CS, by watching Tables (IV, VII) we can find that the highest CA for the leukemia dataset results when using subset of 100 genes at k=10. This again ensures the importance of the proposed combined gene selection technique because further reduction of the number of genes enhances the accuracy. For the lymphoma dataset we can notice that using the two reduced subsets of genes (either 200 or 100) didn't have a great effect on the CAs which means that using a subset of 100 genes is better as the number of genes is reduced to the half.

VI. CONCLUSIONS

In this paper, we introduced two new hybrid classification systems for classifying cancer samples using microarray gene expression datasets. The main target of the proposed systems is to get the highest accuracy when classifying the samples using a small subset of informative genes. A combination of two gene selection techniques is introduced to solve the problem of the microarray high dimensionality. This combined technique presents high performance as it reduces the number of genes by 71.29%. SVM and KNN were chosen for classification as they are very efficient binary classification techniques and usually give good results by attenuating their variety of attributes. The two systems were a result of integrating the proposed gene selection technique once with SVM resulting in the First CS and another time with KNN resulting in the Second CS. First CS and Second CS were applied to two public microarray datasets, leukemia dataset and lymphoma dataset. First CS shows a maximum accuracy on leukemia dataset without any misclassifications and a very small error rate on lymphoma dataset equals to 0.0541. Thus, it reaches its

main objective as it classified the test cancer samples with high classification accuracy (100% on leukemia dataset and 94.59% on lymphoma dataset) using only a set of 100 genes instead of using the original number of genes (7129). Comparison with others verifies the efficiency of the First CS as it results in a perfect CA (without any misclassifications) on leukemia and the highest CA recorded on lymphoma till now. Although Second CS results in good but not excellent results compared to the First CS, it verifies the importance of the proposed gene selection technique. This is because the 100 genes subset (which is the minimum number of genes used in this work) produces the highest results. The First CS was proven to be a powerful system which can be adapted to any microarray gene expression dataset.

REFERENCES

[1] D. T. Larose, "Discovering knowledge in Data: An Introduction to Data Mining," Ed. Hoboken, New Jersey: *John Wiley & Sons, Inc*.,2005.

[2] Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery," 3rd ed., 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

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

[4] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms" *Wiley-IEEE Press*, November 2002.

[5] J. Han, M. Kamber: Data Mining:Concepts and Techniques, 2nd ed., *Morgan Kaufmann Publishers*, 2006.

[6] D.V. Setty, T.M. Rangaswamy, K.N. Subramanya, "A Review on Data Mining Applications to the Performance of Stock Marketing," *International Journal of Computer Applications (0975 – 8887)* 1(3),2010,pp. 25-34.

[7] J.W. Seifert, "Data Mining:An Overview," In: CRS Report for Congress. *Congressional Research Service ˜ The Library of Congress*, 2004.

[8] K. Raza, "Application Of Data Mining In Bioinformatics," *Indian Journal of Computer Science and Engineering,* 1(2),2010, pp.114-118.

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

[10] H.F. Ong, N. Mustapha, M.N. Sulaiman, "Integrative Gene Selection for Classification of Microarray Data" *Computer and Information Science (CCSE),* 4(2),2011, pp.55-63.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring" *Science,* 286,1999,pp.531–537. doi: 10.1126/science.286.5439.531.

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

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

[14] E. J. Moler, M. L. Chow, and I. S. Mian, "Analysis of molecular profile data using generative and discriminative methods,"*Physiological Genomics* , 4(2),2000, pp.109-126.

[15] P. Yang, Z. Zhang, "Hybrid Methods to Select Informative Gene Sets in Microarray Data Classification," In: *Australian Conference on Artificial Intelligence*, Berlin Heidelberg: Springer Verlag, Ed. M.A. Orgun, J.Thornton, 2007,pp. 810-814.

[16] J. G. Zhang, H. W. Deng, "Gene selection for classification of microarray data based on the Bayes error," *BMC Bioinformatics* 8(1),2007,pp. 370-379.

[17] S. Deegalla, H. Boström, "Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods," In. *Intelligent Data Engineering and Automated Learning - IDEAL 2007,* Berlin Heidelberg;:Springer, Ed: H. Yin, P. Tino, E. Corchado, W. Byrne, X. Yao, 2007, vol.4881, pp. 800-809.

[18] 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),2003, pp.593-608.

[19] S. V. Sanden, D. Lin, T. Burzykowski, "Performance of gene selection and classification methods in a microarray setting: A simulation study," *Communications in Statistics-Simulation and Computation,* 37(2),2008,pp. 409-424.

[20] C. Shang, Q. Shen, "Aiding classification of gene expression data with feature selection: a comparative study," *International Journal of Computational Intelligence Research* , I(1),2006,pp. 68-76.

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

[22] E. B. Huerta, B. Duval, J. K. Hao, "A hybrid ga/svm approach for gene selection and classification of microarray data," In: *EvoWorkshops, LNCS 3907,* 2006, pp.34-44.

[23] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol 3,2003, pp.1157-1182.

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

[25] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, G. Rätsch, "Support Vector Machines and Kernels for Computational Biology,"*PLoS Comput Biol,* 4(10), 2008, e1000173. doi:10.1371/ journal.pcbi.1000173

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

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