

# **An Efficient Statistical Model Based Classification Algorithm for Classifying Cancer Gene Expression Data with Minimal Gene Subsets**

Mallika Rangasamy

Sri Ramakrishna College of Arts and Science for Women, India

E-mail:mallikapanneer@hotmail.com

Saravanan Venketraman

Karunya School of Science and Humanities, Karunya University, India

E-mail:tvssaran@hotmail.com

## **ABSTRACT**

Data mining algorithms are extensively used to classify gene expression data, in which prediction of disease plays a vital role. This paper aims to develop a new classification algorithm for cancer gene expression data using minimal number of gene combinations i.e. minimum gene subsets. The model uses classical statistical technique for gene ranking and two different classifiers for gene selection and prediction. The proposed method proves the capability of producing very high accuracy with very minimum number of genes. The methodology was tried with three publicly available cancer databases and the results were compared with the earlier approaches and proven better and promising prediction strength with less computational burden. This paper also focuses on the importance of applying an efficient gene selection method prior to classification can lead to good performance and the results are proven to be the best.

Keywords: Microarray Data, Classification, SVM-OAA, LDA, Prediction, ANOVA  
P-values

## **INTRODUCTION**

Microarray technology has made the modern biological research by permitting the simultaneous study of genes comprising a large part of the genome (Per Broberg, 2003). In response to the rapid development of DNA Micro array technology, classification methods and gene selection techniques are being computed for better use of classification

algorithm in microarray gene expression data (Chai & Domeniconi, 2001; Jaeger, Sengupta, & Ruzzo, 2003).

The goal of classification is to build a set of models, which are able to correctly predict the class of different objects. The input to such models is a set of objects (i.e., training data), the classes, which these objects belong to (i.e., dependant variables), and a set of variables describing different characteristics of the objects (i.e., independent variables). Once such a predictive model is built, it can be used to predict the class of the objects for which class information is not known. The key advantage of supervised learning methods over unsupervised methods like clustering is that by having an explicit knowledge of the classes the different objects belong to, these algorithms can perform an effective feature selection if that leads to better prediction accuracy.

In classification analysis of microarray data, gene selection is one of the critical aspects (Li, Campbell, & Tipping, 2002; Diaz, 2005; Hua, Xiong, Lowey, Suh, & Dougherty, 2005; Jirapech & Aitken, 2005). Efficient gene selection can drastically ease computational burden of the subsequent classification task, and can yield a much smaller and more compact gene set without the loss of classification accuracy (Ben-Dor, Bruhn, Friedman, Nachman, Schummer, & Yakhini, 2000; Blanco, Larranaga, Inza, & Sierra, 2004).

In microarray classification analysis, the main objective of gene selection is to search for the genes, which keep the maximum amount of information about the class and minimize the classification error (Zhang & Deng, 2007).

With the help of gene expression data obtained from microarray technology, heterogeneous cancers can be classified into appropriate subtypes and the challenge of effectiveness in cancer prediction lies when the data is high dimensional.

Supervised machine learning can be used for cancer prediction, which uses a part of the dataset as training set and uses the trained classifier to predict the samples in the rest of the data set to evaluate the effectiveness of the classifier (Wang, Chu, & Xie, 2007).

With DNA microarray data, selecting a compact subset of discriminative genes from thousands of genes is a critical step for accurate classification.

Selection of important genes using statistical technique was carried out in various papers such as Fisher Criterion, Signal-to-Noise, traditional t-test, and Mann-Whitney rank sum statistic (Venu Satuluri, 2007), chi-squared test, Euclidean distance (Saeyns, Inza, & Larrañaga, 2007) and the some of the classification algorithms used were SVMs, k-nn (Chin & Deris, 2005), Genetic algorithms (GA) (Liu & Iba, 2001) Naïve bayes (NB)

(Keller, Schummer, Hood, & Ruzzo, 2000). This paper used the two publicly available cancer dataset Lymphoma, Liver and Leukemia.

In 2003, Tibshirani Hastie, Narasimhan, and Chu successfully classified the lymphoma data set with only 48 genes by using a statistical method called nearest shrunken centroids and used 43 genes for SRBCT data.

