# GECC: Gene Expression Based Ensemble Classification of Colon Samples

Saima Rathore, Mutawarra Hussain, and Asifullah Khan

**Abstract**—Gene expression deviates from its normal composition in case a patient has cancer. This variation can be used as an effective tool to find cancer. In this study, we propose a novel gene expressions based colon classification scheme (GECC) that exploits the variations in gene expressions for classifying colon gene samples into normal and malignant classes. Novelty of GECC is in two complementary ways. First, to cater overwhelmingly larger size of gene based data sets, various feature extraction strategies, like, chi-square, F-Score, principal component analysis (PCA) and minimum redundancy and maximum relevancy (mRMR) have been employed, which select discriminative genes amongst a set of genes. Second, a majority voting based ensemble of support vector machine (SVM) has been proposed to classify the given gene based samples. Previously, individual SVM models have been used for colon classification, however, their performance is limited. In this research study, we propose an SVM-ensemble based new approach for gene based classification of colon, wherein the individual SVM models are constructed through the learning of different SVM kernels, like, linear, polynomial, radial basis function (RBF), and sigmoid. The predicted results of individual models are combined through majority voting. In this way, the combined decision space becomes more discriminative. The proposed technique has been tested on four colon, and several other binary-class gene expression data sets, and improved performance has been achieved compared to previously reported gene based colon cancer detection techniques. The computational time required for the training and testing of 208 $\times$ 5,851 data set has been 591.01 and 0.019 s, respectively.

**Index Terms**—Colon cancer, ensemble classification, gene expressions, PCA, mRMR, F-Score, chi-square

## 1 Introduction

Colon is a major constituent of large intestine and its cancer is quite common worldwide. Colon cancer arises due to abnormal growth of tissues in colon, which may turn into polyps. Polyps are usually benign, but some of them may catch malignancy if not treated in time. Colon cancer is primarily due to low intake of herby diet, and more intake of meat and fatty stuff. Some other factors of colon cancer include older age, chain smoking, and family history of colon cancer [1].

Microscopic inspection of colon biopsy samples is the classical method of cancer detection, however, it is time-consuming and laborious for the histopathologists, and have inter-observer/intra-observer variations in grading [2]. Therefore, automatic colon cancer detection techniques are in high demand. Researchers have been working since decades to propose reliable automatic methods of colon cancer detection. These methods have been summarized in a recent survey reported by Rathore et al. [3]. Some of these methods [4], [5] work on selected bands of hyperspectral colon data, extract a few discriminative features, and classify the sample into normal and malignant classes. There is another exciting method of colon cancer detection that is based on exploitation of heavy differences between composition of normal and malignant blood serum by using laser-induced fluorescence and Raman spectroscopy [6], [7]. The texture of normal and malignant colon samples (images) has also notable contrast, and researchers have utilized it to detect colon cancer. Morphological features [8], statistical features [9], and image fractal dimensions [10] based colon cancer detection schemes are the major representatives of this category. Furthermore, visual analysis of colon biopsy images is another method for colon cancer diagnosis [11].

Another promising method of colon cancer detection, which is also the focus of this research study, is the analysis of genes by using Oligonucleotide and cDNA microarrays (detailed working of microarrays will be explained in Section 2.1). Gene based data sets have been used in various research studies for diagnosis of cancer [12]. Similarly, physicians have also analyzed the human gene expressions for diagnosis of colon cancer by using microarrays, and identified many discerning genes responsible for detection of colon cancer [13], [14], [15]. But, these studies worked purely on finding discriminating genes amongst a pool of genes without an aim to actually identify the type (normal or malignant) of samples.

The discerning genes have been used to classify the samples into normal and malignant classes. In 1999, Backert et al. utilized 588 gene expressions, obtained from three classes (normal, non-mucinous and mucinous) of colon tissues [16]. But, their scheme yielded classification accuracy of slightly above 50 percent. Li et al. used genetic algorithm (GA) to identify discriminative genes, and achieved

- *S. Rathore is with the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan, and the Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan. E-mail: saimarathore_2k6@yahoo.com.*
- *M. Hussain and A. Khan are with the Department of Computer Science and Information Technology, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan. E-mail: {mutawarra, asif}@pieas.edu.pk.*

*Manuscript received 23 Aug. 2013; revised 26 May 2014; accepted 22 June 2014. Date of publication 5 Aug. 2014; date of current version 4 Dec. 2014. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2014.2344655*

classification accuracy of 94.1 percent by using k-nearest neighbor (KNN) classifier [17]. Chen and Li [18] used multiple kernel support vector machine (MK-SVM), where multiple kernels are described as the convex combination of the single kernels. Algorithm was tested on leukemia and colon tumor data sets, and more than 90 percent classification success was achieved for both the data sets. Research continued in this domain, and Shon et al. proposed using wavelet transformation for reduction of feature space [19]. They classified colon cancer data with probabilistic neural network (PNN), and obtained 92 percent accuracy.

Further, Venkatesh et al. proposed an EFJ-neural network for classification of KentRidge colon cancer data set [20]. The data distribution was 70 and 30 percent amongst training and testing data, respectively. Similarly, Kulkarni et al. proposed an evolutionary algorithms based method for detection of colon cancer [21]. In this work, t-statistic and mutual information were employed as feature selection strategies, and genetic programming and decision trees were used as classifiers. The results revealed that the combination of mutual information and genetic programming is promising compared to others.

Recently, Lee et al. proposed a finite impulse response extreme learning machine (FIR-ELM) based colon cancer detection technique [22], and achieved quite promising classification of colon samples. Further, Tong et al. proposed a method of colon cancer detection in which 50 gene pairs were selected using top scoring pair method, and linear SVM classifiers were trained on those pairs [23]. GA was employed to select such an optimal combination of SVM base classifiers that yields maximum possible performance. Tong et al. reported a classification accuracy of 90.30 percent with colon data set.

Microarrays though facilitate to analyze huge volume of data enabling insight into tissues and find the state of the cancer, however, microarrays based gene analysis poses two challenges. First, the major challenge in analysis of human genes is large dimensionality of the feature set under consideration. Therefore, efficient techniques are required to identify meaningful genes amongst a large pool of available genes. Second, many of the classifiers used previously to classify colon cancer data set did not perform reasonably well. Therefore, an efficient and robust classifier is needed that can divide the gene based samples into respective classes by considering selected genes. In this paper, we tackle both the issues quite reasonably. Different feature selection strategies such as minimum redundancy maximum relevancy (mRMR), principal component analysis (PCA), F-Score, and chi-square have been employed to select a discerning feature set quite capable to distinguish the two classes. Moreover, an SVM-ensemble based new approach for classification of colon gene expressions has been proposed, which is named GECC. In the proposed scheme, the individual SVM models are constructed through the learning of different SVM kernel functions such as linear, polynomial, RBF, and sigmoid. The predicted results of individual SVM models are then combined using majority voting. In this way, the combined decision of various models turns out to be better compared to individual decisions. SVM has been used for classification of genes expressions in the past [23], [24]. However, we have experimentally validated that making ensemble of SVM decision models for colon cancer detection is an interesting idea due to the chance of making a more discernible decision space.

TABLE 1
Abbreviations Used in the Text

| Acronym | Abbreviations |
|---|---|
| AUC | Area under the curve |
| CNS | Central nervous system |
| DLBCL | Diffuse large b-cell lymphoma |
| DNA | Deoxyribonucleic acid |
| KNN | K-nearest neighbor |
| mRMR | Minimum redundancy and maximum relevance |
| PCA | Principal component analysis |
| PNN | Probabilistic neural network |
| RBF | Radial basis function |
| RNA | Ribonucleic acid |
| ROC | Receiver operating characteristics |
| SVM | Support vector machines |

The remainder of this paper is organized as follows. Section 2 describes the detail of microarray experiment and a few notations/abbreviations used in the text. Section 3 presents the proposed GECC scheme in detail. Section 4 highlights the performance evaluation measures. Section 5 demonstrates experimental results, and Section 6 concludes the paper.

## 2 Preliminaries

This section provides a healthy supporting material to understand the working of microarrays. Additionally, it also presents different abbreviations, which are used in the document. Table 1 summarizes these abbreviations.

### 2.1 Microarrays

A microarray is a collection of multiple spots on a glass slide and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. Most common method of measuring gene expressions from microarrays is to compare genes expressions of one cell inhibited under certain condition (sample 1) to those of the reference cell maintained in normal condition (sample 2). RNA molecules of both the samples are reverse transcribed into cDNA by using an enzyme reverse transcriptase and nucleotides, labeled with different fluorescent dyes. Once both the samples become uniquely identifiable by using labels, they are hybridized on to the same glass slide, and the locations in the hybridized microarray are excited by a laser. The amount of fluorescence emitted upon excitation corresponds to the amount of bound nucleic acid. The final output of the microarray experiment is an image in which each location that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene. Fig. 1 presents a sample tiff image obtained after a microarray experiment. Each row represents one unique sample, wherein each column represents gene expressions corresponding to the sample. The gene expressions are measured from the image, and are stored in database against respective samples.

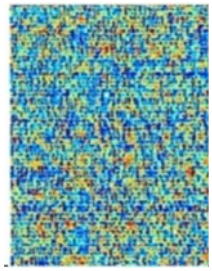

Fig. 1. Tiff image generated by a microarray experiment showing genes expressions for various samples.

## 3 Proposed GECC Scheme

The proposed GECC scheme is an effective combination of feature selection strategy coupled with ensemble classification. It comprises four distinct phases; (1) gene expression based feature vector formulation, (2) feature selection using various feature selection strategies, (3) training/testing data formulation, and finally, (4) ensemble classification of genes based samples into normal and malignant categories. Fig. 2 represents top-level architecture of the GECC, and the following text explains its different phases in detail.

In the proposed GECC scheme, four standard data sets (explained in Section 3.1) have been used. The sequence of operations is identical for all the data sets except for the gene selection phase, which is not executed for BioGPS data set owing to its smaller dimensionality. Once discriminative genes are selected, data processing phase assigns target labels to the samples. Multiple SVM kernels are employed to predict the class of a given sample, and then the predictions of individual SVM models are combined through majority voting. In this work, it has been shown through experimentation that a carefully selected combination of feature selection strategy and ensemble classification for identification of gene based samples prove to be an effective solution.

Next few sections explain different phases of the proposed scheme in detail. Several symbols have been used in subsequent sections, therefore, in order to make the document readable and understandable, commonly used symbols are listed in Table 2.

### 3.1 Data Set

Gene expression based classification has been conducted on four standard colon cancer data sets, namely, KentRidge [25], BioGPS [26], Notterman [27], and E-GEOD-40966 [28]. These data sets have been acquired from publically accessible gene expression databases. These are raw data sets, therefore, two subsequent steps, namely, gene expression based feature vector formulation (Section 3.2), and discerning genes selection (Section 3.3) are applied on these data sets to make them suitable for classification. The following text provides a brief description of these data sets.

