A novel approach for automatic gene selection and classification of gene based colon cancer datasets

[1,2]Saima Rathore, [1,3]Muhammad Aksam Iftikhar, [1]Mutawarra Hussain
[1]DCIS, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan
[2]DCS&IT, University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir
[3]Comsats Institute of Information Technology, Lahore, Pakistan
saimarathore_2k6@yahoo.com, aksam.iftikhar@gmail.com, mutawarra@pieas.edu.pk

Abstract— **Colon cancer heavily changes the composition of human genes (expressions). The deviation in the chemical composition of genes can be exploited to automatically diagnose colon cancer. The major challenge in the analysis of human gene based datasets is their large dimensionality. Therefore, efficient techniques are needed to select discerning genes. In this research article, we propose a novel classification technique that exploits the variations in gene expressions for classifying colon gene samples into normal and malignant classes, and quite intelligently tackles the larger dimensionality of gene based datasets. Previously individual feature selection techniques have been used for selection of discerning gene expressions, however, their performance is limited. In this research study, we propose a feed forward gene selection technique, wherein, two feature selection techniques are used one after the other. The genes selected by the first technique are fed as input to the second feature selection technique that selects genes from the given gene subset. The selected genes are then classified by using linear kernel of support vector machines (SVM). The feed forward approach of gene selection has shown improved performance. The proposed technique has been tested on three standard colon cancer datasets, and improved performance has been observed. It is observed that feed forward method of gene selection substantially reduces the size of gene based datasets, thereby reducing the computational time to a great extent. Performance of the proposed technique has also been compared with existing techniques of colon cancer diagnosis, and improved performance has been observed. Therefore, we hope that the proposed technique can be effectively used for diagnosis of colon cancer.**

Keywords- Colon cancer; Gene expressions; Chi-Square; mRMR

I. INTRODUCTION

Colon is a major constituent of large intestine and its cancer is quite common worldwide. Major reason of colon cancer is low intake of fruits/vegetables, and more intake of meat and fatty stuff. Some other reasons of colon cancer include older age, chain smoking, and family history of colon cancer [1].

Conventionally, colon cancer is diagnosed by the microscopic inspection of histopathological colon samples, but, the process of manual examination is laborious and time-consuming for the pathologists, and also has inter/intra-observer variability in diagnosis [2]. Therefore, there is the need of automatic techniques, which could automatically detect colon cancer, and could provide a reliable secondary opinion to the histopathologists. Researchers have proposed several automatic techniques for colon cancer diagnosis. A recent survey by Rathore et al. [3] summarizes a larger subset of these techniques.

A famous method of colon cancer diagnosis is the analysis of human genes by using cDNA and Oligoneclotide microarrays. The focus of this research work is also the classification of colon samples into normal and malignant samples based on the values of gene expressions. In 1999, Backert et al. used 588 gene expressions for classification of colon samples into normal, non-mucinous and mucinous classes of colon tissues, and yielded classification accuracy of slightly above 50% [4]. Li et al. selected discerning gene expressions by using genetic algorithm (GA), and employed the selected genes for classification by using k-nearest neighbor (KNN) classifier [5]. They obtained an accuracy of 94.10%. Chen et al. [6] used multiple kernel support vector machine, where multiple kernels are described as the convex combination of the single kernels. Chen et al. tested their technique on leukemia and colon cancer datasets, and more than 90% recognition rate was observed for both the datasets. Similarly, Shon et al. proposed a technique in which wavelet transformation was used for reducing the dimensionality of gene datasets, and selecting discriminative genes [7]. They used probabilistic neural network (PNN) for classification, and obtained 92% accuracy on colon cancer dataset.

Further, Venkatesh et al. proposed an EFJ-neural network based system for classification of colon cancer dataset [8], and achieved 94% classification accuracy. Similarly, Kulkarni et al. proposed a technique, wherein t-statistic and mutual information were used as gene selection strategies, and decision trees and genetic programming were used as classifiers [9]. The results proved the combination of mutual information and genetic programming to be most promising compared to others.

Recently, Lee et al. proposed a finite impulse response extreme learning machine based colon cancer detection technique [10], and achieved quite promising classification of colon samples. Furthermore, Tong et al. proposed a method of colon cancer detection, wherein top scoring pair method was

978-1-4799-6089-7/14/$31.00 ©2014 IEEE

employed for selection of 50 gene pairs. The selected pairs were used for training of linear SVM classifiers [11]. GA was then utilized to select such an optimal combination of SVM base classifiers, which yields maximum classification performance. Tong et al. achieved an accuracy of 90.30% with colon dataset.