Wang, Chu, and Xie (2007) proposed an algorithm in finding out minimum number of gene up to 3 genes with best classification accuracy using C-SVM and FNN.

Wong and Hsu (2008), classification method to classify the causality of a disease is of two stages with gene Selection mechanism with individual or gene subset ranking as the first stage and classification tool with or without dimensionality reduction as the second stage.

This paper proposes an efficient classification algorithm using statistical model for individual gene ranking and data mining models for selecting minimum number of gene rather than thousands of genes, which can be used to give good classification accuracy. Further, the paper emphasizes on applying an efficient gene selection method prior to classification improves classification accuracy and the paper also proves for the better result when the same classifiers are used for gene selection and classification than using different classifiers.

## **METHODOLOGY**

### **Algorithm**

- Step 1: Pick half the samples randomly for training and for each gene compute ANOVA P-values.
- Step 2: Sort the genes as per their least p-values. (Gene Ranking)
- Step 3: Pick the top n genes and generate all possible combinations and train the classifier SVM-OAA / LDA using all possible gene combinations. (Gene Selection)
- Step 4: Validate the classifier using 5 fold / 10 fold Cross validation method
- Step 5: Select the gene combinations that achieved 100% CV accuracy
- Step 6: Retrain the classifier
- Step 7: Use the classifier to predict the samples in testing database

### **ANOVA p-values**

ANOVA is a technique, which is frequently used in the analysis of microarray data, e.g. to find if there are any significant difference in the types of disease (cancer), and to

select interesting genes based on P-values (Troyanskaya, Cantor, Sherlock, Brown, Hastie, Tibshirani, Botstein, & Altman, 2001). The ANOVA test is known to be robust and assumes that all sample populations are normally distributed with equal variance and all observations (samples) are mutually independent.

The approach chosen in this paper is the one-way ANOVA that performs an analysis on comparing two or more groups (samples) for each gene and returns a single p-value that is significant if one or more groups are different from others. The most significantly varying genes have the smallest p-values.

$$\text{Within - groups estimate of } \sigma_y^2 = \frac{\sum_{ij} (y_{ij} - \bar{y}_j)^2}{\sum_j (n_j - 1)} = \frac{SS_{WG}}{df_{WG}} = MS_{WG}$$

$$\text{Between - groups estimate of } \sigma_y^2 = \frac{\sum_j n_j (\bar{y}_j - \bar{y}_{..})^2}{(k - 1)} = \frac{SS_{BG}}{df_{BG}} = MS_{BG}$$

$$F(df_{BG}, df_{WG}) = \frac{\text{Between Groups estimate of } \sigma_y^2}{\text{Within Groups estimate of } \sigma_y^2} = \frac{MS_{WG}}{MS_{BG}}$$

Of all the information presented in the ANOVA table, if the p value for the F- ratio is less than the critical value ( $\alpha$ ), then the effect is said to be significant. In this paper the  $\alpha$  value is set at .05, any value less than this will result in significant effects, while any value greater than this value will result in non-significant effects. The very small p-value indicates that differences between the column means (group means) are highly significant.

The probability of the F-value arising from two identical distributions gives us a measure of the significance of the between-sample variation as compared to the within-sample variation. Small p-values indicate a low probability of the between-group variation being due to sampling of the within-group distribution and small p-values indicate interesting genes. The paper uses the p-values to rank the important genes with small values and the sorted numbers of genes are used for further processing.

### Support Vector machine-one-against- all (SVM-OAA)

SVMs are the most modern method applied to classify gene expression data, which works by separating space into two regions by a straight line or hyper plane in higher dimensions. SVMs were formulated for binary classification (2 classes) but cannot naturally extend to more than two classes. SVMs are able to find the optimal hyper plane that minimizes the boundaries between patterns (Song & Rajasekaran, 2007). SVMs are power tools used widely to classify gene expression data (Marchiori & Sebag, 2005; Lee & Lee, 2003). How to effectively extend SVM for a multi-class classification is still an ongoing research issue (Hsu & Lin, 2002). This paper gives an effective method to classify multi-class problems. For extending SVMs to classify to multi-class problems, SVMs were designed with SVM one against one, SVM one against all. This paper efficiently uses SVM-one against all method (SVM-OVA) with RBF kernel. The SVM-OAA constructs 'n' binary SVM classifier with the  $i^{\text{th}}$  class separating from all other classes. Each binary SVM classifier creates a decision boundary, which can separate the group it represents from the remaining groups.