#### 3.1.1 KentRidge Data Set

This data set [25] comprises 40 malignant and 22 normal samples. Its dimensionality is 2,000. KentRidge is a preprocessed data set in which 2,000 (out of 6,500) gene expressions have already been selected in a clinical study [13]. However, to further reduce the computational burden and

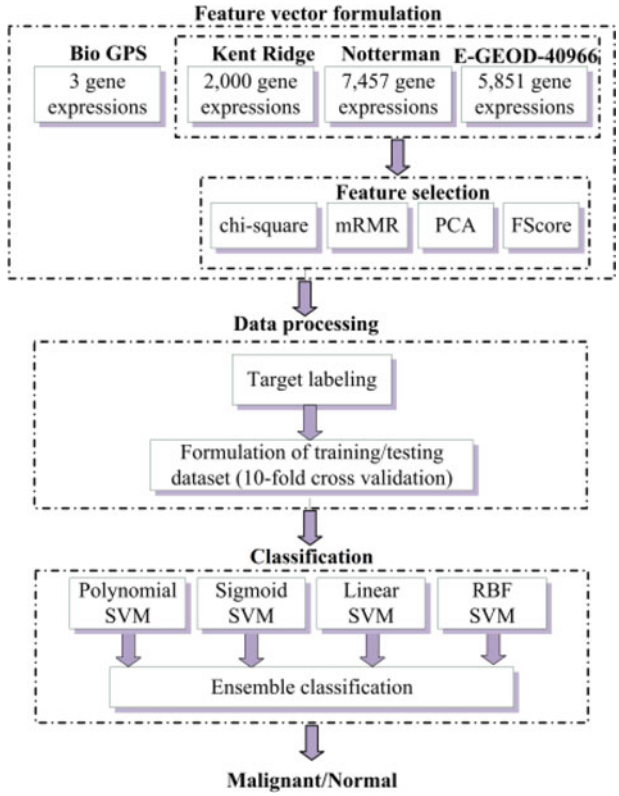


Fig. 2. Top level layout of the proposed GECC technique.

TABLE 2
Symbols Used in the Text

| Symbol | Description |
|---|---|
| $\mathbf{X}$ | Data set comprising S samples |
| $t$ | Target label vector $t = [t_1, t_2, \ldots, t_S]^T$ comprising S labels corresponding to $S$ samples in $\mathbf{X}$; $t_s \in \{-1, +1\}, s = 1, 2, \ldots, S$ |
| $\mathbf{X}^N$ | Data set comprising normal samples of $\mathbf{X}$ |
| $\mathbf{X}^M$ | Data set comprising malignant samples of $\mathbf{X}$ |
| S | Number of samples in $\mathbf{X}$ |
| $S^N$ | Number of samples in $\mathbf{X}^N$ |
| $S^M$ | Number of samples in $\mathbf{X}^M$ |
| $S^P$ | Number of support vectors |
| $\mathbf{X}_n^N(j)$ | $j$th feature value of the $n$th sample in the data set $\mathbf{X}^N$ |
| $\mathbf{X}_m^M(j)$ | $j$th feature value of the $m^t$th sample in the data set $\mathbf{X}^M$ |
| $\mu^S(j)$ | Mean of $j$th feature of all the samples in $\mathbf{X}$ *where* $j = 1, 2, \ldots, J$ |
| $\mu^N(j)$ | Mean of $j$th feature of normal samples in $\mathbf{X}^N$ *where* $j = 1, 2, \ldots, J$ |
| $\mu^M(j)$ | Mean of $j$th feature of malignant samples in $\mathbf{X}^M$ *where* $j = 1, 2, \ldots, J$ |
| T | Number of partitions for a given feature |
| $P_t^S(j)$ | Number of samples in partition $t$ for $j$th feature in $\mathbf{X}$where $t = 1, 2, \ldots, T$ |
| $P_t^N(j)$ | Number of normal samples in partition $t$ for $j$th feature in $\mathbf{X}^N$where $t = 1, 2, \ldots, T$ |
| $P_t^M(j)$ | Number of malignant samples in partition $t$ for $j$th feature in $\mathbf{X}^M$ where $t = 1, 2, \ldots, T$ |
| F | Number of folds for SVM decision models |
| N | Number of neighbors for KNN decision model |

to select most meaningful genes, feature selection process is once again applied on the data set.

#### 3.1.2 BioGPS Data Set

BioGPS data set [26] comprises 131 samples out of which 37 are normal and 94 are malignant. This data set has only three discriminative gene expressions. Therefore, no gene selection process is applied on it due to its relatively smaller size compared to other data sets.

#### 3.1.3 Notterman Data Set

Notterman colon cancer data set [27] has been taken from gene expression project of Princeton University, New Jersey, USA. This data set comprises 36 samples out of which 18 samples are normal, and remaining 18 samples are malignant. Each sample has 7,457 genes.

#### 3.1.4 E-GEOD-40966 Data Set

E-GEOD-40966 data set [28] doesn't comprise ordinary malignant and normal samples just like other data sets. Rather, this data set has 463 malignant colon samples of different cancer stages. E-GEOD-40966 data set has been acquired from ArrayExpress repository of gene expressions. We have picked 208 stage 2 and 142 stage 3 patients from the data set in order to develop a binary-class data set. Dimensionality of E-GEOD-40966 data set is also very large; each sample within the data set has 5,851 gene expressions.

### 3.2 Genes Expressions Based Feature Vector Formulation

The main objective of this stage is to formulate a feature vector for every colon sample. A feature vector is desired to be as small as possible but at the same time should contain the features, which are discriminative enough to classify given samples into their respective classes with good accuracy. In this particular research study, we are dealing with gene expressions, therefore, feature vector has been developed by aligning values of gene expressions for each sample in sequence. One gene expression means one feature in the feature vector, therefore, features, genes, and gene expressions will be used interchangeably in the text. Likewise, gene expression profile comprising multiple gene expressions means one feature vector.

### 3.3 Gene Expression Profile Reduction Methods

Gene expressions found through clinical research studies are usually larger in number. These larger and imbalance gene expressions generally create problems for the decision models in accurately predicting the samples if they are used for the classification without any pre-processing. Therefore, prior to classification, dimensionality of gene expression profile must be reduced in order to provide only the meaningful gene expressions to the decision models. Several advanced gene selection techniques have been proposed in the contemporary literature [29], [30], but they are not fully established. Therefore, we have employed the following four state-of-the-art feature selection strategies.

#### 3.3.1 mRMR

mRMR selects gene expressions (features) while trying to maximize the inter class (genes of two classes) and minimize the intra class (genes of one class) proximities. It accomplishes this by selecting genes, which show maximum relevancy to target labels and have minimum redundancy amongst them [31]. Usually, mutual information amongst the genes as well as amongst the genes and the target labels can be utilized to calculate relevancy and redundancy scores of genes.

For a given data set $\mathbf{X}$ comprising S training samples of J gene expressions each, the redundancy of the data set $R(\mathbf{X})$ is the average value of all mutual information values between all the gene pairs.

$$R(\mathbf{X}) = \frac{1}{\mathrm{J}^2}\sum_{i,j=1}^{\mathrm{J}} I(\boldsymbol{g}_i, \boldsymbol{g}_j); \quad where \ \ \boldsymbol{g}_i, \boldsymbol{g}_j \in \mathbf{X}. \tag{1}$$

Where $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ represent the $i$th and $j$th gene expression vectors in $\mathbf{X}$, and $I(\boldsymbol{g}_i, \boldsymbol{g}_j)$ represents the mutual information between the genes $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$, which can be calculated using the following expression.

$$I(\boldsymbol{g}_i, \boldsymbol{g}_j) = \sum_{x,y} p(g_{i,x}, g_{j,y}) \log\left(\frac{p(g_{i,x}, g_{j,y})}{p(g_{i,x})p(g_{j,y})}\right); \quad where \quad x, y = 1, 2, 3, \ldots, \mathrm{S}. \tag{2}$$

Where $g_{i,x}$ and $g_{j,y}$ are $x$th and $y$th elements of gene expression vector $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$, respectively. $p(g_{i,x}, g_{j,y})$ shows the joint probability density function of $g_{i,x}$, and $g_{j,y}$. The terms $p(g_{i,x})$ and $p(g_{j,y})$ represent marginal probability density functions of $g_{i,x}$ and $g_{j,y}$, respectively.

Similarly, for the target label vector $\boldsymbol{t}$ comprising labels of S samples, the relevance of the data set $\mathbf{X}$ with $\boldsymbol{t}$, denoted by $V\mathbf{X}, \mathbf{t}$, is defined by the average value of all mutual information values between individual gene expressions $\boldsymbol{g}_i$ and the label vector $\boldsymbol{t}$ as follows.

$$V(\mathbf{X}, \mathbf{t}) = \frac{1}{\mathrm{J}}\sum_{i=1}^{\mathrm{J}} I(\boldsymbol{g}_i, \boldsymbol{t}); \quad where \ \boldsymbol{g}_i \in \mathbf{X}. \tag{3}$$

$I(\boldsymbol{g}_i, \boldsymbol{t})$ denotes the mutual information between the gene expression $\boldsymbol{g}_i$ and label vector $\boldsymbol{t}$. It can be calculated using the following equation:

$$I(\boldsymbol{g}_i, \boldsymbol{t}) = \sum_{x} p(g_{i,x}, t_x) \log\left(\frac{p(g_{i,x}, t_x)}{p(g_{i,x})p(t_x)}\right); \quad where \ \ x = 1, 2, 3, \ldots, \mathrm{S}. \tag{4}$$

Here $p(g_{i,x}, t_x)$ is joint probability density function of $g_{i,x}$ and label $t_x$. The terms $p(g_{i,x})$ and $p(t_x)$ show the marginal probability density function of $g_{i,x}$, and the marginal probability mass function of $t_x$, respectively.

The objective is to select the set of gene expressions which yields maximum relevance $V$ and minimum redundancy $R$. As both the objectives are usually not achievable simultaneously, therefore, Equation (5) establishes a trade-off between the two objectives by combining Equations (1) and (3) as follows:

$$\begin{aligned} mRMR &= \max_{\mathrm{X}}[V(\mathbf{X}, \boldsymbol{t}) - R(\mathbf{X})] \\ &= \max_{\mathbf{X}}\left[\frac{1}{\mathrm{J}}\sum_{i=1}^{\mathrm{J}} I(\boldsymbol{g}_i, \boldsymbol{t}) - \frac{1}{\mathrm{J}^2}\sum_{i,j=1}^{\mathrm{J}} I(\boldsymbol{g}_i, \boldsymbol{g}_j)\right]. \end{aligned} \tag{5}$$

Let $m_i$ be the set membership indicator function for gene expression vector $\boldsymbol{g}_i$, so that $m_i = 1$ indicates presence and $m_i = 0$ indicates absence of the gene expression $\boldsymbol{g}_i$ in the globally optimal gene set, then Equation (5) may be written as an optimization problem as follows:

$$mRMR = \max_{m \in \{0,1\}^J} \left[ \frac{\sum_{i=1}^{\mathrm{J}} I(\boldsymbol{g}_i, \boldsymbol{t}) m_i}{\sum_{i=1}^{\mathrm{J}} m_i} - \frac{\sum_{i,j=1}^{\mathrm{J}} I(\boldsymbol{g}_i, \boldsymbol{g}_j) m_i m_j}{\left(\sum_{i=1}^{\mathrm{J}} m_i\right)^2} \right]. \tag{6}$$

Thus, the set of gene expressions determined using mRMR is expected to contain values which not only bear maximum possible relevancy to the target labels, but are non-redundant as well [31].

