Egyptian Informatics Journal 18 (2017) 151–159

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

Gene expression based cancer classification

Sara Tarek *, Reda Abd Elwahab, Mahmoud Shoman

Faculty of Computers and Information, Department of Information Technology, Cairo University, Egypt

ARTICLE INFO

Article history:
Received 4 September 2016
Revised 23 October 2016
Accepted 6 December 2016
Available online 20 December 2016

Keywords:
Microarrays
Cancer
Classification
Gene expression
Feature selection
Ensemble
K-NN
Bioinformatics
Computer science
Machine learning

ABSTRACT

Cancer classification based on molecular level investigation has gained the interest of researches as it provides a systematic, accurate and objective diagnosis for different cancer types. Several recent researches have been studying the problem of cancer classification using data mining methods, machine learning algorithms and statistical methods to reach an efficient analysis for gene expression profiles.

Studying the characteristics of thousands of genes simultaneously offered a deep insight into cancer classification problem. It introduced an abundant amount of data ready to be explored. It has also been applied in a wide range of applications such as drug discovery, cancer prediction and diagnosis which is a very important issue for cancer treatment. Besides, it helps in understanding the function of genes and the interaction between genes in normal and abnormal conditions. That is done by monitoring the behavior of genes -gene expression data- under different conditions.

In this paper, an effective ensemble approach is proposed. Ensemble classifiers increase not only the performance of the classification, but also the confidence of the results. The motivations beyond using ensemble classifiers are that the results are less dependent on peculiarities of a single training set and because the ensemble system outperforms the performance of the best base classifier in the ensemble.

1. Introduction

Deoxyribonucleic acid or DNA stores genetic information required by all livings in order to build, function and develop. DNA is said to be the blueprint of all living organisms since its components encode all the information needed for maintaining life. This genetic information are preserved and passed from one cell to another during the process of cell division in which a parent cell splits into two new daughter cells. DNA molecules form a double twisted helix bonded together and arranged in a very precise order. The basic four molecular units forming the DNA helix are then sequenced in a particular arrangement such that each component on one strand can only bond with a certain component in the other strand. DNA replicates by breaking the bond between its two strands -the double twisted helix- and each strand form a matching strand, re-bond and re-twist once again.

The genome -the entire DNA sequence- provides a template for the synthesis of a variety of RNA molecules. The main types of RNA are messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). One of the major functions of the DNA is to construct proteins which are responsible for carrying out most of cell functions. The process of constructing proteins consists of two major steps; namely, transcription stage in which DNA molecule is transcribed into messenger RNA or mRNA (which is a type of ribonucleic acid RNA); and translation stage, where mRNA is translated into proteins' amino acid sequences to perform cell functions. Once protein is constructed, the gene is said to be *expressed*. The standard technique for measuring gene expression is to measure the mRNA instead of proteins. The reason behind using mRNA sequences is that they hybridize with their complementary RNA or DNA sequences while this property lacks in proteins.

Gene expression level represents the amount of RNA produced in a cell under different biological states. So during cell division process, if the cells suffer from diseases -i.e. cancer or malignant tumors- that cause alteration or mutations in genes, the uncontrollable behavior of the gene will be transmitted to daughter cells.

* Corresponding author.
E-mail addresses: sarah.tarek1@gmail.com (S. Tarek), r.abdelwahab@fci-cu.edu.eg (R. Abd Elwahab), m.essmael@fci-cu.edu.eg (M. Shoman).
Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2016.12.001
1110-8665/

Moreover, certain gene expression values will be affected and hence expression levels can be realized by monitoring the RNA.

The expression levels of thousands of genes can be simultaneously measured under particular experimental environments and conditions due to the significant advancement of DNA microarray technology. This technology made it possible to understand life on the molecular level, and enables to generate large-scale gene expression data. Besides, it has led to many analytical insights because it produced large amount of gene data ready to be analyzed rapidly and precisely by managing them using several statistical and machine learning processes.

In order to transform the DNA microarray samples from their analog form, which is DNA sequences printed in a high density array on a glass microscopic slide, into a digital form, which is the gene expression data matrix that can be handled and manipulated, several steps should be done.