For training data  $t = (X_1, Y_1), (X_2, Y_2) \dots (X_t, Y_t)$ ;  $X_n \in R^n$  and  $n=1 \dots t$ ,  $Y_n = 1 \dots K$  is the class labels corresponding to  $X_n$ . The  $n^{\text{th}}$  SVM solves

$$\begin{aligned} \text{Min } & \frac{1}{2} (w^i)^T w^i + C \sum_{j=1}^t \xi_j^i w^i b_i \in^i, \\ (w^i)^T \phi(x_j) + b^i & \geq 1 - \xi_j^i, \text{ if } y_i = i, \\ (w^i)^T \phi(x_j) + b^i & \leq -1 + \xi_j^i, \text{ if } y_i \neq i, \\ \xi_j^i & \geq 0, j = 1 \dots K, t, \end{aligned}$$

Where each data  $x_j$  is mapped to the feature space by function  $\phi$  and  $c$  is the penalty parameter. The Radial basis function (RBF) is the most popular choice of kernel functions used in Support Vector Machines, which can be represented by

$$K(X_i, X_j) \equiv e^{-\gamma |X_i - X_j|^2}$$

Where  $X_i$  is the support vector of the  $i^{\text{th}}$  class and  $X_j$  is the support vector for the new higher dimensional space and  $\gamma$  is the tuning parameter. This paper uses SVMs with RBF kernel because they can be able to give the same decision as that of RBF network. (Hsu & Lin, 2002).

### Linear Discriminant Analysis (LDA)

LDA otherwise known as FLDA (Fishers Linear Discriminant Analysis calculates a straight line or hyperplane that separates 2 known classes (groups). Unlike SVM where the hyperplane is chosen to minimize the misclassification errors, LDA chooses hyperplane to minimize within class variance on either side of the line and minimize the between-class variance. Then the side of the line or hyperplane determines the class of the unknown sample. The disadvantage in using LDA is that it can perform classification well only for linearly separable data. LDA cannot extend to more than 2 classes. This paper used the basic LDA for the 2 class datasets such as Liver and Leukemia and compared the results with multiclass SVM.

Dudoit, Fridlyand, and Speed (2002) used LDA to classify cancer gene expression data. In a training set  $T$  of size  $n$ ,  $(t_i, c_i)$  is the representation of each tuple, where  $t_i$  is in the form  $(t_i.X_1, t_i.X_2, \dots, t_i.X_m)$ , a vector of expression values of  $m$  number of genes in tuple  $i$  and class  $C_i$  is the class label for the corresponding  $t_i$ . LDA tries to find the linear combination  $M\alpha$  of  $M$  samples that has large ratio of between-class variance ( $S_e$ ) to within-class variance ( $S_n$ ),  $\alpha$  being the transformation matrix denoted by

$$M\alpha = \alpha' S_e \alpha / \alpha' S_n \alpha$$

The extreme values of  $M$  is obtained by the eigen values and eigenvectors of the matrix  $S_n^{-1} - S_e$  which has corresponding eigen vectors  $v_1, v_2, \dots, v_h$ . For any sample  $t$ , the discriminant variables are defined as  $u_k = tv_k$ , where  $l = 1, 2, \dots, h$  and  $h = \min(K-1, m)$  for  $K$  number of classes and  $v_1$  maximizes  $\alpha' S_e \alpha / \alpha' S_n \alpha$ .

## RESULT AND DISCUSSION

### Description of the Database used

There are many microarray datasets, which are publicly available from published cancer gene expression studies. Among various datasets, this paper uses the three datasets Lymphoma, Liver and Leukemia. The following table is the short description of the datasets used in this paper (see Table1).