### 3.3.2 F-Score

F-Score [32] is the simplest of all feature selection strategies. It tries to simultaneously minimize the intra-class distance and maximize the inter-class distance. Given S data samples, if the number of normal and malignant samples are $\mathrm{S}^{\mathrm{N}}$ and $\mathrm{S}^{\mathrm{M}}$, respectively, then F-Score of the $j$th feature is given in the following equation:

$$FScore_j = \frac{(\mu^{\mathrm{N}}(j) - \mu^{\mathrm{S}}(j))^2 + (\mu^{\mathrm{M}}(j) - \mu^{\mathrm{S}}(j))^2}{\frac{1}{\mathrm{S}^N - 1}\sum_{n=1}^{\mathrm{S}^{\mathrm{N}}} (X_n^{\mathrm{N}}(j) - \mu^{\mathrm{N}}(j))^2 + \frac{1}{\mathrm{S}^{\mathrm{M}} - 1}\sum_{m=1}^{\mathrm{S}^{\mathrm{M}}} (\mathrm{X}_m^M(j) - \mu^{\mathrm{M}}(j))^2}, \tag{7}$$

where $j = 1, 2, \ldots, \mathrm{J}$. The terms $\mu^{\mathrm{N}}(j)$,$\mu^{\mathrm{M}}(j)$ and $\mu^{\mathrm{S}}(j)$ are the average of the $j$th feature of the normal, malignant and total samples, respectively. The terms $\mathrm{X}_n^N(j)$ and $\mathrm{X}_m^M(j)$ correspond to individual values of the $j$th features for $n$th and $m$th samples of normal and malignant classes, respectively. The larger the F-Score, the more discriminative the feature is.

### 3.3.3 Principal Component Analysis

Principal component analysis, initially proposed by Pearson, is a mathematical (orthogonal) transformation that transforms a data set comprising correlated variables into a set of linearly uncorrelated variables [33]. These linearly uncorrelated variables are called principal components. Such an orthogonal transformation makes sure that first principal component bears maximum variability in data, and each succeeding component also bears maximum possible variability while maintaining orthogonality with preceding principal components.

### 3.3.4 Chi-Square

Chi-square is another promising method of feature selection that evaluates different features on the basis of their chi-square statistic with respect to the classes in the data set. For the given data set **X**, the chi-square score of the $j$th feature (gene) is given by Equation (8).

$$Chi - Square_j = \sum_{t=1}^{\mathrm{T}} \frac{(\mathrm{P}_t^{\mathrm{N}}(j) - \mathrm{E}_t^{\mathrm{N}}(j))^2}{\mathrm{E}_t^{\mathrm{N}}(j)} + \sum_{t=1}^{\mathrm{T}} \frac{(\mathrm{P}_t^{\mathrm{M}}(j) - \mathrm{E}_t^{\mathrm{M}}(j))^2}{E_t^{\mathrm{M}}(j)}. \tag{8}$$

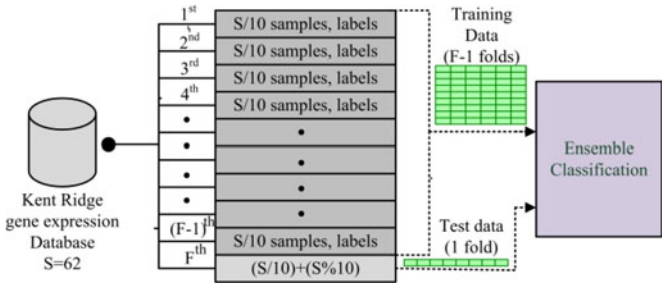


Fig. 3. Formulation of training and testing data through Jackknife cross-validation.

Where $\mathrm{P}_t^{\mathrm{N}}(j)$and $\mathrm{P}_t^{\mathrm{M}}(j)$as defined in Table 2, respectively, are the number of samples of normal and malignant classes lying in partition $t$. Likewise, $\mathrm{E}_t^{\mathrm{N}}(j)$ and $\mathrm{E}_t^{\mathrm{M}}(j)$, respectively, are the expected frequencies of $\mathrm{P}_t^{\mathrm{N}}(j)$ and $\mathrm{P}_t^{\mathrm{M}}(j)$, and are calculated using Equation (9). 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 $j$ are represented by $t$ where $t = 1, 2, \ldots, T$ in equation (8).

$$\mathrm{E}_t^{\mathrm{N}}(j) = \frac{\mathrm{S}^{\mathrm{N}}}{\mathrm{S}} \times \mathrm{P}_t^{\mathrm{S}}(j), \mathrm{E}_t^{\mathrm{M}}(j) = \frac{\mathrm{S}^{\mathrm{M}}}{\mathrm{S}} \times \mathrm{P}_t^{\mathrm{S}}(j). \tag{9}$$

Once chi-square scores for the entire gene expressions are calculated, gene expressions are sorted in the descending order of their chi-square values. Desired number of top most gene expressions are selected as larger the chi-square score, better discerning the gene is.

## 3.4 Training/Testing Data Formulation

Once discerning gene expressions are selected, next comes the issue of data formulation. In this work, Jackknife cross-validation technique has been employed for classification. It is a commonly practiced technique that has been successfully used in the past to validate the accuracy of prediction [34]. In Jackknife test, data are divided into F folds. F-1 folds participate in training, and the classes of the samples belonging to the remaining fold are predicted based on the training performed on F−1 folds. This sampling process is repeated F times and the class of each sample is predicted. In this work, 10-fold cross-validation scheme has been employed for classification. Fig. 3 presents Jackknife cross-validation process, whereby data have been divided into 10 folds. Each fold hosts S/10 samples except the last one that may house less than S/10 samples.

## 3.5 Decision Modeling

SVM was originally proposed by Vapnik [35], and has been successfully used in medical diagnosis [36], [37]. In this work, linear, RBF, sigmoid and polynomial kernels of SVM have been employed. SVM kernels have been chosen after an exhaustive experimental process. We have evaluated several other classifiers such as decision trees, KNN, and PNN in addition to SVM, and observed that SVM kernels yield superior results compared to others. Therefore, for optimized performance, we have selected four variants of SVM. The following text briefly describes the working of SVM.

Consider a training data set $\mathbf{Q} \in \mathbf{X}$ comprising Z training sample $\boldsymbol{q_1}, \boldsymbol{q_2}, \ldots, \boldsymbol{q_Z}$, and target labels $\boldsymbol{t} = [t_1, t_2, \ldots, t_Z]^{\mathrm{T}}$ where $\boldsymbol{t}_{\mathrm{Z}} \in \{-1, +1\}, z = 1, 2, \ldots, \mathrm{Z}$. The data to be classified may be linearly or non-linearly separable. For linearly separable data, the aim of classification is to design a linear decision surface (given in Equation (10)), which correctly classifies the training samples.

$$f(\boldsymbol{q}) = \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{q} + b\,ias = 0. \tag{10}$$

However, such a decision surface, defined by its direction/weight vector $\boldsymbol{w}$ and position (bias) in the space, may not be unique. Therefore, the objective is to select a direction $\boldsymbol{w}$ such that the distance of the surface to the nearest points of the two classes is maximum. The nearest points are called support vectors, and the distance of the nearest points from the surface is called margin. For binary classification of samples into normal and malignant classes, candidate decision surfaces are normalized in such a way that value of $f(\boldsymbol{q})$ for the support vectors is equal to $+1$ for malignant class, and $-1$ for normal class. The problem can be solved using optimization techniques for non-linear objective function subjected to linear inequalities [38].

$$\begin{aligned} &minimize\, \|\boldsymbol{w}\|^2 \\ &\quad subject\ to\ t_{\mathrm{Z}}\big(\boldsymbol{w}^T \boldsymbol{q}_{\mathrm{Z}} + bias\big) \geq 1; \quad z = 1, 2, \ldots, \mathrm{Z}, \end{aligned}$$

where $\boldsymbol{w}$ is a linear combination of the support vectors.

For linearly non-separable data, there are three cases. First, training points may fall on correct side of the decision surface and behind margin. Second, training points may fall on correct side of the surface, but inside margin. Third, training points may fall on wrong side of the surface. The objective is to select a decision surface such that second and third cases could be minimized. A penalty term is added for this purpose. Let $\xi = [\xi_1, \xi_2, \ldots, \xi_z]$ be a vector comprising error terms corresponding to Z training samples in the data set. Therefore, the problem for linearly non-separable data may be formulated as:

$$\begin{aligned} &minimize \|\boldsymbol{w}\|^2 + c \sum_{z=1}^{\mathrm{Z}} \xi_{\mathrm{z}} \\ &subject\ to\ t_z(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{q}_{\mathrm{z}} + bias) \geq 1 - \xi_z; \quad \mathrm{z} = 1, 2, \ldots, \mathrm{Z}, \end{aligned}$$

where $\xi_z = 0$, $0 < \xi_z < 1$, and $\xi_z > 1$ for points corresponding to first, second and third cases, respectively. The term $c$ is the penalty parameter associated with the penalty term $\sum_{z=1}^{\mathrm{Z}} \xi_z$.

When data are not linearly separable, SVM, a non-linear classifier, may be used that maps the data from lower dimension J to a higher dimension $\mathrm{J}^*$ through a non-linear mapping $\Phi(\boldsymbol{q})$ so that $\Phi : R^{\mathrm{J}} \rightarrow R^{\mathrm{J}^*}$, $\mathrm{J}^* \gg \mathrm{J}$. Suppose, $\boldsymbol{q}$ and $\boldsymbol{r}$ are the training samples, a non-linear decision surface $f(\boldsymbol{q})$ between the classes can be constructed in terms of kernel functions [38]:

$$f(\boldsymbol{q}) = \sum_{z=1}^{\mathrm{S^P}} \alpha_z t_z K(\boldsymbol{q}, \boldsymbol{r}) + bias = \sum_{z=1}^{\mathrm{S^P}} \alpha_z t_z \Phi(\boldsymbol{q}) \cdot \Phi(\boldsymbol{r}) + bias, \tag{11}$$

where $\mathrm{S^P}$ is the number of support vectors. $\alpha_z$ and $t_z$, respectively, are the Lagrange multipliers and target labels associated with the support vectors.

The kernel functions of SVM are either local or global. For local kernels, only the data samples that are in proximity of each other influence the kernel values. Whereas, in case of global kernels, samples far away from each other still have an influence on the kernel values. To introduce diversity in the ensemble classifier, we have used local (RBF) as well as global kernels (linear, polynomial and sigmoid). Linear, RBF, sigmoid and polynomial kernels can be mathematically defined by Equations (12), (13), (14) and (15), respectively.

$$K(\boldsymbol{q}, \boldsymbol{r}) = \boldsymbol{q}^{\mathrm{T}} \cdot \boldsymbol{r}, \tag{12}$$

$$K(\boldsymbol{q}, \boldsymbol{r}) = \exp(-\gamma \|\boldsymbol{q} - \boldsymbol{r}\|^2), \tag{13}$$

$$K(\boldsymbol{q}, \boldsymbol{r}) = \tanh(\gamma \boldsymbol{q}^{\mathrm{T}} \cdot \boldsymbol{r} + r), \tag{14}$$

$$K(\boldsymbol{q}, \boldsymbol{r}) = [\gamma \boldsymbol{q}^{\mathrm{T}} \cdot \boldsymbol{r} + r]^g. \tag{15}$$

All these SVM kernel functions share one common cost parameter $c$, which is the constraint violation cost associated with the data point occurring on wrong side of the boundary. The parameter $\gamma$ in the RBF, sigmoid and polynomial kernel functions controls the shape of the separating hyper plane. Increasing $\gamma$ usually increases number of support vectors. The parameter $g$ is the degree of polynomial kernel, and $r$ is the offset of polynomial and sigmoid kernels. Selection of optimal values of SVM parameters will be discussed in Section 5.2.1.