Microarrays though facilitate to analyze huge volume of data enabling insight into tissues and find the state of the cancer, however, a major challenge in microarrays based gene analysis is large dimensionality of the gene set under consideration. Hence, proficient techniques are required to identify discerning genes amongst a large pool of available genes. In this paper, we tackle the issues quite reasonably. Two feature selection strategies, namely, minimum redundancy maximum relevancy (mRMR), and chi-square have been employed in a feed-forward fashion to select a discerning gene set quite capable to distinguish the two classes. The discerning genes are selected initially using Chi-Square, which returns a larger subset of gene expressions. The mRMR method is then applied on the selected genes in order to further refine the selection process. The gene set selected by mRMR is small compared to that selected by Chi-Square. Individual feature selection techniques have been used for selecting discerning genes in the past. However, we have experimentally validated that using two effective feature selection methodologies in a feed-forward manner is an interesting idea due to the chance of selecting a more discernible gene set.

The remaining of the text is organized as follows. Proposed methodology is described in Section II. Section III describes the performance evaluation measures. Section IV presents experimental results, and Section V concludes this research work.

II. PROPOSED METHODOLOGY

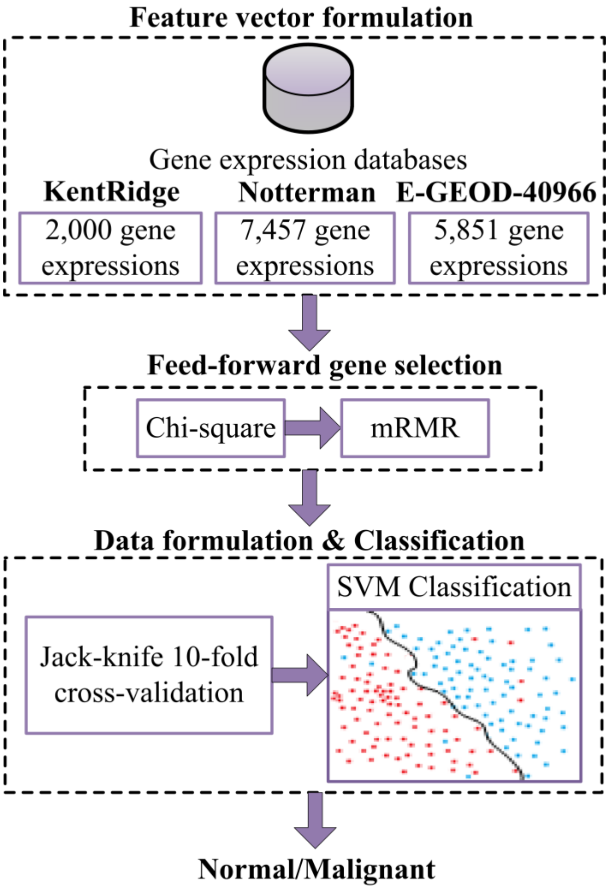
There are four main phases of the proposed technique, namely, (1) feature vector formulation, (2) gene selection, (3) training/testing data formulation, and lastly, (4) classification of colon gene samples into normal and malignant classes. Figure 1 portrays main phases of the proposed technique, and the subsequent text explains these phases in detail.

A. Feature vector formulation

The gene (also called gene expression) based datasets used in this research work are in raw format (simple database entries), therefore, gene expression values are aligned in the form of a feature vector for each sample. The resultant feature vectors are combined to develop a full-fledge dataset. One gene expression corresponds to one feature in the feature vector.

B. Gene selection

Gene expression based datasets have large dimensionality, therefore, a feed-forward gene selection approach comprising two feature selection strategies (Chi-Square and mRMR) has been adopted to reduce the dimensionality of the datasets by selecting meaningful and discerning gene expressions. These feature selection techniques have been explained in the following text.



Feature vector formulation
Gene expression databases
KentRidge
Notterman
E-GEOD-40966
2,000 gene expressions
7,457 gene expressions
5,851 gene expressions
Feed-forward gene selection
Chi-square
mRMR
Data formulation & Classification
SVM Classification
Jack-knife 10-fold cross-validation
Normal/Malignant

Figure 1: Schematic of the proposed technique

- mRMR

mRMR selects gene expressions, which show maximum relevancy to the corresponding target labels and possess minimum redundancy amongst them [12]. Usually, mutual information amongst the genes, and amongst the genes and the target labels is used to compute redundancy and relevancy of genes.