The standard technique is to transcribe mRNA from two cells and reverse transcribe both of them into cRNA, label them using fluorescent dyes (red if cancerous and green if benign). Both samples are distributed over the whole microarray in order to hybridize with their corresponding cDNA (labeled cDNA try to bind to their complementary cDNA on the microarray in order to form a double stranded molecule in the process called hybridization). Hybridization thus acts like a detector of the presence of a certain gene. The slide is then scanned to obtained numerical intensities of each dye. Then, the substrate can be scanned as an image that can be manipulated by image processing techniques; the intensity of the colors corresponds to the number of mRNA transcribed for each gene. By comparing the intensity of the colors for a gene under two different experimental conditions, gene expression levels can be monitored. For all genes on a single chip, gene expression value is: $\log_2(I_R/I_G)$ where I_R is the intensity of the red dye and I_G is the intensity of the green dye.

DNA microarray technology provided a considerable amount of important data ready to be explored. This massive amount of data suffers from several issues. First, the process of extracting samples under the right conditions is extremely difficult to meet and it involves high level of noise. Second, there are thousands of gene expressions versus few dozens of samples, which require excluding the irrelevant genes before actual classification. The choice of the discriminating genes between different cancer types or classes is a major research field. Third, several tradeoffs have been revealed such as maintaining accuracy rate versus ensuring generalization, controlling complexity versus improving classifier's performance, improving the performance versus memory requirements. Those factors have affected the efficiency of the cancer classification algorithms.

2. Related work

Okun [1] suggested an ensemble system implemented over Colon dataset. Filter feature selection models are used to mitigate the effect of overfitting. Three Different gene selection methods were implemented; namely, Backward Elimination Hilbert-Schmidt Independence Criterion "BAHSIC" [2], Extreme Value Distribution based gene selection "EVD" [3] and Singular Value Decomposition Entropy gene selection "SVDEntropy" [4]. The ensemble consists of five base classifiers, each utilizes K-nearest neighbor "K-NN" with different values of K "either 3 or 5 nearest neighbors". The choice of K-NN classifier was justified as it doesn't require training which is more suitable for usage with the Colon dataset because of the nature of the microarray data (few dozens of samples with high dimensionality).

Because of the small sample size, bolstered resubstitution error estimation "BRE" is used [5]. The bolstered resubstitution estimator is based on the theory that, more confidence should be attributed to points far from the decision boundary than points near it. Bolstered resubstitution error estimators are low variance and generally low bias as well. It is very competitive in comparison with cross validation and bootstrap error estimators especially for small sample problems [6].

Although the results of the as-is system looks appealing; however, it can be further improved by taking some points into consideration. First, due to the sensitivity of the cancer classification domain, more accurate output results are desired such that the error is minimized. In this existing system, the total ensemble error of the as-is ensemble system is 6.5%, which is relatively high. Second, the existing system has been tested against Colon dataset only, while a more comprehensive system that is tested against various cancer datasets is required. Third, in the case of small sample size where substitution is extremely low biased due to over-fitting, bolstering or spreading the error of the misclassified instances is not practically suitable. This is because it contributes in increasing the bias; specially with over-fitting rules classifiers. Fourth, the choice of the method of combining votes of the base classifiers into one predicted decision should consider the accuracy -or the weight- of the ensemble member. This is in order to determine to what extent an ensemble member should participate in the final decision. Also, the size of the dataset on which the decision has been taken should be considered (e.g. Naïve Bayes combination). Moreover, the choice of the feature selection methods that preserve the semantics of the features 'genes' and further select more informative and relevant features (e.g. Markov Blanket) is a more suitable from biological point of view. Currently, some feature selection techniques on microarray data have been proposed [5–7].

3. Proposed system

We propose an ensemble system which is a set of individually trained classifiers whose decisions are combined typically with majority voting, weighted voting or other relatively simple techniques such as stacking or Naïve Bayes combination. Researches show that generally ensemble classifier outperforms the performance of the best member classifier in the squad [8–10].