### Lymphoma datasets

The algorithm was applied to the lymphoma dataset with 4026 genes and 62 samples with 3 classes namely DLBCL, CLL and FL, the subtype's of Lymphoma cancer.

**Table1: Description of the dataset used**

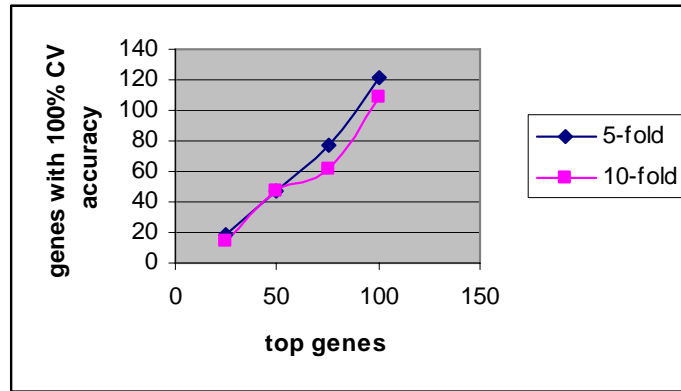
Database	Number of genes	Number of samples	No of class (groups)
Lymphoma	4026	62	3
Liver	1648	156	2
Leukemia	7129	72	2

The dataset was with few missing data. The K-Nearest neighbour (KNN) algorithm as used by Troyanskaya et al. (2001) with  $k=3$  was used and the missing data were filled. KNN algorithm finds the nearest distance of the 'n' known samples and imputes the missing value with the class of the nearest 'n' sample. Half of the samples were picked randomly for training i.e. 31 samples from 3 groups and all the samples for testing (62 samples). For the training dataset with 4026 X 31 dimensions ANOVA p-value was calculated for each gene and the top ranked genes were selected. All possible combinations of the top n genes were generated. For n number of top genes all possible gene combinations are  $n(n+1)/2!$ . Using these combinations SVM-OAA were trained.

The performance of the classifier was validated using cross validation (CV) technique with 5-folds and 10-folds. For 5 folds, the samples in the training dataset was randomly divided into 5 equal parts, classification was performed for 5 runs, using the 4 parts as training and the other as testing. Each time the classifier is trained, a different test set was used, so that over 5 runs of the classified, all the samples were used as test set. The average 5-fold accuracy for each run was calculated and average error rate in training were calculated.

The gene subsets that achieved 100% CV accuracy from the training samples were selected and the classifier was retrained. Then the classifier was used to predict the samples in the testing dataset.

First, the method used all possible gene subsets from top ranked 25, 50, 75 and 100 genes and obtained a good accuracy for 5 fold and 10 fold. Fig.1 depicts the number of gene pairs that obtained 100% CV accuracy using 5-fold and 10-folds.



**Figure 1: No of genes achieved 100% CV accuracy for Lymphoma dataset**

The SVM-OAA classifier was retrained using the gene pairs that achieved 100% CV accuracy and their corresponding testing accuracy were measured. Table 2 shows the best testing accuracy for the top ranked genes. Though it is recommended to use n-fold Cross validation for estimating the accuracy, the paper uses minimum of 5-fold Cross validation test as it gave a good result and to avoid computational burden in dividing the dataset into 10 equal parts.