Consider a given dataset **D** having S training samples with each sample having G gene expressions, the redundancy within the dataset $R(\mathbf{D})$ is computed by averaging mutual information values between all the gene pairs.

$$R(\mathbf{D}) = \frac{1}{G^2}\sum_{i,j=1}^{G} I(\boldsymbol{e}_i,\boldsymbol{e}_j)\ ;\ where\ \boldsymbol{e}_i,\boldsymbol{e}_j \in \mathbf{D} \tag{1}$$

Where $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are the i^{th} and j^{th} gene expression vectors in **D,** and the term $I(\boldsymbol{e}_i, \boldsymbol{e}_j)$ shows the mutual information between the genes $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$. The mutual information is calculated using the following expression.

$$I(\boldsymbol{e}_i,\boldsymbol{e}_j) = \sum_{a,b} p(e_{i,a},e_{j,b})\log\left(\frac{p(e_{i,a},e_{j,b})}{p(e_{i,a})p(e_{j,b})}\right)\ ;\ where\ \mathrm{a,b} = 1,2,3,\ldots,\mathrm{S} \tag{2}$$

Where $e_{i,a}$ and $e_{j,b}$ are a^{th} and b^{th} elements of gene expression vector $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, respectively. The term $p(e_{i,a},e_{j,b})$ is the joint probability density function of $e_{i,a}$, and $e_{j,b}$. The terms $p(e_{i,a})$ and $p(e_{j,b})$ are marginal probability density functions of $e_{i,q}$ and $e_{j,b}$, respectively.

Similarly, the relevance $V(\mathbf{D},\boldsymbol{l})$ of the dataset **D** with label vector $\boldsymbol{l}$ is calculated by averaging all mutual information values between individual gene expressions $\boldsymbol{e}_i$ and the label vector $\boldsymbol{l}$ as follows.

$$V(\mathbf{D},\boldsymbol{l}) = \frac{1}{\mathrm{G}}\sum_{i=1}^{\mathrm{G}} I(\boldsymbol{e}_i,\boldsymbol{l}) \text{ ; } \textit{where } \boldsymbol{e}_i \in \mathbf{D} \tag{3}$$

$I(\boldsymbol{e}_i,\boldsymbol{l})$ is the mutual information between the gene expression $\boldsymbol{e}_i$ and label vector $\boldsymbol{l}$. It can be calculated using Equation (4).

$$I(\boldsymbol{e}_i,\boldsymbol{l}) = \sum_a p(e_{i,a},l_a)\log\left(\frac{p(e_{i,a},l_a)}{p(e_{i,a})p(l_a)}\right) \quad ; \textit{ where } \mathrm{a} = 1,2,3,\ldots,\mathrm{S} \tag{4}$$

Where $p(e_{i,a},l_a)$ is joint probability density function of $e_{i,a}$ and label l_a. The terms $p(e_{i,a})$ and $p(l_a)$ show the marginal probability density function of $e_{i,a}$, and the marginal probability mass function of l_a, respectively.

The objective of mRMR method is to find a set of gene expressions that results in maximum relevance V and minimum redundancy R. It is not possible to simultaneously achieve both the objectives, hence, Equation (5) develops a tradeoff between the objectives by uniting Equations (1) and (3) as follows.

$$mRMR = \max_{\mathbf{D}}\left[V(\mathbf{D},\boldsymbol{l}) - R(\mathbf{D})\right] = \max_{\mathbf{D}}\left[\frac{1}{\mathrm{G}}\sum_{i=1}^{\mathrm{G}} I(\boldsymbol{e}_i,\boldsymbol{l}) - \frac{1}{\mathrm{G}^2}\sum_{i,j=1}^{\mathrm{G}} I(\boldsymbol{e}_i,\boldsymbol{e}_j)\right] \tag{5}$$

Consider s_i be a set membership variable for gene vector $\boldsymbol{e}_i$, such that s_i=1 means presence and s_i=0 means absence of the gene $\boldsymbol{e}_i$ in the globally optimal set of genes, then Equation (5) may be converted to an optimization problem as follows.

$$mRMR = \max_{s\in\{0,1\}^{\mathrm{G}}}\left[\frac{\sum_{i=1}^{\mathrm{G}} I(\boldsymbol{e}_i,\boldsymbol{l})s_i}{\sum_{i=1}^{\mathrm{G}} s_i} - \frac{\sum_{i,j=1}^{\mathrm{G}} I(\boldsymbol{e}_i,\boldsymbol{e}_j)s_i s_j}{\left(\sum_{i=1}^{\mathrm{G}} s_i\right)^2}\right] \tag{6}$$

Thus, the gene set selected by mRMR is supposed to comprise genes, which not only possess maximum possible relevancy to the target labels, but have least possible redundancy as well [12].