In order to validate the effectiveness of the selected decision models for the task at hand, performance of GECC has also been compared with a few state-of-the-art decision models, like, PNN, KNN and decision tree. Detailed information on these decision models can be found in Duda's classical book [39].

### 3.6 Majority Voting Based Ensemble Classification

Recently, ensemble classification has become popular in medical diagnosis due to its pre-eminence over single classifier based systems [40]. The major advantage of ensemble classification is that it utilizes diversity of individual models. The proposed ensemble scheme has been developed by using the concept of stacking the predicted labels of individual SVM models. This way, a new decision space has been constructed that is expected to be more discerning compared to the original one.

In this research study, the individual SVM models are trained on data set $\mathbf{X}$ and their predictions are noted down. Suppose the predicted labels of linear, RBF, sigmoid and polynomial decision models for the input data set $\mathbf{X}$ are $\boldsymbol{l}^{\mathrm{L}}$, $\boldsymbol{l}^{\mathrm{R}}$, $\boldsymbol{l}^{\mathrm{O}}$ and $\boldsymbol{l}^{\mathrm{Y}}$ column vectors of size S. Since, there are two class types (–1 for normal and +1 for malignant), therefore, each individual element of the predicted label vectors has either value $-1$ or $+1$. Majority voting algorithm is then used to combine the predictions of individual SVM models. In this step, $-1$ and $+1$ labels assigned by the individual decision models for each sample are counted, and then

based on the majority of votes, $-1$ or $+1$ label is assigned to the sample. However, an equal value of votes for $-1$ and $+1$ means that a tie exists amongst decision models. These samples are termed as hard samples, and the concept of weighted majority voting [41] has been introduced in GECC in order to tackle these samples.

In weighted majority voting, weights have been determined for different SVM decision models based on their individual performances. Genetic algorithm has been designed to run for 200 iterations, and to find such an optimal combination of weights for different decision models that gives optimal ensemble classification results. The weights are selected such that the summation of all the weights remains equal to one. The optimal weights found by GA for linear, RBF, sigmoid and polynomial kernels are 0.210, 0.240, 0.295, and 0.255, respectively. The classes of the samples have been found by adding weights of the decision models separately for normal and malignant labels. Two conclusions can be drawn from the optimized weights of SVM kernels. First, weight of any single SVM kernel never exceeds the sum of weights of the other three kernels. Second, weight of sigmoid kernel in combination with any other kernel exceeds the sum of weights of rest of the two kernels. Therefore, voting preference has been given to the sigmoid SVM in case of these hard samples. Let $\boldsymbol{p}$ (a column vector of size S) represents the result of ensemble classification, and $\boldsymbol{y}^{-1}$ and $\boldsymbol{y}^{+1}$ vectors contain the number of normal and malignant votes for all the samples, respectively. The predicted label $p_s$ is assigned to the $s$th sample depending upon the values of $y_s^{+1}$ and $y_s^{-1}$, and in case of tie, the label predicted by sigmoid SVM for $s$th sample i.e. $l_s^O$ is used. The whole process is summarized as follows:

$$p_s = \begin{cases} +1, & if \ \mathrm{y}_s^{+1} > \mathrm{y}_s^{-1} \\ -1, & if \ \mathrm{y}_s^{-1} > \mathrm{y}_s^{+1}; \\ l_s^O, & if \ \mathrm{y}_s^{-1} = \mathrm{y}_s^{+1} \end{cases} \quad s = 1, 2, \ldots, \mathrm{S}.$$

## 4 Performance Evaluation Measures

Results have been evaluated using well-known performance measures such as accuracy, sensitivity, specificity, Matthews's correlation coefficient (MCC), and F-Measure. The calculation of these measures involves number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TN and TP are the number of correctly classified negative and positive samples. FN and FP are the number of incorrectly classified positive and negative samples.

Accuracy is a measure of overall effectiveness of the classification scheme. It can be calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100. \tag{16}$$

Sensitivity is the ability of a classifier to recognize patterns of positive class. It can be calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}. \tag{17}$$

Specificity is the ability of a classifier to recognize patterns of negative class. It can be calculated as follows:

$$Specificity = \frac{TN}{TN + FP}. \tag{18}$$

MCC serves as a measure of classification in binary class problems. Its value ranges from $-1$ to $+1$. $+1$ means classifier always predicts a right label, whereas $-1$ means classifier always commits a mistake. However, 0 means random prediction. MCC can be calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FN)(TP + FP)(TN + FN)(TN + FP))}}. \tag{19}$$

F-Measure makes use of precision and recall to estimate accuracy of classification.

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}.$$

F-measure can be calculated by using Equation (20). It is a weighted average of precision and recall values. Its value ranges between 0 and 1, where 0 is the worst possible score and 1 is the best possible.

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{20}$$

Accuracy, sensitivity, specificity, and F-Measure will be abbreviated as acc, sens, spec, and FM at some places in Section 5. Similarly, sigmoid and polynomial SVM will be shortened as sigm and poly in Section 5.

## 5 Experimental Results

The proposed GECC technique has been tested on various standard colon cancer data sets. The gene expressions, selected by various feature selection strategies, are given to the algorithm as an input for ensemble classification through majority voting scheme. All the computations have been performed on Intel Core i7 with 3.4 GHz processor and 12 GB RAM.

### 5.1 Selection of Discriminative Gene Expressions from Data Sets Having High Dimensionality

KentRidge, Notterman and E-GEOD-40966 data sets have high dimensionality. Therefore, experimentation starts with the selection of meaningful gene expressions from these data sets. In this connection, four different feature selection strategies have been adopted. Overall purpose of each feature selection strategy is to select discriminative gene expressions, but the underlying method of selection is entirely different. This is why gene expressions nominated by different feature selection strategies vary in number and are different as well. Weka, a machine learning tool, has been used to find F-score and chi-square based features for the data sets. For each data set, Weka returns an optimal set of gene expressions, which in turn leads to maximum classification accuracy for the data set. The process of features selection through PCA and mRMR is slightly different. Therefore, the process has been explained in the following text for KentRidge data set as an example.

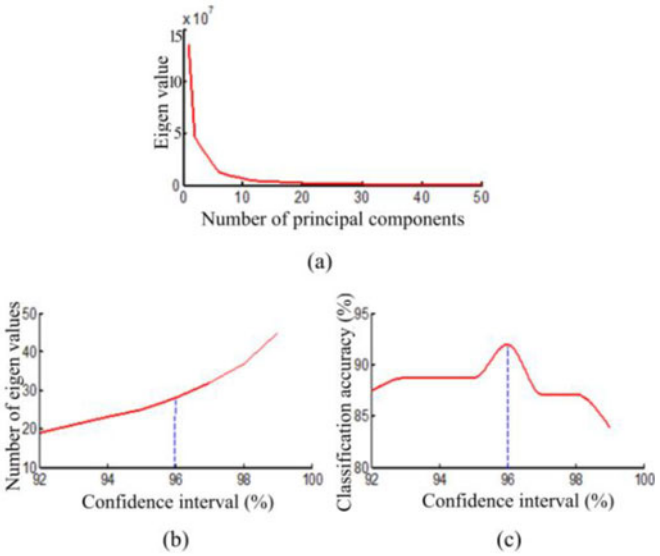


Fig. 4. Plot of (a) eigenvalues, (b) number of eigenvalues required for a particular confidence interval, (c) confidence interval versus classification accuracy of GECC on KentRidge data set.

For features selection through PCA, eigenvalues of the data set are analyzed, and principal components regarding most discerning eigenvalues are selected. Fig. 4a presents a plot of first 50 eigenvalues for KentRidge data set. It is observed from the figure that only first few eigenvalues have discerning power, therefore, we can cut off the remaining eigenvalues to reduce computational burden. In order to select the number of eigenvalues to be used, eigenvalues and corresponding classification accuracy of GECC have been measured as a function of confidence interval. Confidence interval is actually the area under the curve of eigenvalues as shown in Fig. 4a, and is usually normalized to 100 percent. In this research work, in order to determine the optimal number of eigenvalues (which yield maximum classification accuracy), eigenvalues at various confidence levels have been used. In this context, confidence interval has been varied in the range of 92-99 percent, and corresponding number of eigenvalues, which lie within the given confidence interval, have been determined from Fig. 4a. The principal components corresponding to these eigenvalues have been used for the classification. Figs. 4b and 4c demonstrate the number of eigenvalues (determined form Fig. 4a) and classification accuracy corresponding to various values of confidence intervals. It is observed from Fig. 4c that classification accuracy increases up to 96 percent confidence interval, but, it deteriorates beyond this point. Therefore, in this work, 28 principal components lying within 96 percent confidence interval have been used for classification.

Likewise, mRMR selects an ordered list of genes in terms of their discriminating power. Therefore, to find an optimal set of gene expressions, we have selected multiple sets (varying in size) of genes, and analyzed their effect on the classification accuracy achieved by the decision models. Fig. 5 reveals corresponding results, which show that accuracy gradually increases up to the subset of data comprising 50 gene expressions, and after that accuracy either decreases or maintains the same value. Therefore, we have used gene subset comprising 50 genes for classification of KentRidge data set.

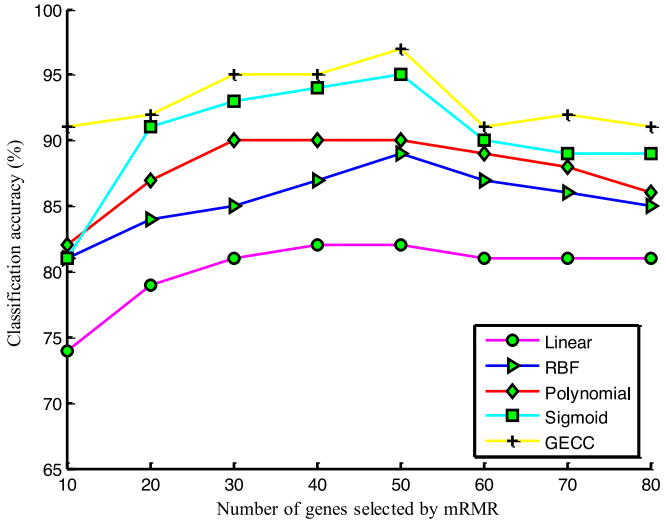


Fig. 5. Classification accuracy of various decision models for gene data sets selected by mRMR.

A similar process of feature selection through PCA and mRMR has also been applied on Notterman and E-GEOD-40966 data sets. Table 3 summarizes the number of genes selected by various feature selection strategies for different data sets. Principal components are shown in case of PCA in Table 3.

## 5.2 Parameter Selection for Various Decision Models Pertaining to Gene Selection Strategies

Performance of decision models depends on several parameters. Therefore, it is desirable to find optimal values of these parameters prior to classification. Analysis of optimal values of parameters has been divided into two sections; one, dealing with SVM models, and the second, dealing with rest of the models.