**Table 2: Best testing accuracy for Lymphoma dataset for 5-fold and 10-fold cross validation**

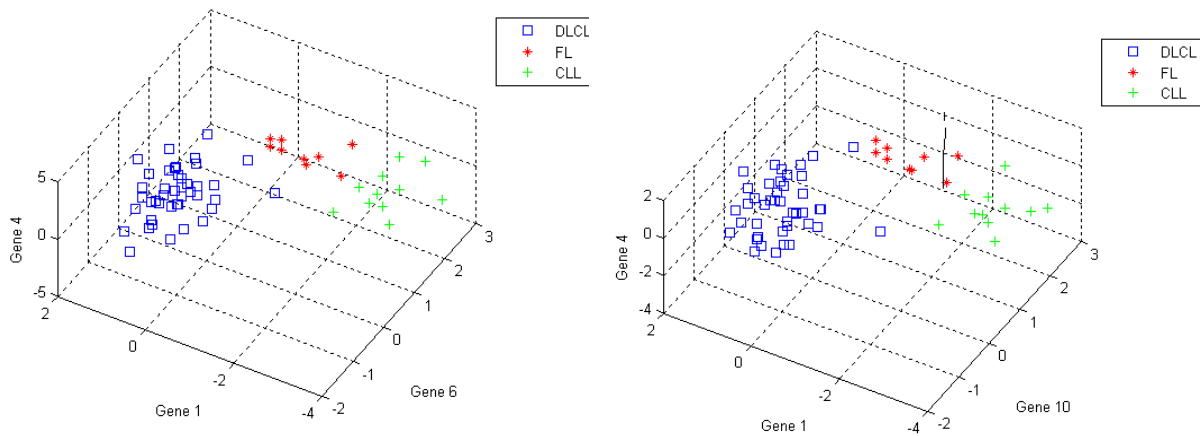
Top gene	Accuracy
	5-f 10-f
25	95.16
	96.77
50	96.77
	98.38
75	96.77
	98.38
100	96.77
	98.38

The proposed method then aimed to try the same procedure for gene subset of 3 aiming to increase the number of genes with 100% training accuracy. In comparison with the 2 gene subsets, good results were achieved for gene subset of 3. Table 3 shows the number of combinations that achieved 100% CV accuracy and their corresponding testing accuracies.

**Table 3: Lymphoma dataset: Number of gene combinations that achieved 100% accuracy (for 5-fold) in training and their corresponding test accuracy**

Top genes	Combinations	No. of gene combinations with 100% CV accuracy	Best Testing accuracy
10	120	25	100%
25	2300	684	96.77%
30	4060	447	98.38
50	19600	2963	100%

The subsets that achieved 100% accuracy for both training and testing were plotted in Figure 2. All the plots show a good separation of 3 classes (groups). These gene subsets may allow doctors to predict the type of Lymphoma for a patient.



**Figure 2: Plots of gene expression level of the combination (1,4,5) and (1,4,10) showing clear separation of 3 classes**

Furthermore, in Table 4, the proposed method has proven for the promising result with less computation burden in training and minimal number of genes needed for prediction.

**Table 4: Comparison of results of previous work with minimal number of genes, best testing accuracy and computation time for Lymphoma dataset**

Method	No of genes	Best testing accuracy	Computation time (in seconds)
Proposed method	3	100%	215
SVM (Song & Rajasekaran, 2007)	5	100%	675
ELM (Zhang, Huang, Sundararajan, & Saratchandran, 2007)	10	97.33%	386.73
Bayesian learning with local SVM (Marchiori & Sebag, 2005)	30	93%	-

Confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different groups. A confusion matrix for the gene combination (1,16,33) that achieved the 100% testing accuracy is shown in Figure 3

		Predicted class		
		DLBCL	FL	CLL
Actual class	DLBCL	42	0	0
	FL	0	9	0
	CLL	0	0	11

**Figure 3: Confusion matrix for (1,16,33) from Lymphoma dataset**

### Liver Dataset

Liver dataset (<http://genome-www.stanford.edu/hcc>) has 2 classes HCC and non-tumor liver for 1648 genes for 156 observations in which 82 are from HCC and 74 from non-tumor livers. Unlike the Lymphoma dataset, which were with the prediction of subtypes in Lymphoma cancer, the Liver dataset were from the cancerous and non-cancerous tissue. The problem is to predict whether the samples are from cancerous or non-cancerous tissue. The proposed algorithm was applied. Randomly half of the samples were selected for training and all samples for testing. Results are depicted in Table 4 showing the average training accuracy and the corresponding best test accuracy for gene subset of 2 and 3 genes. Best results were achieved with 3 gene subsets.