- Chi-Square

Chi-square is an important method of feature selection. It assigns Chi-Square score to different features on the basis of their chi-square statistic, which is calculated with respect to the classes in the dataset. For the given dataset **D**, the chi-square score of the i^{th} gene expression is given by Equation (7).

$$Chi-Square_i = \sum_{r=1}^{\mathrm{R}} \frac{(\mathrm{P}_r^{\mathrm{N}}(i) - \mathrm{F}_r^{\mathrm{N}}(i))^2}{\mathrm{F}_r^{\mathrm{N}}(i)} + \sum_{r=1}^{\mathrm{R}} \frac{(\mathrm{P}_r^{\mathrm{M}}(i) - \mathrm{F}_r^{\mathrm{M}}(i))^2}{\mathrm{F}_r^{\mathrm{M}}(i)} \tag{7}$$

Where $\mathrm{P}_r^{\mathrm{N}}(i)$ and $\mathrm{P}_r^{\mathrm{M}}(i)$ are the number of normal and malignant samples lying in partition r, respectively. Likewise, $\mathrm{F}_r^{\mathrm{N}}(i)$ and $\mathrm{F}_r^{\mathrm{M}}(i)$ are the expected frequencies of $\mathrm{P}_r^{\mathrm{N}}(i)$ and $\mathrm{P}_r^{\mathrm{M}}(i)$, respectively, and are calculated using Equation (8). Chi-square, when operates on numeric attributes, requires the range of the attribute to be discretized into multiple partitions. Partitions of a single gene expression i are represented by r where r=1,2,…,R in Equation (7).

$$\mathrm{F}_r^{\mathrm{N}}(i) = \frac{\mathrm{S}^{\mathrm{N}}}{\mathrm{S}} \times \mathrm{P}_r^{\mathrm{S}}(i), \mathrm{F}_r^{\mathrm{M}}(i) = \frac{\mathrm{S}^{\mathrm{M}}}{\mathrm{S}} \times \mathrm{P}_r^{\mathrm{S}}(i) \tag{8}$$

S, S^{N} and S^{M} are the number of total samples, normal and malignant samples, respectively. Once chi-square scores for all the genes are determined, the genes are sorted in the descending order of their chi-square values, and a pre-defined number of top most genes are selected.

C. Training and Testing data formulation

This is an important phase of overall classification framework. In this article, Jackknife 10-fold cross-validation technique, which has been widely used in the past to measure the accuracy of prediction [13], has been used for formulation of training/testing data. In Jackknife test, data is divided into 10 folds out of which 9 folds involve in training, and the classes of the samples in the 10th fold are predicted based on the training performed on 9 folds. The process is performed 10 times, and the classes of all the samples are predicted.

D. Classification

The genes selected through feed-forward gene selection methodology are given as input to the classifier for classification of colon biopsy images into normal and malignant classes.

SVM classifier [14] has been widely used in the past for classification of medical images [15, 16]. It finds a decision surface amongst a set of candidate decision surfaces that possesses maximum distance to the closest points of the two classes in the training data set. In this work, linear kernel of SVM has been employed as classifier. Linear is a global kernel of SVM that caters the influence of far away points in classification. Let x_i and x_j be the training samples of the dataset, the linear kernel is defined as follows.

$$K(x_i,x_j) = x_i^T.x_j$$

Linear kernel has one adjustable parameter (c) that is the cost of the constraint violation associated with the data points occurring on the wrong side of decision surface.

III. Performance Evalaution Measures

Performance of the proposed classification technique has been measured in terms of the following well-known performance measures.

Accuracy is the ratio of number of correctly classified samples to the total samples [17]. It can be calculated by

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100$$

where true positive (TP), true negative (TN) are the number of correctly classified positive and negative samples, respectively. Similarly, false positive (FP) and false negative (FN) are incorrectly classified negative and positive samples, respectively.

The valid value of accuracy lies between 0 and 100, where 0 and 100, respectively, correspond to worst and best accuracy values.

Sensitivity is the capability of a classifier to identify positive samples [17]. It can be calculated by

$$Sensitivity = \frac{TP}{TP+FN}$$

Specificity is the capability of a classifier to identify negative samples [17]. It can be calculated by

$$Specificity = \frac{TN}{TN+FP}$$

The valid values of sensitivity and specificity lie between 0 and 1, where 0 is the worst and 1 is the best value.