The proposed system addresses the first three drawbacks of the existing system; namely, enhancing result accuracy, applying the ensemble technique to more cancer types, and mitigating the effect of over-fitting. Block diagram of the proposed system is presented in Fig. 2. The shaded blocks in the diagram indicate a contribution is done in this block. The following points explain the function of each module as well as the modifications done at each one:

- **Gene Expression Dataset:** In this module the dataset that the system will run is defined. Sequence of actions to be carried out on those datasets is defined amongst file reading, loading and connecting to dataset repository. The proposed ensemble system has been applied to Colon dataset as the case in the existing system introduced by [11]. In addition, the proposed system has been modified, tuned and tested against Leukemia and Breast cancer datasets to boost the confidence in applying the proposed system to different cancer types and to emphasize the suitability of the proposed system to be applied in this sensitive domain.
- **Preprocessing Module:** According to the dataset defined in the gene expression dataset module, preprocessing module prepares the dataset to be manipulated. Preparation includes filtering, thresholding, logarithmic transformation and data normalization. Those procedures are essential to be done before actual classification take place.

- **Gene Selection Module:** In terms of cancer classification, this massive number of microarray data doesn't bring more discriminative power; degrade the accuracy of the classifier. Thus, features need to be decreased to a feature subset with the most significant features or genes that are capable of discriminating different classes. So, the objective of feature selection -also known as subset selection- is to get rid of redundant or irrelevant features in order to decrease the complexity of classification, reduce computational time and storage requirements, and improve the prediction of the classifier. This, in turn, improves classification accuracy, facilitate data visualization and provide a better understanding of the underlying data by revealing interrelationship among features. Another advantage of feature selection is that it mitigates the effect of over-fitting. Lack of generalization caused by over-fitting leads to a high accuracy level on training phase; however, a very bad performance on unseen samples and also results. This happens because the classifier learned all about the fine details of the training data (i.e., memorized it) and hence its performance is unexpected while predicting the class membership of new samples. Feature selection can be categorized into three main categories based on the way of combing the selected feature subset with the classification model: Wrappers, Filters, and Embedded methods. The proposed system utilizes three different feature selection algorithms:
  - Backward Elimination Hilbert-Schmidt Independence Criterion "BAHSIC" [2].
  - Extreme Value Distribution based gene selection "EVD" [3].
  - Singular Value Decomposition Entropy gene selection "SVDEntropy" [4].
- **Classification Ensemble Module:** The proposed ensemble system consists of 5 base classifiers; all base classifiers implements 3-NN algorithm. Each classifier utilizes its own feature selection parameters in order to ensure the diversity of the ensemble. The first three classifiers utilize BAHSIC feature selection algorithm with different number of genes to select. The number of genes selected from genes' pool is a user-defined parameter and it was set to 50, 5 and 25, respectively. The forth base classifier utilizes EVD gene selection algorithm with automated algorithm for defining the number of genes to select; number of genes selected by the algorithm was 49 genes for Colon dataset, 224 genes for Leukemia dataset and 5127 genes for the Breast cancer dataset; the last base classifier utilizes SVD entropy gene selection algorithm; the number of genes returned by the algorithm was 240 genes for Colon dataset, 187 genes for Leukemia dataset and 1236 genes for the Breast cancer dataset. As depicted in the block diagram of the proposed system, the parameters of the second and third classification members of the ensemble have been altered to improve the overall ensemble performance and to increase the diversity in the ensemble squad.
- **Post-processing Module:** In order to improve the accuracy, and hence, the certainty of the decision made by the ensemble, certain modifications have been applied over the existing system. The post-processing module can be broken into two main sub-modules:
  - ***Error Estimation Module:*** In gene expression data, evaluating the performance of a machine learning algorithm with an unbiased classification error estimator is an important issue. Braga-Neto [8] proposed an error estimator called "Bolstered Resubstitution Error" (BRE) that is based on bolstering the data instances by setting suitable kernels at each data instance. BRE is less prone to biasness and considered to be a very fast error estimator compared with other error estimators such as bootstrap. It is computed by a small number of Monte Carlo samples. Bolstering has the effect of reducing the variance of the error estimation and sometimes the bias too [1]. Unlike cross validation, BRE make use of the whole available data by generating M artificial test samples based on the available ones on the training set. It has been proven that BRE is more accurate and faster than traditional error estimation techniques such as LOO and the 0.632 bootstrap [5]. Bolstered resubstitution error can be defined as:

$$\varepsilon^{\wedge *} = \frac{1}{N}\sum_{i=1}^{N}\left(\int_{A_1} f_i^*(x - x_i)dx I_{y_i=0} + \int_{A_0} f_i^*(x - x_i)dx I_{y_i=1}\right)$$