### 5.2.1 *Optimal Parameter Values for SVM Kernels*

Individual SVM models are trained on optimal parameters, however, there is no standard way to find optimal values of these parameters. In this research study, grid search method [42] has been employed by carefully setting grid range and step size. Polynomial kernel involves four parameters; $r$, $\gamma$, $c$ and $g$. In order to simplify the problem, $g$ and $r$ have been set to 3 and 1, respectively. Similarly, $r = 1$ has been fixed for sigmoid kernel. The optimal value of $c$ parameter has been obtained by adjusting the grid range of $c = [0, \ldots, 100]$ with $\Delta c = 1$ for all the kernels. Similarly, the optimal value of $\gamma$ has been found by setting the grid range of $\gamma = [0.001, \ldots, 0.1]$ with $\Delta\gamma = 0.002$ for RBF, polynomial, and sigmoid kernels. For each combination of $c$ and $\gamma$, 10-fold cross-validation has been applied on the data sets, and

TABLE 3
Number of Gene Expressions Selected by Various Feature Selection Strategies for Different Data Sets

| | KentRidge | Notterman | E-GEOD-40966 |
|---|---|---|---|
| PCA | 28 | 33 | 25 |
| mRMR | 50 | 120 | 140 |
| F-Score | 26 | 95 | 130 |
| Chi-square | 135 | 185 | 165 |

TABLE 4
Optimal Values of Parameters for Different SVM Decision Models

| | Chi-Sq/ original | | F-score | | mRMR | | PCA | |
|---|---|---|---|---|---|---|---|---|
| | $c$ | $\gamma$ | $c$ | $\gamma$ | $c$ | $\gamma$ | c | $\gamma$ |
| **BioGPS data set** | | | | | | | | |
| Linear | 36 | — | — | — | — | — | — | — |
| RBF | 16 | 0.009 | — | — | — | — | — | — |
| Sigm | 27 | 0.011 | — | — | — | — | — | — |
| Poly | 17 | 0.001 | — | — | — | — | — | — |
| **KentRidge data set** | | | | | | | | |
| Linear | 19 | — | 59 | — | 01 | — | 01 | — |
| RBF | 01 | 0.059 | 25 | 0.021 | 10 | 0.011 | 25 | 0.053 |
| Sigm | 26 | 0.057 | 17 | 0.071 | 38 | 0.033 | 18 | 0.065 |
| Poly | 04 | 0.011 | 13 | 0.075 | 68 | 0.013 | 17 | 0.021 |
| **Notterman data set** | | | | | | | | |
| Linear | 05 | — | 01 | — | 01 | — | 01 | — |
| RBF | 08 | 0.001 | 05 | 0.005 | 27 | 0.011 | 19 | 0.061 |
| Sigm | 32 | 0.011 | 28 | 0.001 | 24 | 0.021 | 06 | 0.055 |
| Poly | 11 | 0.001 | 13 | 0.021 | 20 | 0.037 | 40 | 0.029 |
| **E-GEOD-40966 data set** | | | | | | | | |
| Linear | 25 | — | 04 | — | 01 | — | 01 | — |
| RBF | 01 | 0.077 | 02 | 0.063 | 28 | 0.001 | 07 | 0.059 |
| Sigm | 03 | 0.001 | 12 | 0.041 | 32 | 0.003 | 18 | 0.043 |
| Poly | 13 | 0.001 | 34 | 0.085 | 01 | 0.021 | 06 | 0.001 |

classification accuracy has been calculated based on the predicted labels. The combination of parameter values where maximum classification accuracy is achieved has been selected to be optimal.

Table 4 summarizes the optimal values of the parameters for BioGPS data set, and different variants of other data sets corresponding to individual feature selection strategies. The results in Table 4 reveal that optimal values do not depend on the type of SVM model rather they depend on the nature of data.

Another parameter known as number of folds (F) is used for cross-validation strategy, which may also influence the performance of SVM models somehow. Therefore, we have measured the classification accuracy and time of SVM models over a potential range of F for all the data sets. Fig. 6 demonstrates the results for KentRidge and BioGPS data sets. Analysis of Figs. 6b and 6d reveals that the classification time exponentially increases as data are partitioned into more folds, and vice versa. The classification accuracy presented in Figs. 6a and 6c, on the other hand, seems independent of F. We observed almost similar behavior for Notterman and E-GEOD-40966 data sets. Therefore, a smaller value of 10 has been chosen for F owing to the exponential increase in time with increasing value of F.

#### 5.2.2 *Optimal Parameter Values for PNN and KNN Decision Models*

Performance of PNN kernel depends on spread of the Gaussian function. Optimal value of the spread has been found by varying it in a suitable range of 0.5-1, and by measuring corresponding classification accuracy and time. Fig. 7 presents the results. Fig. 7a demonstrates that smaller level of spread is suitable as far as the classification accuracy is concerned. Contrary, the level of spread has no influence

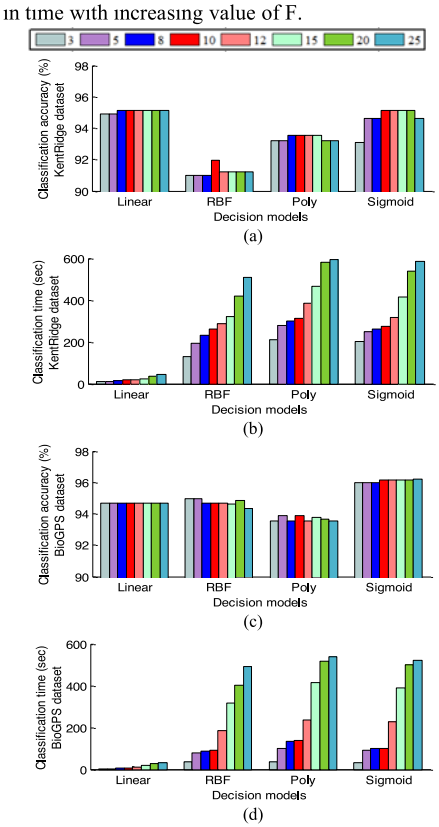


Fig. 6. F versus (a) classification accuracy and (b) time of KentRidge data, F versus (c) classification accuracy and (d) time of BioGPS data.

on the classification time. It is more or less the same for all the spread levels.

Likewise, performance of KNN classifier depends on number of neighbors (N) to be used in the classification process. Therefore, in this research study, effect of N on classification accuracy and time has been studied. The results are shown in Fig. 8. Fig. 8b demonstrates an increase in the classification time with the increase in N. Fig. 8a reflects that classification accuracy only increases upto $N = 3$, and beyond this point classification performance either deteriorates or remains the same.

### 5.3 Selected Gene Expressions, and Optimized SVM Classifiers Based Decision Modeling

Selected gene expressions are given as input to the optimized decision models for classification, and individual predictions of various models are collected. Individual SVM predictions are stacked, and then majority voting scheme is applied to find ensemble prediction. For instance, Fig. 9

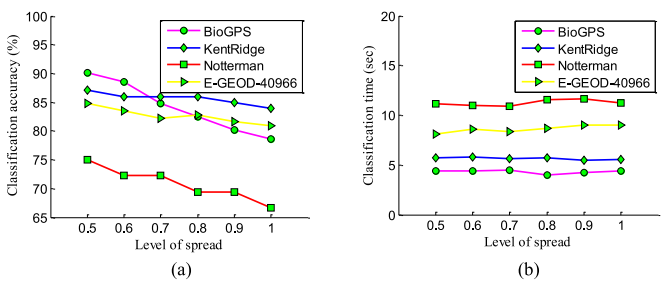


Fig. 7. Spread versus (a) classification accuracy and (b) classification time elapsed in 10-fold cross-validation through PNN classifier.

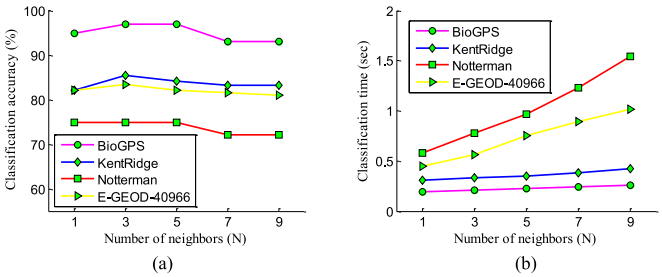


Fig. 8. Number of nearest neighbors (N) versus (a) classification time elapsed in 10-fold cross-validation through KNN classifier.

demonstrates predictions of individual SVM models and the GECC for BioGPS data set. In Fig. 9, data samples are along the horizontal axis, whereas stacked labels are along the vertical axis. First 37 data samples are normal, whereas remaining 94 samples are malignant. Each normal sample will be considered as wrongly classified if +1 label is assigned to it, and vice versa. For instance, fifth and seventh samples in Fig. 9 are actually normal, but have been misclassified by the linear classifier, and have been assigned +1 label.

There is a significant improvement in the performance of GECC compared to individual SVM models. It is evident from Fig. 9 that samples which were hard to classify for individual SVM models are better classified by GECC. This is mainly due to the ability of ensemble GECC to learn from the predictions of various decision models. Fig. 9 also demonstrates that the hard samples of BioGPS data set (fifth, 16th, 34th and 43rd sample in Fig. 9) are correctly classified by the proposed GECC technique due to the concept of weighted majority voting.

### 5.4 Performance on the Standard Data Sets

To measure the efficacy of the proposed GECC, various performance measures have been calculated. Table 5 shows the values of these measures for individual SVM decision

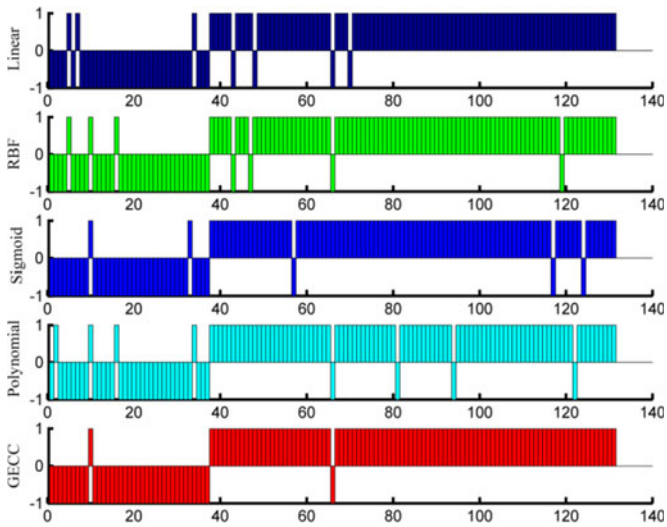


Fig. 9. Predictions of individual SVM models (collected through 10-fold cross-validation) and the proposed GECC for BioGPS data set.

models as well as for the proposed GECC model when used for the classification of different data sets.