MCC is another parameter that measures the performance of classification in binary class problem [18]. Its value ranges from -1 to +1. +1, -1 and 0 mean perfect, worst and random prediction. MCC can be calculated as follows.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP+FN)(TP+FP)(TN+FN)(TN+FP))}}$$

IV. Experimental Results

The proposed gene expression based classification has been carried out on three standard colon cancer datasets, namely, Notterman [19], KentRidge [20], and E-GEOD-40966 [21] datasets. The datasets have been obtained from publically available databases of gene expression based datasets. Brief details of the datasets have been summarized in Table I.

TABLE I. Datasets

Statistics for the datasets				
Dataset	***Ref.***	***Dimensionality***	***Number of samples***	
			Normal	***Malignant***
KentRidge	[19]	2000	22	**40**
Notterman	[20]	7457	18	**18**
E-GEOD-40966	[21]	5851	208	**142**

The gene expressions have been used for the classification of samples into normal and malignant classes. All the computations have been carried out on Core I7 Intel machine with 3.4GHz Processor and 12GB RAM. Matalb computational toolbox and Weka software have been used in the experiments.

Two different types of experiments have been carried out on the given datasets. In the first experiment, the proposed technique has been evaluated on the given datasets, and in the second experiment, previously published colon cancer detection techniques have been compared with the proposed technique. The experiments are described in the following text.

A. Experiment-I

In the first experiment, discerning gene expressions have been selected from the given datasets by using the proposed feed-forward gene selection mechanism.

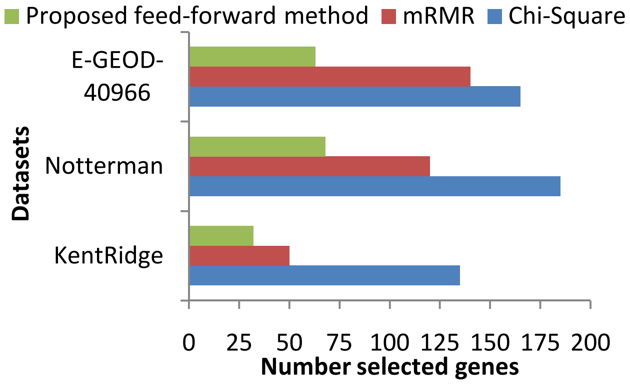


Figure 2: Number of genes selected by mRMR, Chi-Square, and the proposed feed-forward gene selection approach

The number of selected genes are shown in Figure 2. The genes have also been independently extracted by using mRMR and Chi-Square, and are shown in Figure 2 in order to demonstrate the effectiveness of the proposed gene selection mechanism. The results in the figure show that the proposed gene selection technique has considerably reduced the size of the datasets by selecting discerning genes compared to individual feature selection techniques.

The selected genes have been used for classification of samples into normal and malignant classes by using linear SVM. The performance of SVM classifier depends on the value of '*c*' parameter. The optimal value of parameter *'c'* has been obtained through grid search method for all the datasets pertaining to the genes selected by mRMR, Chi-Square and the proposed feed-forward gene selection method. The values obtained this way are later used for classification. Figure 3 shows the selection of optimal value of '*c*' for all the gene sets selected by the proposed technique.

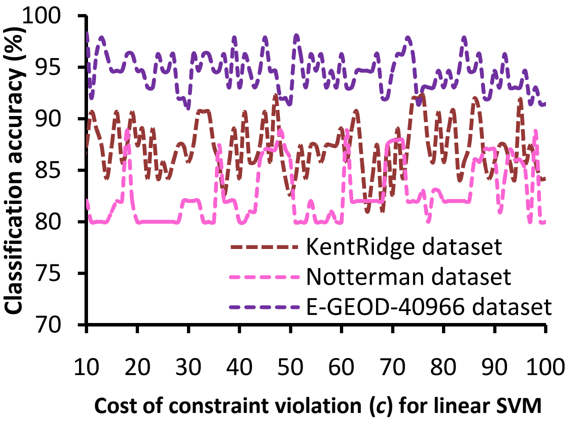


Figure 3: Selection of optimal value of *'c'* parameter of linear SVM for different datasets

Jack-knife 10-fold cross-validation technique has been used to model the testing and training data. The classification performance of the proposed technique for different datasets has been calculated in terms of various performance evaluation measures presented in Section III. The classification results are reported in Table II. The classification results have also been shown for genes selected by mRMR, and Chi-Square in order to demonstrate the effectiveness of the proposed gene selection mechanism.