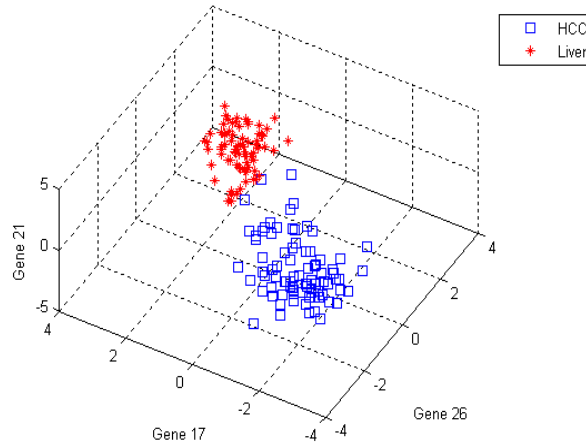
**Table 4: Results of Liver dataset showing train and test accuracy for 3 and 2 gene subsets**

Top gene	Average training accuracy		Best testing accuracy	
	3 gene	2 gene	3 gene	2 gene
25	97.78%	96.14%	98.71%	98.71%
50	97.42%	94.21%	99.39%	98.71%

The algorithm was tried with 3 gene subsets, since the number of gene subsets that achieved 100% CV accuracy was very low compared with 2 gene subsets. Figure 4 shows the confusion matrix for the gene combination (7,19,50), which gave 99.39% prediction strength i.e. 1 error out of 156 samples and the plot in Figure 5 shows a clear separation of the 2 classes HCC and non-tumor liver (represented as Liver and HCC).

		Predicted class	
		HCC	Liver
Actual class	HCC	81	1
	Liver	0	74

**Figure 4: Confusion matrix for liver dataset Pertaining to the gene combination (7,19,50)**



**Figure 5: View of gene expression levels of (17, 26, 21) that achieved 100% training accuracy for Liver dataset**

For the Liver database, linear discriminant analysis (LDA) was also applied for gene selection and SVM-OAA to classify and predict the samples in the testing dataset (LDA: SVM-OAA). Since the LDA used in this paper can classify only two class (group) dataset it was not used for the Lymphoma dataset .The LDA: SVM-OAA method showed a good performance in computation time in training for the Liver dataset, but the number of gene subsets that achieved 100% CV accuracy was low compared to SVM-OAA gene selection method. Table 6 shows the comparison results

**Table 6: Comparison of SVM-OAA method and LDA:SVM-OAA method for Liver dataset.**

Top gene	SVM-OAA	LDA: SVM-OAA
	No of gene subsets with 100% CV accuracy	No of gene subsets with 100% CV accuracy
10	24	23
25	582	476
30	824	624
50	4015	2513

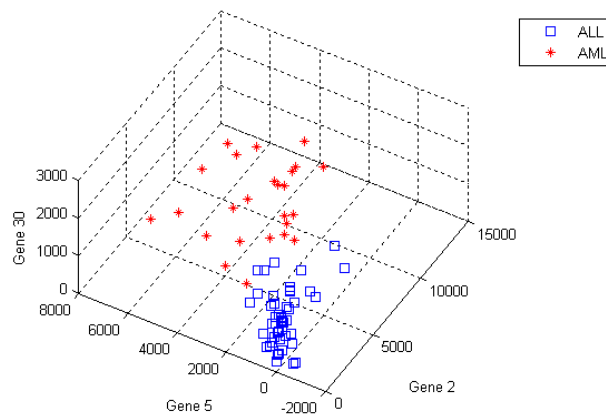
The average training accuracy for 3 gene subsets was almost same in both the methods and the computational time in training was considerably very less in the LDA: SVM-OAA method, but the number of gene subsets that achieved 100% accuracy was less in LDA: SVM-OAA method. Best testing accuracy of 99.39% were achieved for 25 gene subsets using SVM-OAA. Comparisons of testing results for the Liver database with the top 50 selected genes are shown in Table 7.