Such that N is the number of samples and $f^*(x)$ is a pairwise bolstering kernel smoothing function intends to reduce the variance by taking a value in the interval [0,1], and can be denoted as

$$f_i^* = \frac{1}{(2\pi)^{D/2}\sigma_i^D}\exp\left(-\frac{x^2}{2\sigma_i^2}\right)$$

where σ_i's the standard deviation are automatically computed by the bolstered error, not a user defined parameter.

Fig. 1 shows how to generate M random normally distributed samples using Marsaglia polar method. Marsaglia polar method aims to generate Gaussian pseudo-random numbers from a uniform pseudo-random number. That is done by:

a. For M sample points, generate a pair of uniformly distributed numbers using Box-Müller transform.
b. Transform the points from the interval [0,1] to a new interval [−1,1] by the following simple equations: $u_i = 2u_i - 1, i = 1,2; r = u_1^2 + u_2^2$
c. If r ⩾ 1 or r = 0, go to (a.)
d. Else if r < 1, which means the pair of points (u_1, u_2) are uniformly distributed points inside the unit circle
e. Transform (u_1, u_2) to normally distributed points $z_i = \frac{u_i\sqrt{-2\ln r}}{r}, i = 1,2$
f. Obtain the desired normally distributed numbers based on the current mean and standard deviation of the training point in question $z_i' = \mu + z_i\sigma, i = 1,2$

The First modification is in the implementation of BRE. It is more convenient not to propagate the error of incorrectly classified training instances as this further increases the biasness. By setting σi = 0 -which means no bolstering is made for point i, the biasness is decreased. However, it increases the variance of the error estimation. In situations where low variance and high bias classifiers are used (e.g. K-NN, Naïve Bayes, CART, etc.), increasing the variance helps decreasing the tendency of the classifier to over-fit (third drawback of the existing system). This version of BRE is called semi-bolstered re-substitution error "sRRE". sBRE yields better results over BRE in dealing with small size problems such as

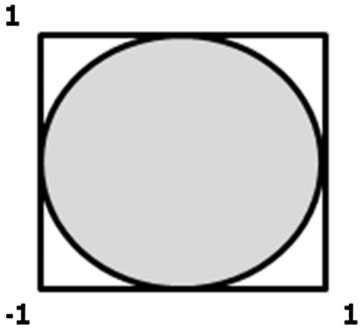

Figure 1. Gaussian distribution random samples drawn from uniform distribution inside the unit circle.