TABLE II. Classification performance of datasets reduced using Chi-Square, mRMR and the proposed gene seelction method

	Accuracy	***Sensitivity***	***Specificity***	***MCC***
Chi-square				
KentRidge	82.26	0.80	0.86	0.64
Notterman	80.56	0.88	0.72	0.62
E-GEOD-40966	89.71	0.86	0.92	0.76
mRMR				
KentRidge	88.71	0.85	0.95	0.78
Notterman	86.11	0.89	0.83	0.72
E-GEOD-40966	90.57	0.88	0.92	0.80
Proposed feed-forward gene selection method				
KentRidge	91.94	0.90	0.95	0.83
Notterman	88.89	0.94	0.83	0.78
E-GEOD-40966	94.29	0.92	0.96	0.88

The classification results in Table II show that the proposed technique yields promising classification of gene expression based colon samples. The classification accuracy on KentRidge, Notterman, and E-GEOD-40966 datasets is 91.94%, 88.89% and 94.29%, respectively, by using the proposed technique. These accuracy values are higher compared to the accuracy achieved by using mRMR (KentRidge=88.71, Notterman=86.11, E-GEOD-40966=90.57) and Chi-Square (KentRidge=82.26, Notterman=80.56, E-GEOD-40966=89.71). The results show that mRMR and Chi-Square individually show good performance, but when the genes are selected in a cascade style fashion i.e. mRMR is applied on the genes selected by Chi-Square, the classification accuracy is further boosted. The results also reveal that selection of genes in a feed-forward manner not only reduces the size of the gene set, but also selects more discerning gene set compared to the gene sets selected by individual feature selection techniques.

We have also investigated the CPU time required for the classification of various datasets selected by mRMR, Chi-square and the proposed feed-forward gene selection technique. Figure 4 shows the results when linear SVM has been used for the classification of selected gene sets. The results show that classification time taken by linear SVM classifier depends on the size of the gene set. The gene sets selected by Chi-square are larger compared to those selected by mRMR, therefore, take more time for classification. Further, the results show that the gene set selected by the proposed feed-forward gene selection technique is not only discerning, but also takes lesser time for classification owing to its smaller size.

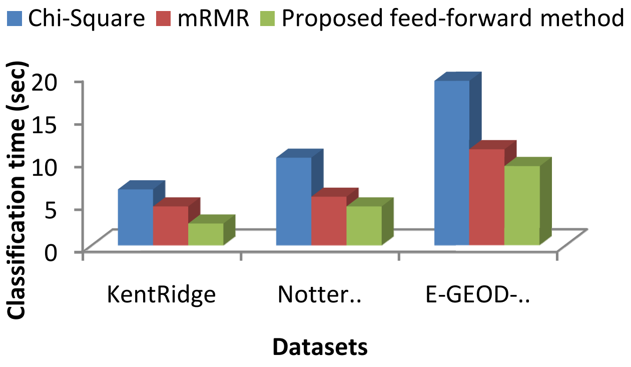
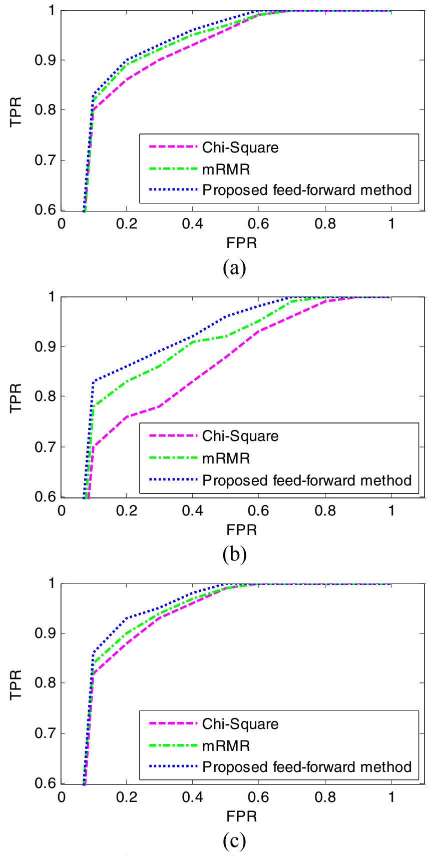


Figure 4: CPU time elapsed during classification of gene subsets selected by individual and the proposed feed-forward gene selection techniques

Performance of the mRMR, Chi-Square and the proposed gene selection method has also been investigated in terms of ROC curves. In this context, ROC curves for different datasets have been drawn, and are shown in Figure 5 for all the datasets. The results show that the ROC curve produced by the proposed gene selection technique is well above the ROC curves produced by individual gene selection mechanisms regardless of the underlying gene expression based dataset. This shows the supremacy of the proposed feed-forward gene selection method over individual feature selection methods of mRMR and Chi-Square.