**Table 7: Average Testing Accuracy Comparison table for liver dataset**

Method	Average testing accuracy
SVM-OAA	99.39%
LDA: SVM-OAA	98.72%
FNN (Wang, Chu, & Xie, 2007)	98.1%

### Leukemia Dataset

The publicly available Leukemia dataset contains 7129 genes and 2 classes, with the gene expression data ALL, and AML, subtypes of leukemia. The same methodology used for Lymphoma and Liver dataset were applied. LDA: SVM-OAA was tried and promising results were achieved. Plots in Figure 6 shows a clear separation of two classes AML and ALL for the gene subset (2, 5, 30).



**Figure 6: View of gene expression levels of (2, 5, 30) that achieved 100% train accuracy for leukemia dataset**

Table 8 shows the results achieved for the Leukemia dataset using SVM-OAA and LDA: SVM. The average error rate in training and the number of gene subsets that achieved 100% training accuracy, the best testing accuracy are depicted in the table. The best testing accuracy was 97.22 with the SVM-OAA method. In comparison with Dudoit

et al. (2002)'s work with best accuracy of 95% and with Furey, Christianini, Duffly, Bednarski, Schummer, and Hauessler (2000) with 94% accuracy, the proposed method achieved the best.

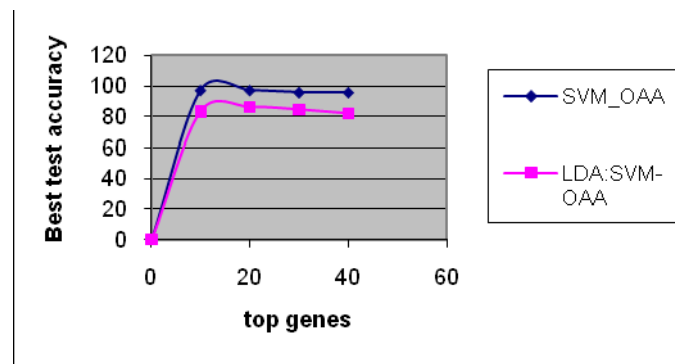
**Table 8: Results of Leukemia dataset for the two methods**

Top genes	Method	Average error rate	Average Training accuracy	No of genes achieved 100% CV accuracy	Best Testing accuracy
10	SVM-OAA	0.0555	94.45	5	97.22%
	LDA-SVM-OAA	0.0643	93.57	1	83.33%
25	SVM-OAA	0.0902	90.08	33	97.22%
	LDA-SVM-OAA	0.0719	92.81	9	86.11%
30	SVM-OAA	0.0886	91.19	73	95.83
	LDA-SVM-OAA	0.0709	92.91	65	94.44
50	SVM-OAA	0.0886	91.14	644	95.83
	LDA-SVM-OAA	0.0831	91.69	343	81.94

Furthermore Figure 7 shows the confusion matrix for the best gene subset (6,9,22) that achieved the maximum testing accuracy of 97.22% i.e. 2 errors in 72 samples

		predicted class	
		ALL	AML
Actual class	ALL	45	2
	AML	0	25

**Figure 7: Confusion matrix for the Gene subset (6,9,22) of Leukemia dataset**



**Figure 8: Comparative chart for SVM-OAA and LDA: SVM-OAA Method applied for Leukemia dataset.**

The graph in Figure 8 proves for better results achieved for the Leukemia dataset when the SVM-OAA with RBF kernel function were used to select important genes as

well as classification than that of using LDA for gene selection and SVM-OAA for classification. Since Lymphoma dataset was with 3 classes, LDA was not applied.

### CONCLUSION

The paper proposed an efficient approach for cancer classification based on microarray gene expression data giving out the best prediction strength using minimal gene subsets and the results have proven to be the best with the SVM-OAA method than with that of LDA. The method was designed to address the importance of gene ranking and selection prior to classification, which improves the prediction strength of the classifier. The paper focused on promising accuracy results with very few number of gene subsets enabling the doctors to predict the type of cancer. The results on Leukemia and Liver datasets shows the importance of the same classifier used for both the gene selection and classification can improve the strength of the model.