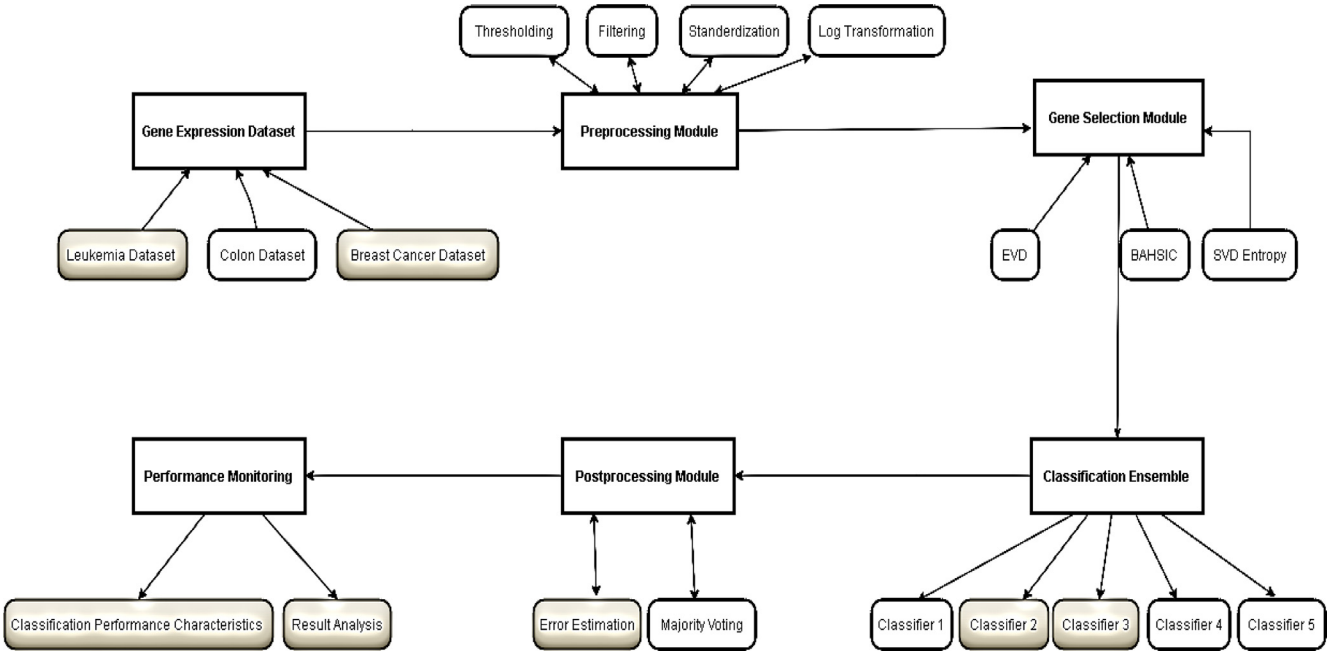

Figure 2. Block diagram for "Gene Expression Based Cancer Classification" proposed system.