(a)

(b)

(c)

Figure 5: ROC curve of (a) KentRidge, (b) Notterman, and (c) E-GEOD-40966 datasets for genes selected by individual and the proposed gene selection method

AUC is another parameter that can be measured from ROC. It shows the effectiveness of a classification technique. In this connection, we have also calculated the value of AUC from ROC curves drawn for different datasets. Figure 6 shows the AUC pertaining to gene sets selected by individual and the proposed gene selection technique. The figure shows higher values of AUC for the proposed technique.

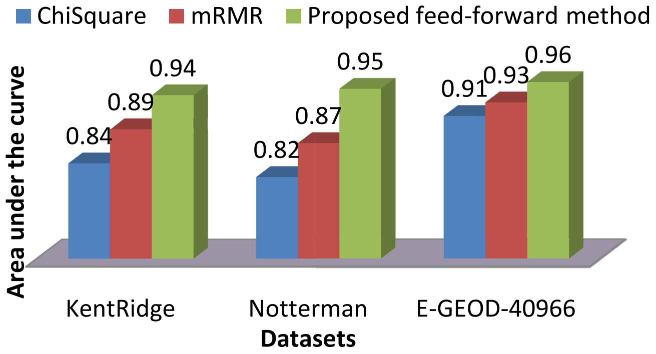


Figure 6: AUC for the classification based on gene sets selected by mRMR, Chi-Square and the proposed gene selection method

B. Experiment-II

In the second experiment, the performance of the proposed technique has been compared with previously published gene expression based techniques of colon cancer detection. To this end, three techniques, namely, Li et al. [5], Lee et al. [10], and

Tong et al. [11] have been implemented in Matlab, and evaluated on the given datasets. Table III provides the corresponding results.

TABLE III. COMPARISON OF THE PROPOSED TECHNIQUE WITH PREVIOUSLY PUBLISHED TECHNIQUES

	Ref.	*Accuracy*	*Sensitivity*	*Specificity*	*MCC*
KentRidge dataset					
Li et al.	[5]	89.01	0.88	0.90	0.87
Lee et al.	[10]	76.85	0.77	0.79	0.70
Tong et al.	[11]	90.32	0.82	0.95	0.79
Proposed technique	---	91.94	0.90	0.95	0.83
Notterman dataset					
Li et al.	[5]	80.56	0.78	0.83	0.61
Lee et al.	[10]	83.33	0.83	0.83	0.67
Tong et al.	[11]	88.89	0.89	0.89	0.78
Proposed technique	---	88.89	0.94	0.83	0.78
E-GEOD dataset					
Li et al.	[5]	88.57	0.89	0.88	0.77
Lee et al.	[10]	88.86	0.89	0.89	0.77
Tong et al.	[11]	91.14	0.93	0.88	0.82
Proposed technique	---	94.29	0.92	0.96	0.88

The results in Table III reveal that the proposed technique yields superior results compared to previous techniques for the given datasets. The proposed technique has achieved accuracy of 91.94%, 88.89% and 94.29% for KentRidge, Notterman and E-GEOD-40966 datasets, respectively. These results are better compared to the results achieved by previous techniques except the case of Tong et al. for Notterman dataset where similar performance has been obtained. Therefore, we can conclude that the proposed technique is better able to model the challenging problem of dimensionality reduction for gene based datasets compared to various previous techniques.

V. CONCLUSION

In this paper, we have proposed a gene expressions based classification technique that exploits the variations in gene expressions for classification of colon gene samples into normal and malignant classes, and reduces the dimensionality of gene based datasets in a simple and straightforward manner. The proposed gene selection method works in a feed forward manner, wherein two feature selection techniques are used one after the other. The genes selected by the first technique are fed as input to the second technique that selects genes from the given gene subset. The selected genes are then classified by using linear SVM. The feed forward approach of gene selection has shown improved performance on three standard datasets of colon gene expressions. It is observed that feed forward method of gene selection substantially reduces the feature space of KentRidge dataset from 2000 to 29 gene expressions compared to gene expressions selected by mRMR (50) and Chi-Square (135), thereby reducing the computational time from 6.54 sec (Chi-Square) and 4.35 sec (mRMR) to 2.54 sec. The proposed technique has also been compared with existing colon cancer detection techniques, and improved performance has been observed. Therefore, we hope that the proposed technique can be used as an effective tool for diagnosis of colon cancer. This research can further be extended into two directions. First, malignant colon samples may be classified into various cancer stages. Second, making ensemble of classifiers may potentially provide better performance.