The results demonstrate better performance of GECC compared to individual models. Ensemble classification accuracy for BioGPS data set is 98.67 percent, which is 2.49 percent higher compared to individual best accuracy of 96.18 percent, achieved by using sigmoid kernel. Likewise, the results prove clear superiority of GECC over individual SVM models regardless of the choice of gene selection strategy for KentRidge, Notterman and E-GEOD-40966 data sets. The best possible accuracy of GECC is 98.78, 97.22 and 97.71 percent, respectively, for F-score based selected sets of KentRidge, Notterman and E-GEOD-40966 data sets. These accuracy values, respectively, are 3.62, 2.78, and 3.71 percent higher compared to individual best achieved by using sigmoid kernel for these data sets. Moreover, there is also an improvement

TABLE 5
Performance Analysis for Various Combinations of Decision Models and Feature Selection Strategies

| | Linear | RBF | Sigm | Poly | GECC | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc | Acc | Acc | Acc | Acc | Sens | Spec | MCC | FM |
| **BioGPS data set** | | | | | | | | | |
| | 94.63 | 94.66 | ***96.18*** | 93.89 | **98.67** | 0.97 | 0.98 | 0.96 | 0.98 |
| **KentRidge data set** | | | | | | | | | |
| mRMR | 88.71 | 90.32 | 93.54 | 92.32 | 97.03 | 0.98 | 0.96 | 0.93 | 0.98 |
| F-Score | 90.32 | 91.94 | ***95.16*** | 93.54 | **98.78** | 0.98 | 0.97 | 0.96 | 0.98 |
| Chi-sq | 82.26 | 87.10 | 93.55 | 91.93 | 97.01 | 0.98 | 0.95 | 0.93 | 0.98 |
| PCA | 82.26 | 87.10 | 85.48 | 85.48 | 91.94 | 0.90 | 0.95 | 0.83 | 0.94 |
| **Notterman data set** | | | | | | | | | |
| mRMR | 86.11 | 86.11 | 91.67 | 88.89 | 94.44 | 0.94 | 0.94 | 0.89 | 0.94 |
| F-Score | 91.67 | 91.67 | ***94.44*** | ***94.44*** | **97.22** | 0.94 | 1.00 | 0.95 | 0.97 |
| Chi-sq | 80.56 | 83.33 | 88.89 | 83.33 | 91.67 | 0.89 | 0.94 | 0.83 | 0.91 |
| PCA | 77.78 | 80.56 | 86.11 | 83.33 | 88.89 | 0.94 | 0.83 | 0.78 | 0.89 |
| **E-GEOD-40966data set** | | | | | | | | | |
| mRMR | 90.57 | 91.43 | 93.14 | 92.86 | 97.14 | 0.97 | 0.98 | 0.94 | 0.98 |
| F-Score | 92.29 | 93.43 | ***94.00*** | 93.71 | **97.71** | 0.96 | 0.99 | 0.95 | 0.97 |
| Chi-sq | 89.71 | 90.29 | 92.00 | 90.86 | 95.71 | 0.95 | 0.97 | 0.91 | 0.96 |
| PCA | 86.57 | 87.71 | 90.29 | 88.86 | 94.29 | 0.93 | 0.96 | 0.88 | 0.95 |

*–Simple and italic bold face entries, respectively, correspond to the best performance of GECC and individual classifiers on a given data set.*

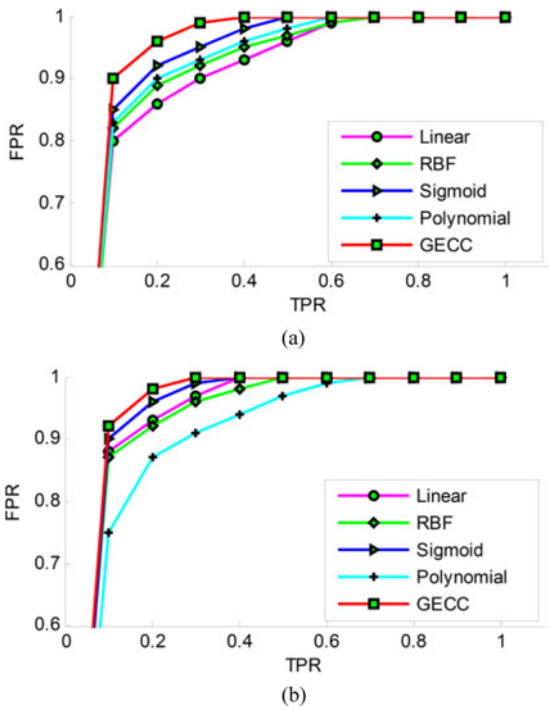


Fig. 10. ROC curves of various decision models for (a) KentRidge data set (using F-Score), and (b) BioGPS data set.

in terms of other performance measures for all the data sets. Further, results demonstrate that F-Score performs better because it achieves maximum possible separability between the instances of normal and malignant classes. The better performance of F-Score may be attributed to the characteristics of the underlying data set, which may have good separation between the values of gene expressions for the samples of two classes.

ROC curves [43] are an important parameter to measure the performance of classifiers. They are created by plotting true positive rate (TPR) against false positive rate (FPR). In order to plot ROC curve, the values of TPR and FPR of the entire test samples are obtained by applying threshold H in the range of [0-1]. In this study, the performance of the individual models and the proposed GECC on several colon data sets has also been measured in terms of ROC curves, and corresponding area under the curve (AUC). Fig. 10 presents ROC curves for various decision models when applied on BioGPS data set and F-Score based variant of KentRidge data set. It is evident from the figure that the proposed GECC has better ROC curve compared to other decision models. A similar behavior has been observed for Notterman and E-GEOD-40966 data sets, however, ROC curves have been shown only for two data sets keeping simplicity in mind.

AUC is usually measured from ROC curves. It is a scalar value that represents an overall performance of the decision model. A decision model is near to optimal if the value of AUC is close to one. Table 6 presents AUC of ROC curves for various decision models. AUC of GECC is higher compared to AUC of individual decision models for BioGPS and four variants of other data sets. The better AUC is largely due to the integrated decision power of various SVM models in one GECC.

TABLE 6
Performance Comparison of Decision Models in Terms of AUC

| | Linear | RBF | Sigmoid | Poly | GECC |
|---|---|---|---|---|---|
| **BioGPS data set** | | | | | |
| Original data | 0.9356 | 0.9323 | ***0.9516*** | 0.8763 | **0.9759** |
| **KentRidge data set** | | | | | |
| mRMR | 0.8765 | 0.8965 | 0.9385 | 0.9084 | 0.9797 |
| F-Score | 0.9052 | 0.9064 | ***0.9552*** | 0.9450 | **0.9886** |
| Chi-square | 0.8321 | 0.8865 | 0.9377 | 0.9036 | 0.9645 |
| PCA | 0.8232 | 0.8865 | 0.8612 | 0.8615 | 0.9254 |
| **Notterman data set** | | | | | |
| mRMR | 0.8693 | 0.8752 | 0.9210 | 0.8962 | 0.9452 |
| F-Score | 0.9199 | 0.9215 | 0.9387 | ***0.9398*** | **0.9799** |
| Chi-square | 0.8164 | 0.8497 | 0.8751 | 0.8439 | 0.9200 |
| PCA | 0.7991 | 0.8135 | 0.8586 | 0.8465 | 0.9029 |
| **E-GEOD-40966 data set** | | | | | |
| mRMR | 0.8987 | 0.9235 | 0.9463 | 0.9287 | 0.9705 |
| F-Score | 0.9165 | 0.9487 | ***0.9586*** | 0.9535 | **0.9832** |
| Chi-square | 0.8757 | 0.8865 | 0.9365 | 0.9174 | 0.9607 |
| PCA | 0.8598 | 0.8654 | 0.9174 | 0.8978 | 0.9386 |

*– Simple and italic bold face entries, respectively, correspond to the best performance of the proposed GECC and individual classifiers on a given data set.*

### 5.5 Computational Complexity of the GECC Technique

In this section, we have calculated CPU time elapsed from the start to the end of different phases of the proposed GECC technique such as gene selection, parameter optimization with 10-fold cross-validation, and weight optimization for majority vote has been measured in seconds. Table 7 shows the corresponding results.

TABLE 7
Computational Time Requirements of GECC (sec)

| | BioGPS | KentRidge | Notterman | E-GEOD-40966 |
|---|---|---|---|---|
| **Gene selection** | | | | |
| mRMR | 3.56 | 6.77 | 13.58 | 15.89 |
| Chi-square | 0.33 | 0.72 | 0.98 | 1.06 |
| F-Score | 0.68 | 0.99 | 1.58 | 3.01 |
| PCA | 0.048 | 7.76 | 16.02 | 32.25 |
| **Parameter optimization using 10-fold cross-validation** | | | | |
| Linear | 1.25 | 2.22 | 2.62 | 3.30 |
| RBF | 56.02 | 120.02 | 136.23 | 179.23 |
| Sigmoid | 63.54 | 124.25 | 140.28 | 184.28 |
| Poly | 65.87 | 129.98 | 142.56 | 188.56 |
| GECC | 186.68 | 376.47 | 421.69 | 555.37 |
| **Weight optimization for majority vote** | | | | |
| GECC | 0.3121 | 0.4080 | 0.3356 | 0.3897 |
| **Training on optimal parameter values** | | | | |
| Linear | 0.0100 | 0.0212 | 0.0252 | 0.0330 |
| RBF | 0.0116 | 0.0236 | 0.0272 | 0.0346 |
| Sigmoid | 0.0123 | 0.0252 | 0.0280 | 0.0360 |
| Poly | 0.0127 | 0.0260 | 0.0286 | 0.0374 |
| GECC | 0.0466 | 0.0960 | 0.1090 | 0.1410 |
| **Testing on optimal parameter values** | | | | |
| Linear | 0.0024 | 0.0031 | 0.0037 | 0.0040 |
| RBF | 0.0026 | 0.0033 | 0.0040 | 0.0042 |
| Sigmoid | 0.0027 | 0.0038 | 0.0041 | 0.0044 |
| Poly | 0.0030 | 0.0037 | 0.0043 | 0.0045 |
| GECC | 0.0117 | 0.0148 | 0.0182 | 0.0191 |