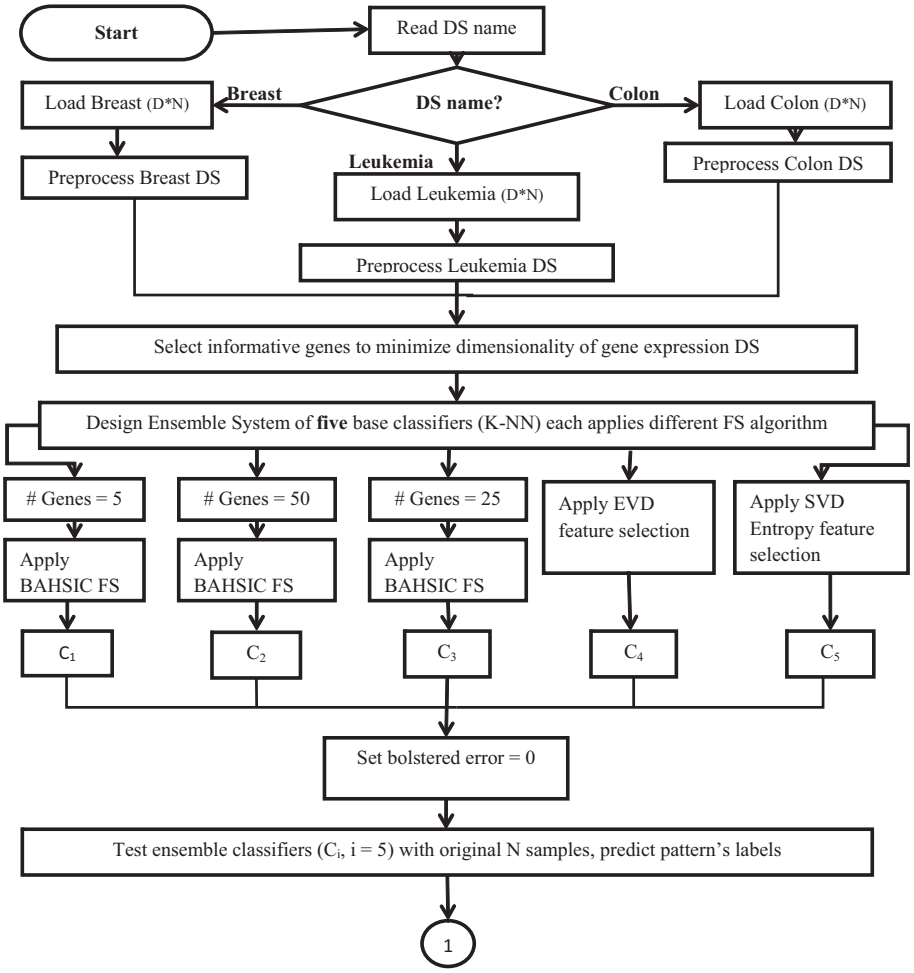

Figure 3. Flowchart showing the execution logic of the proposed system.

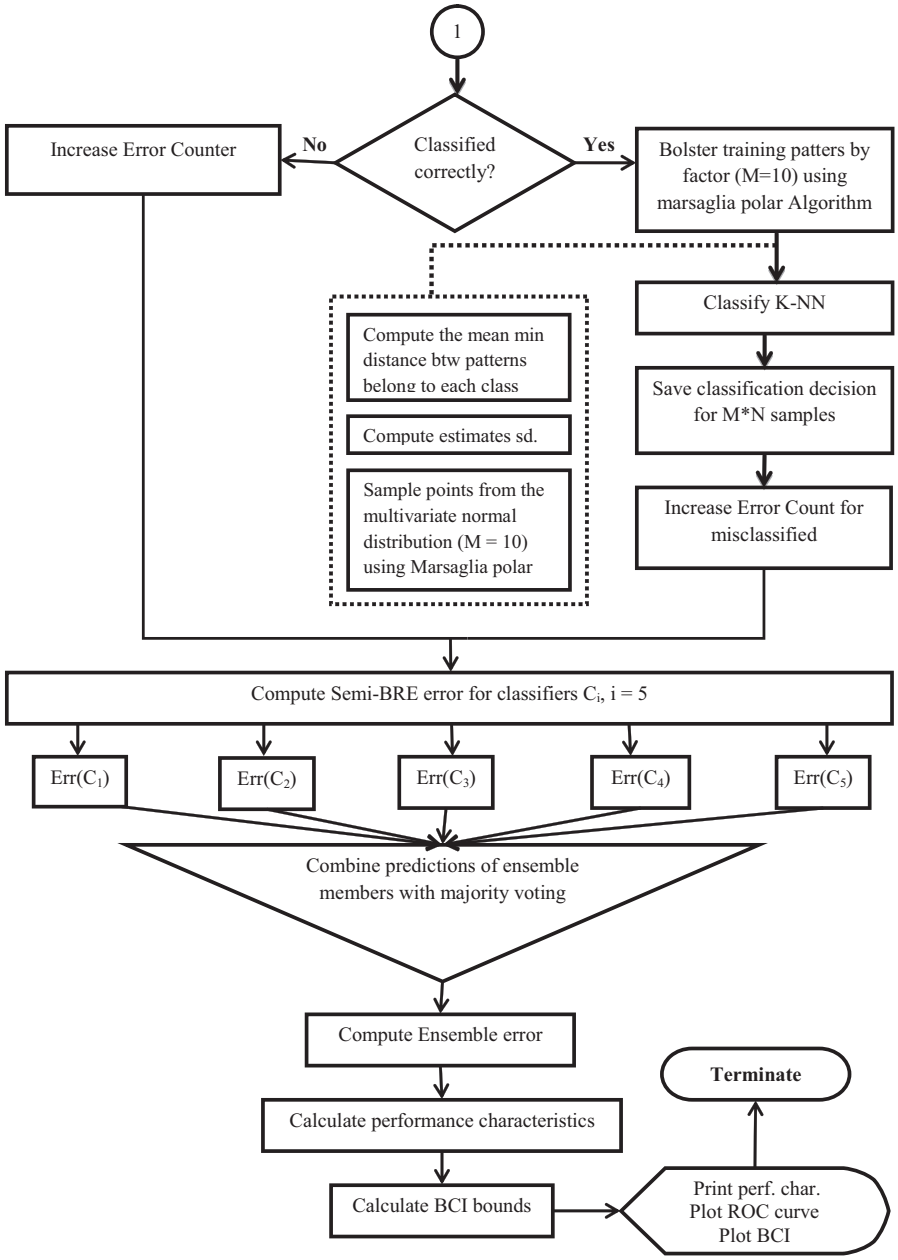

Fig. 3 (*continued*)

microarray data in conjunction with nearest neighbor classifier as recommended by [6]. Besides, it is even thousands of times faster than other error estimators [11,6,12].