REFERENCES

[1] "Colon Cancer Risk Factors," http://www.ccalliance.org/colorectal_cancer/riskfactors.html.

[2] G. D. Thomas, M. F. Dixon, N. C. Smeeton et al., "Observer Variation in the Histological Grading of Rectal Carcinoma," Journal of Clinical pathology, vol. 36, no. 4, pp. 385-391, 1983.

[3] S. Rathore, M. Hussain, A. Ali, A. Khan, "A recent survey on colon cancer detection techniques," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, pp. 545-563, 2013.

[4] S. Backert, M. Gelos, U. Kobalz *et al.*, "Differential Gene Expression in Colon Carcinoma Cells and Tissues Detected with a Cdna Array," *International Journal of Cancer,* vol. 82, pp. 868-874, 1999.

[5] L. Li, C. R. Weinberg, T. A. Darden *et al.*, "Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method," *Bioinformatics,* vol. 17, no. 12, pp. 1131-1142, 2001.

[6] Z. Chen, and J. Li, "A Multiple Kernel Support Vector Machine Scheme for Simultaneous Feature Selection and Rule-Based Classification," *Proc. 11th Pacific-Asia conference on Advances in knowledge discovery and data mining*, pp. 441-448, 2007.

[7] H. S. Shon, G. Sohn, K. S. Jung *et al.*, "Gene Expression Data Classification Using Discrete Wavelet Transform," *Proc. International Conference on Bioinformatics & Computational Biology*, pp. 204-208, 2009.

[8] E. T. Venkatesh, P. Thangaraj, and S. Chitra, "An Improved Neural Approach for Malignant and Normal Colon Tissue Classification from Oligonucleotide Arrays," *European Journal of Scientific Research,* vol. 54, pp. 159-164, 2011.

[9] A. Kulkarni, N. Kumar, V. Ravi *et al.*, "Colon Cancer Prediction with Genetics Profiles Using Evolutionary Techniques," *Expert Systems with Applications,* vol. 38, pp. 2752–2757, 2011.

[10] K. Lee, Z. Man, D. Wang *et al.*, "Classification of Bioinformatics Dataset Using Finite Impulse Response Extreme Learning Machine for Cancer Diagnosis," *Neural Computing and Applications,* vol. 22, pp. 457-468, 2013.

[11] M. Tong, K.H. Liu, C. Xu *et al.*, "An Ensemble of SVM Classifiers Based on Gene Pairs," *Computers in biology and medicine,* vol. 43, pp. 729-737, 2013.

[12] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226-1238, 2005.

[13] Saima Rathore, Mutawarra Hussain, Muhammad Aksam Iftikhar, Abdul Jalil, "Ensemble classification of colon ibopys images based on information rich hybrid descriptors," Computers in Biology and Medicine, vol. 47, pp. , 2013.

[14] V. Vapnik, "Statistical Learning Theory," Wiley-Interscience, 1998.

[15] Saima Rathore, Muhammad Aksam Iftikhar, Mutawarra Hussain, Abdul Jalil, "A novel clustering approach for colon biopsy images", IEEE conference on Frontiers of Information Technology, 2013, Islamabad, Pakistan.

[16] Saima Rathore, Muhammad Aksam Iftikhar, Mutawarra Hussain, Abdul Jalil, "Classification of colon biopsy images based on novel structural features," IEEE International Conference on Emerging Technologies, 2013, Islamabad, Pakistan.

[17] I.H. Witten, E. Frank, M.A. Hall, Data mining: Practical machine learning tools and techniques, in, Morgan Kaufmann Publishers, London, 2nd edition, 2005.

[18] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phase lysozyme, BioChemistry, BioPhysics, Acta 1975, vol. 405, pp. 442-451, 1975.

[19] "Colon Cancer Dataset Kent Ridge," 2013; http://datam.i2r.a-star.edu.sg/datasets/krbd/ColonTumor/ColonTumor.html.

[20] D. A. Notterman, U. Alon, A. J. Sierk *et al.*, "Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays," *Cancer Research,* vol. 61, no. 7, pp. 3124-3130, 2001.

[21] L. Marisa, A. d. Reyniès, A. Duval *et al.*, "Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value," *PLoS Med,* vol. 10, no. 5, 2013.