### REFERENCES

- Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., & Yakhini, Z. (2000). *Tissue classification with gene expression profiles*. Proceedings of the Fourth Annual International Conference on Computational molecular biology, Tokyo, Japan.
- Blanco, R., Larranaga, P., Inza, I., & Sierra, B. (2004). Gene selection for cancer classification using wrapper approaches. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(8), 1373-1390.
- Chai, H., & Domeniconi, C. (2001). *An Evaluation of Gene Selection Methods for Multi-class Microarray Data Classification*. Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, Fairfax, VA.
- Diaz, U.R. (2005). *Supervised methods with genomic data: a review and Cautionary view*. In F. Azuaje & J. Dopazo (Eds.), *Data analysis and visualization in genomics and proteomics*. New York: Wiley.
- Dudoit, S., Fridlyand, J., & Speed, T. (2002). Comparison of discrimination methods for the classification of tumor using gene expression data. *Journal of the American Statistical Association*, 97, 77-87.
- Furey, T.S., Christianini, N., Duffly, N., Bednarski, W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906-914.

- Jaeger, J., Sengupta, R., & Ruzzo, W.L. (2003). Improved gene selection for classification of Microarrays. *Pacific Symposium on Biocomputing*, 8, 53-64.
- Jirapech, U.T., & Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, 6, 148.
- Keller, A. D., Schummer, M., Hood, L., & Ruzzo, W. L. (2000). *Bayesian Classification of DNA Array Expression Data* (Tech. Rep. No. UW-CSE-2000-08-01). Seattle: University of Washington, Department of Computer Science & Eengineering.
- Lee, Y., & Lee, C.K. (2003). Classification of multiple cancer types by Multicategory support vector machines using gene expression data. *Bioinformatics*, 19(9), 1132-1139.
- Li, Y., Campbell, C., & Tipping, M. (2002). Bayesian automatic relevance Determination algorithms for classifying gene expression data. *Bioinformatics*, 18(10), 1332-1339.
- Liu, J., & Iba, H. (2001). Selecting Informative Genes with Parallel Genetic algorithms in Tissue Classification. *Genome Informatics*, 12, 14-23.
- Hsu, C.W. , & Lin, C.J. (2002). A comparison of methods for multiclass Support Vector machines. *IEEE transactions on neural network*, 13(2), 415 - 425.
- Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E.R. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*, 21(8), 1509-1515.
- Marchiori, E., & Sebag, M. (2005). Bayesian learning with local support vector Machines for cancer classification with gene expression data. *Lecture Notes in Computer Science*, 3449, 74-83.
- Per Broberg, P. (2003). Statistical methods for ranking differentially expressed Genes. *Genome Biology*, 4(6), R41.
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.
- Song, M., & Rajasekaran, S. (2007). A Greedy Correlation-Incorporated SVM-Based Algorithm for Gene Selection. *21st International Conference on Advanced Information Networking and Applications Workshops*, 1, 657-661.
- Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2003). Class Prediction by Nearest Shrunken Centroids with Applications to DNA Microarrays. *Statistical Science*, 18, 104-117.

- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R.B. (2001). Missing values estimation methods for DNAMicroarrays. *Bioinformatics*, 17, 520-525.
- Venu Satuluri, V. (2007). A survey of parallel algorithms for classification.
- Chin, Y.L., & Deris, S. (2005). A study on gene selection and classification Algorithms for classification of microarray Gene expression data. *Jurnal Teknologi*, 43, 111-124.
- Wang, L., Chu, F., & Xie, W. (2007). Accurate Cancer Classification Using Expressions of Very Few Genes. *IEEE/ACM Transactions on computational biology and bioinformatics*, 4(1), 40-53.
- Wong, T.T., & Hsu, C.H. (2008). Two-stage classification methods for microarray data. *Expert Systems with Applications*, 34(1), 375-383.
- Zhang, J.G., & Deng, H.W. (2007). Gene selection for classification of microarray data based on the Bayes error. *BMC Bioinformatics*, 8, 370.
- Zhang, R., Huang, G.B., Sundararajan, N., & Saratchandran, P. (2007). Multicategory classification using an Extreme Learning machine for Microarray Gene Expression Cancer Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3), 485- 495.