TABLE 8
Performance Comparison of GECC with Some Existing Schemes and Decision Models in Terms of Classification Accuracy

| | Ref. | Acc | Sens | Spec | MCC | FM |
|---|---|---|---|---|---|---|
| **BioGPS data set** | | | | | | |
| KNN | — | 91.15 | 0.89 | 0.88 | 0.90 | 0.90 |
| PNN | — | 90.08 | 0.65 | 0.87 | 0.75 | 0.79 |
| Decision tree | — | 90.84 | 0.92 | 0.93 | 0.82 | 0.87 |
| Li et al. | [17] | 91.23 | 0.89 | 0.90 | 0.88 | 0.88 |
| Venkatesh et al. | [20] | 91.25 | 0.89 | 0.93 | 0.90 | 0.89 |
| Kulkarni et al. | [21] | ***94.45*** | 0.96 | 0.96 | 0.97 | 0.97 |
| Lee et al. | [22] | 80.23 | 0.79 | 0.81 | 0.77 | 0.81 |
| Tong et al. | [23] | 93.55 | 0.86 | 0.98 | 0.86 | 0.90 |
| GECC | — | **98.67** | 0.97 | 0.98 | 0.96 | 0.98 |
| **KentRidge data set** | | | | | | |
| KNN | — | 85.48 | 0.77 | 0.90 | 0.68 | 0.79 |
| PNN | — | 87.09 | 0.77 | 0.93 | 0.71 | 0.81 |
| Decision tree | — | 85.48 | 0.73 | 0.95 | 0.71 | 0.80 |
| Li et al. | [17] | 89.01 | 0.88 | 0.90 | 0.87 | 0.88 |
| Venkateshet al. | [20] | 94.40 | 0.92 | 0.93 | 0.93 | 0.92 |
| Kulkarni et al. | [21] | ***98.33*** | 0.99 | 0.98 | 0.97 | 0.97 |
| Lee et al. | [22] | 76.85 | 0.77 | 0.79 | 0.70 | 0.74 |
| Tong et al. | [23] | 90.32 | 0.82 | 0.95 | 0.79 | 0.86 |
| GECC | — | **98.78** | 0.98 | 0.97 | 0.96 | 0.98 |
| **Notterman data set** | | | | | | |
| KNN | — | 75.00 | 0.72 | 0.78 | 0.50 | 0.74 |
| PNN | — | 75.00 | 0.67 | 0.83 | 0.51 | 0.73 |
| Decision tree | — | 77.78 | 0.67 | 0.89 | 0.57 | 0.75 |
| Li et al. | [17] | 80.56 | 0.78 | 0.83 | 0.61 | 0.80 |
| Venkaetsh et al. | [20] | 86.11 | 0.83 | 0.89 | 0.72 | 0.86 |
| Kulkarni et al. | [21] | ***91.67*** | 0.89 | 0.94 | 0.83 | 0.91 |
| Lee et al. | [22] | 83.33 | 0.83 | 0.83 | 0.67 | 0.83 |
| Tong et al. | [23] | 88.89 | 0.89 | 0.89 | 0.78 | 0.89 |
| GECC | — | **97.22** | 0.94 | 1.00 | 0.95 | 0.97 |
| **E-GEOD-40966 data set** | | | | | | |
| KNN | — | 83.71 | 0.89 | 0.76 | 0.66 | 0.87 |
| PNN | — | 84.57 | 0.86 | 0.83 | 0.68 | 0.87 |
| Decision tree | — | 86.29 | 0.88 | 0.85 | 0.72 | 0.88 |
| Li et al. | [17] | 88.57 | 0.89 | 0.88 | 0.77 | 0.90 |
| Venkaetsh et al. | [20] | 90.57 | 0.91 | 0.90 | 0.81 | 0.92 |
| Kulkarni et al. | [21] | ***92.29*** | 0.94 | 0.90 | 0.84 | 0.94 |
| Lee et al. | [22] | 88.86 | 0.89 | 0.89 | 0.77 | 0.90 |
| Tong et al. | [23] | 91.14 | 0.93 | 0.88 | 0.82 | 0.93 |
| GECC | — | **97.71** | 0.96 | 0.99 | 0.95 | 0.97 |

*–Simple and italic bold face entries, respectively, correspond to the best performance of the proposed GECC and previous techniques on a given data set.*

The parameter optimization time of GECC, shown in the second section of the table, is merely the sum of the optimization times of individual models. Once optimal values of parameters and weights for the majority votes have been computed, they are directly used for the training and testing of samples. The training and testing time on optimal values is shown in the last two sections of Table 7. The testing time involves an overhead of weighted majority voting in addition to the time consumed by individual models. The results in Table 7 show that the proposed GECC technique is computationally tractable even the maximum training time for the largest data set (E-GEOD-40966) is $35.25 + 555.37 + 0.39 = 591.01$ sec only when PCA has been used as the underlying gene selection strategy.

TABLE 9
Performance Comparison of GECC with Some Existing Schemes in Terms of CPU Time Requirements (sec)

| | Ref. | Training time | Testing time |
|---|---|---|---|
| **BioGPS data set** | | | |
| Li et al. | [17] | 0.3055 | ***0.0105*** |
| Venkatesh et al. | [20] | ***0.0401*** | 0.0224 |
| Kulkarni et al. | [21] | 0.3564 | 0.0302 |
| Lee et al. | [22] | 0.2552 | 0.0125 |
| Tong et al. | [23] | 0.3870 | 0.0826 |
| GECC | — | **0.0466** | **0.0117** |
| **KentRidge data set** | | | |
| Li et al. | [17] | 0.5252 | ***0.0123*** |
| Venkatesh et al. | [20] | ***0.1150*** | 0.0225 |
| Kulkarni et al. | [21] | 0.5650 | 0.0312 |
| Lee et al. | [22] | 0.4230 | 0.0156 |
| Tong et al. | [23] | 0.5692 | 0.0869 |
| GECC | — | **0.0960** | **0.0148** |
| **Notterman data set** | | | |
| Li et al. | [17] | 0.5826 | ***0.0130*** |
| Venkatesh et al. | [20] | ***0.1057*** | 0.0228 |
| Kulkarni et al. | [21] | 0.6239 | 0.0316 |
| Lee et al. | [22] | 0.4805 | 0.0132 |
| Tong et al. | [23] | 0.5974 | 0.0898 |
| GECC | — | **0.1090** | **0.0182** |
| **E-GEOD-40966 data set** | | | |
| Li et al. | [17] | 0.5902 | 0.0159 |
| Venkatesh et al. | [20] | ***0.2135*** | 0.0227 |
| Kulkarni et al. | [21] | 0.6903 | 0.0341 |
| Lee et al. | [22] | 0.5237 | ***0.0128*** |
| Tong et al. | [23] | 0.6704 | 0.0967 |
| GECC | — | **0.1410** | **0.0191** |

*–Simple and italic bold face entries, respectively, correspond to the time of GECC and the minimum time of previous techniques.*

### 5.6 Performance Comparison of GECC with Some Existing Schemes and Classifiers

The performance of GECC has been compared with several existing gene based colon cancer detection techniques in terms of classification accuracy and CPU time requirements. In this context, five techniques [17], [20], [21], [22], [23] have been selected. These techniques have been implemented in Matlab, and have been tested on the same system. Further, the effectiveness of GECC has also been compared with a few commonly used decision models, like, KNN, PNN and decision trees. Table 8 presents a comparison of the classification accuracy of the GECC with previously published techniques and classifiers.

GECC exhibits performance improvement in terms of all the performance measures. Classification accuracy of GECC on BioGPS, KentRidge, Notterman and E-GEOD-40966 data set, respectively, is 4.22, 0.45, 5.55 and 5.42 percent higher compared to the best accuracy achieved by previous techniques. Further, GECC produces superior classification results compared to KNN, PNN and decision tree. Such a noteworthy performance improvement validates the efficacy of the proposed GECC technique for detection of colon cancer, and also encourages its use for classification of other complex gene based data sets.

Moreover, the performance of the GECC technique has been compared with previously published techniques in terms of the CPU time requirements. In this

TABLE 10
Binary-Class Gene Expression Data Sets

| Data set | Ref. | No. of Genes | Samples of $C_1$ | Samples of $C_2$ | Total samples |
|---|---|---|---|---|---|
| CNS | [44] | 7,129 | 39 | 21 | 60 |
| DLBCL | [45] | 7,129 | 58 | 19 | 77 |
| Leukemia | [46] | 7,129 | 25 | 47 | 72 |
| Lung-I | [47] | 12,533 | 150 | 31 | 181 |
| Lung-II | [48] | 7,129 | 86 | 10 | 96 |
| Prostate-I | [49] | 12,600 | 52 | 50 | 102 |
| Prostate-II | [50] | 12,625 | 38 | 50 | 88 |
| Prostate-III | [51] | 12,626 | 24 | 09 | 33 |

context, the training and testing time of these techniques has been measured on optimal parameter values. Table 9 shows the results which reveal that the proposed GECC technique is computationally tractable, and consumes quite reasonable CPU time. The training time of GECC is smaller compared to other techniques except the Venkatesh et al. [20], which shows comparable results.

### 5.7 Performance Analysis of GECC on Other Complex Gene Expression Data Sets

In this section, we have validated the effectiveness of the proposed GECC on other complex binary-class gene expression data sets. This has been done to demonstrate the applicability of GECC as a general purpose gene based cancer detection technique. Table 10 provides a brief description of these data sets.

These data sets have enormously large dimensionality. Therefore, feature selection strategies (PCA, F-Score, chi-square and mRMR) have been applied to select meaningful features. Table 11 summarizes the number of features selected by various feature selection strategies. In case of PCA, confidence interval has been provided in Table 11, and principal components within the given confidence interval have been used for classification.

Linear, RBF, sigmoid, polynomial and ensemble GECC have been employed for the classification of data sets. It is noteworthy that features highlighted in Table 11 have been used for the classification, and corresponding classification accuracies have been reported in Table 12. It can reasonably be concluded that GECC can effectively detect several cancer types.

TABLE 11
Number of Features Selected by Feature Selection Strategies for Some Binary-Class Gene Expression Data Sets

| | Original | mRMR | PCA | Chi-square | F-score |
|---|---|---|---|---|---|
| CNS | 7,129 | 175 | 96 | 180 | **165** |
| DLBCL | 7,129 | 160 | 96 | 210 | **155** |
| Leukemia | 7,129 | 180 | 97 | 220 | **135** |
| Lung-I | 12,533 | **280** | 97 | 295 | 235 |
| Lung-II | 7,129 | 130 | 96 | 195 | **140** |
| Prostate-I | 12,600 | 235 | 96 | 410 | **220** |
| Prostate-II | 12,625 | 285 | 97 | 395 | **235** |
| Prostate-III | 12,626 | 285 | 96 | 425 | **280** |

*–Bold face entries correspond to the features where maximum accuracy is achieved by GECC for a given data set.*

TABLE 12
Performance of GECC on Some Binary-Class Gene Expression Data Sets

| | Linear | RBF | Sigmoid | Polynomial | GECC |
|---|---|---|---|---|---|
| CNS | 91.67 | 93.33 | 95.00 | **96.67** | 98.33 |
| DLBCL | 89.61 | 92.21 | **94.81** | 93.51 | 98.70 |
| Leukemia | 91.67 | 93.06 | **97.22** | 95.83 | 98.61 |
| Lung-I | 92.27 | 93.92 | **97.79** | 96.13 | 99.45 |
| Lung-II | 81.25 | 82.29 | **85.42** | 83.33 | 88.54 |
| Prostate-I | 89.22 | 91.98 | **94.12** | 93.14 | 96.08 |
| Prostate-II | 88.64 | 89.77 | **92.05** | 90.91 | 94.32 |
| Prostate-III | 90.91 | 93.94 | **96.97** | **96.97** | 100.00 |

*–Bold face entries correspond to the individual best performance of a classifier for a given data set.*

## 6 Conclusion

The primary focus of our work was the development of a robust and accurate classification technique, called GECC, for gene expression based prediction of colon cancer. The proposed GECC technique employs an ensemble of various SVM decision models for classification. The experiments have been conducted on four standard colon cancer data sets. To reduce the large size of the data sets, four different feature selection strategies have been employed. Analysis reveals that genes selected by F-Score are better able to classify different data sets compared to the genes selected by other techniques. Amongst the multiple individual decision models, sigmoid SVM performs best for all the data sets. Ensemble SVM greatly increases performance compared to individual SVM decision models with a slight increase in computational time. Classification accuracy of GECC for BioGPS, KentRidge, Notterman and E-GEOD-40966 data set is 98.67, 98.78, 97.22 and 97.71 percent, respectively. These values are better compared to the individual best accuracies of 96.18, 95.16, 94.44 and 94.00 percent achieved by sigmoid SVM, respectively, for BioGPS, KentRidge, Notterman and E-GEOD-40966 data set. Performance of GECC has also been validated on several other complex gene expression data sets, and quite promising classification results have been achieved. Therefore, we can reasonably conclude that the proposed GECC can help biologists not only in accurately predicting the cancer of colon, but of other body parts as well. There are two possible future directions along this study. First possibility is to unite multiple feature selection methods i.e. selecting genes using one technique and then applying another on the selected genes. Second option is to assign weight to each gene by gene selection techniques and then apply maximum voting for selecting discerning genes.

## References

[1] (2012). Colon cancer risk factors. [Online]. Available: http://ccalliance.org/colorectal_cancer/riskfactors.html
[2] A. Andrion, C. Magnani, P. G. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M. Botta, and B. Terracini, "Malignant mesothelioma of the pleura: Inter observer variability," *J. Clin. Pathol.*, vol. 48, pp. 856–860, 1995.
[3] S. Rathore, M. Hussain, A. Ali, and A. Khan, "A recent survey on colon cancer detection techniques," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 10, no. 3, pp. 545–563, Jul. 2013.
[4] K. Masood and N. Rajpoot, "Texture based classification of hyperspectral colon biopsy samples using clbp," in *Proc. Int. Symp. Biomed. Imaging: From Nano Macro.*, 2009, pp. 1011–1014.

[5] K. Rajpoot and N. Rajpoot, "SVM optimization for hyperspectral colon tissue cell classification," in *Proc. 7th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2004, pp. 829–837.

[6] A. Molckovsky, L. M. Song, M. G. Shim, N. E. Marcon, and B. C. Wilson, "Diagnostic potential of near-infrared raman spectroscopy in the colon: Differentiating adenomatous from hyperplastic polyps," *Gastrointest. Endosc.*, vol. 57, pp. 396–402, 2003.

[7] Z. Huang, "Laser-induced auto fluorescence microscopy of normal and tumor human colonic tissue," *Int. J. Oncol.*, vol. 24, pp. 59–63, 2004.

[8] K. Masood, N. Rajpoot, H. Qureshi, and K. Rajpoot, "Co-occurrence and morphological analysis for colon tissue biopsy classification," in *Proc. 4th Int. Workshop Front. Info. Technol.*, 2006, pp. 211–216.

[9] S. Rathore, M. Hussain, M. A. Iftikhar, and A. Jalil, "Ensemble classification of colon biopsy images based on information rich hybrid features," *Comput. Biol. Med.*, vol. 47, pp. 76–92, 2014.

[10] A. N. Esgiar, R. N. G. Naguib, B. S. Sahrif, and M. K. Bennett, "Fractal analysis in the detection of colonic cancer images," *IEEE Trans. Info. Technol. Biomed.*, vol. 6, no. 1, pp. 54–58, Mar. 2002.

[11] A. G. Todman, R. N. G. Naguib, and M. K. Bennett, "Orientational coherence metrics: Classification of colon images based on human form perception," in *Proc. IEEE Can. Conf. Elec. Comput. Eng.*, 2001, pp. 1379–1385.

[12] M. Y. Wu, D. Q. Dai, Y. Shi, H. Yan, and X. F. Zhang, "Biomarker identification and cancer classification based on microarray data using Laplace Naive Bayes model with mean shrinkage," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 6, pp. 1649–1662, Nov./Dec. 2012.

[13] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. National Acad. Sci. USA*, vol. 96, pp. 6745–6750, Nov./Dec. 1999.

[14] M. Grade, P. Hörmann, S. Becker, A. B. Hummon, D. Wangsa, S. Varma, R. Simon, T. Liersch, H. Becker, M. J. Difilippantonio, B. M. Ghadimi, and T. Ried, "Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas," *Cancer Res.*, vol. 67, pp. 41–56, 2007.

[15] K. Kim, U. Park, J. Wang, J. Lee, S. Park, S. Kim, D. Choi, C. Kim, and J. Park, "Gene profiling of colonic serrated adenomas by using oligonucleotide microarray," *Int. J. Colorectal Dis.*, vol. 23, pp. 569–580, 2008.

[16] S. Backert, M. Gelos, U. Kobalz, M. L. Hanski, C. Bohm, B. Mann, N. Lovin, A. Gratchev, U. Mansmann, M. P. Moyer, E. O. Riecken, and C. Hanski, "Differential gene expression in colon carcinoma cells and tissues detected with a cDNA Array," *Int. J. Cancer*, vol. 82, pp. 868–874, 1999.

[17] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "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.

[18] Z. Chen and J. Li, "A multiple kernel support vector machine scheme for simultaneous feature selection and rule-based classification," in *Proc. 11th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, 2007, pp. 441–448.

[19] H. S. Shon, G. Sohn, K. S. Jung, S. Y. Kim, E. J. Cha, and K. H. Ryu, "Gene expression data classification using discrete wavelet transform," in *Proc. Int. Conf. Bioinformat. Comput. Biol.*, 2009, pp. 204–208.

[20] E. T. Venkatesh, P. Thangaraj, and S. Chitra, "An improved neural approach for malignant and normal colon tissue classification from oligonucleotide arrays," *Euro. J. Sci. Res.*, vol. 54, pp. 159–164, 2011.

[21] A. Kulkarni, N. Kumar, V. Ravi, and U. S. Murthy, "Colon cancer prediction with genetics profiles using evolutionary techniques," *Expert Syst. Appl.*, vol. 38, pp. 2752–2757, 2011.

[22] K. Lee, Z. Man, D. Wang, and Z. Cao, "Classification of bioinformatics dataset using finite impulse response extreme learning machine for cancer diagnosis," *Neural Comput. Appl.*, vol. 22, pp. 457–468, 2013.

[23] M. Tong, K. H. Liu, C. Xu, and W. Ju, "An ensemble of SVM classifiers based on gene pairs," *Comput. Biol. Med.*, vol. 43, pp. 729–737, 2013.

[24] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[25] (2013). Colon cancer data set Kent ridge. [Online]. Available: http://datam.i2r.a-star.edu.sg/datasets/krbd/ColonTumor/ColonTumor.html

[26] (2013). Colon cancer data set Biogps. [Online]. Available: http://biogps.org/dataset/1352/stage-ii-and-stage-iii-colorectal-cancer/

[27] D. A. Notterman, U. Alon, A. J. Sierk, and A. J. Levine, "Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays," *Cancer Res.*, vol. 61, no. 7, pp. 3124–3130, 2001.

[28] L. Marisa, A. D. Reyniès, A. Duval, J. Selves, M. P. Gaub, L. Vescovo, M. C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J. F. Fléjou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. L. P. Mail, and V. Boige, "Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value," *PLoS Med.*, vol. 10, no. 5, 2013.

[29] M. Re, M. Mesiti, and G. Valentini, "A fast ranking algorithm for predicting gene functions in biomolecular networks," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 6, pp. 1812–1818, Nov./Dec. 2012.

[30] J. C. Rajapakse and P. A. Mundra, "Multiclass gene selection using pareto-fronts," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 10, no. 1, pp. 87–97, Jan./Feb. 2013.

[31] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[32] K. Polat and S. Günes, "A new feature selection method on classification of medical data sets: Kernel f-score feature selection," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10367–10373, 2009.

[33] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosoph. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.

[34] M. Hassan, A. Chaudhry, A. Khan, and J. Y. Kim, "Carotid artery image segmentation using modified spatial fuzzy c-means and ensemble clustering," *Comput. Methods Programs Biomed.*, vol. 108, pp. 1261–1276, 2012.

[35] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley-Interscience, 1998.

[36] W. B. Sampaioa, E. M. Diniza, A. C. Silvaa, A. C. Paivaa, and M. Gattassb, "Detection of masses in mammogram images using CNN, geostatistic functions and SVM," *Comput. Biol. Med.*, vol. 41, no. 8, pp. 653–664, 2011.

[37] A. Subasi, "Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders," *Comput. Biol. Med.*, vol. 43, no. 5, pp. 576–586, 2013.

[38] S. Theodoridis and K. Koutroumbas, "Pattern recognition," *Fourth ed, Academic Press*, 2008.

[39] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2001.

[40] S. Osowski, K. Siwek, and R. Siroic, "Neural system for heartbeats recognition using genetically integrated ensemble of classifiers," *Comput. Biol. Med.*, vol. 41, no. 3, pp. 173–180, 2011.

[41] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Info. Comput.*, vol. 108, pp. 212–261, 1994.

[42] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector machines," Dept. Comput. Sci. Inf. Eng., National Taiwan Univ., Taiwan, 2003.

[43] S. Rathore, M. A. Iftikhar, M. Hussain, and A. Jalil, "A novel approach for ensemble clustering of colon biopsy images," in *Proc. 11th Int. Conf. Front. Inf. Technol.*, 2013, pp. 25–30.

[44] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Pogio, S. Mukharjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[45] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gassenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Merisov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Gloub, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nat. Med.*, vol. 8, no. 1, pp. 68–74, 2002.

[46] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Merisov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[47] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res.*, vol. 62, pp. 4963–4967, 2002.

[48] D. G. Beer, S. L. R. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Lannettoni, M. B. Orringer, and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nat. Med.*, vol. 8, no. 8, pp. 816–823, 2002.

[49] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amino, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp. 203–209, 2002.

[50] R. O. Stuart, W. Wachsman, C. C. Berry, J. Wang-Rodriguez, L. Wasserman, I. Klacansky, D. Masys, K. Arden, S. Goodison, M. McClelland, Y. Wang, A. Sawyers, I. Kalcheva, D. Tarin, and D. Mercola, "In silico dissection of cell-type-associated patterns of gene expression in prostate cancer," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 615–620, 2004.

[51] J. Welsh, L. Sapinoso, A. Su, S. Kern, J. Wang-Rodriguez, C. Moskaluk, H. Frierson, and G. Hampton, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Cancer Res.*, vol. 61, no. 16, pp. 5974–5978, 2001.

**Saima Rathore** received the BS degree in software engineering from Fatima Jinnah Women University, Rawalpindi, Pakistan, in 2006, and the MS degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2008. She is currently working toward the PhD degree at the Pakistan Institute of Engineering and Applied Sciences, Islamabad. She possesses more than eight years of research and teaching experience at the university level. Her research interests include medical image analysis, segmentation, and classification.

**Mutawarra Hussain** received the MSc degree in computer science and the MS degree in nuclear engineering from Quaid-i-Azam University, Islamabad, Pakistan, in 1979 and 1981, respectively, and the PhD degree in the field of medical image analysis from the School of Computer Science, University of Birmingham, United Kingdom, in 2000. He worked as a postdoctoral researcher in the field of medical image analysis at the School of Computer Science, University of Birmingham, in 2008. He has more than 30 years of research experience and is currently working as the dean of research at the Pakistan Institute of Engineering and Applied Sciences, Islamabad. His research interests include medical image analysis, image segmentation, and classification.

**Asifullah Khan** received the MS and PhD degrees in computer systems engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan, in 2003 and 2006, respectively. He has completed two years of postdoctoral research at the Signal and Image Processing Lab, Department of Mechatronics, Gwangju Institute of Science and Technology, Korea. He has more than 15 years of research experience and is working as an associate professor in the Department of Computer and Information Sciences at Pakistan Institute of Engineering and Applied Sciences, Nilore, Pakistan. His research interests include digital watermarking, pattern recognition, image processing, and evolutionary algorithms.