- ***Majority Voting Module:*** The last step in ensemble classification process is to combine the predictions made by ensemble member classifiers. It varies from very simple forms of vote count or majority voting which assign a sample to the class that takes largest number of votes predicted by ensemble members (ties are broken arbitrary), weighted votes which take into consideration the weight or the accuracy of the ensemble member in order to determine to what extends an ensemble member participates in the final decision. The proposed system applies simple majority voting as a combining technique.
- ***Classification Performance Characteristics:*** In order to measure the performance of the algorithms used for classifying any dataset, it's reasonable to unify the performance characteristics -such as classification accuracy or true or false positive rate- that will be used to compare and evaluate results attained during runs by different algorithms. Receiver operating characteristic curve (ROC), area under ROC and Bayesian credible interval (BCI) and other performance measures are introduced to cover more aspects of the result obtained. To construct a ROC curve, the classifier need to provide a score or probability estimate represent class membership of each instance in the test set used. These probabilities are used along with the actual class labels and the predicted class labels for the test set instances to calculate the TPR and FPR. One way of constructing the ROC list is by sorting the probabilistic outputs in a descending order of magnitude, moving along the sorted list and updating the counters of false and true positive. The Area under curve (AUC) is a way to summarize the ROC curve. It is calculated by cumulatively adding the areas under ROC curve between each two following points on the curve. As AUC is an area, it's a scaler

value varies between zero and one. The AUC of random guessing is 0.5 and represented by a line from the origin (0, 0) to the point (1, 1).

For the implementation, Matlab was used as it provides a rich library of useful toolboxes such as Bioinformatics Toolbox™ and Statistics Toolbox™. This, in turn, facilitates the computations and analysis required throughout this research.

The provided flowchart (Fig. 3) illustrates the process of cancer classification based on gene expression dataset in an orderly manner and includes various stages of processing the data and classifying the provided samples into its corresponding class.

The process of cancer classification initiated with reading the name of the dataset need to be manipulated; in the case of the proposed system, there are three datasets are taken into account each has its own preprocessing steps -as explained in later sections- preceding the classification process itself.

Typically, gene expression datasets suffers from high dimensionality. Thus, gene selection mechanism must be taken into consideration. It involves finding the most informative genes in among the massive number of available genes. The proposed system uses three different feature selection methods for selecting the best gene subsets and passes it over for classification. The first three members of the ensemble apply BAHSIC gene selection algorithm with different predefined number of genes to select, the forth apply EVD algorithm and the fifth apply SVD algorithm.

After gene selection take place, the dataset become ready to be explored by the classification system. The selected subsets works as training sets for the classifier. The proposed system applies a committee of classifiers whose decision is then combined called ensemble system. The designed ensemble system consists of five base classifiers. In the proposed system, K-NN algorithm was selected in the design of the base classifiers for its simplicity and because generally similarity based classifiers don't ignore the correlation and gene interactions. Although K-NN suffers from scalability problem; however, works fine with data with small size as the case with cancer's gene expression datasets in use.

The five base classifiers work in parallel as follows:

- Each sample in the selected subset -generated by one of the three used gene selection algorithm- is passed to the corresponding 3-NN classifier. The classifier in question aims to predict the correct label of the sample.
- If the sample is classified correctly, semi-BRE start generating ten artificial test samples based on this correctly classified sample. In other words, BRE start bolstering each training sample by factor (M = 10) using Marsaglia polar Algorithm (as explained in the previous section), test the classifier with those samples to calculate the error, update the error count for statistical purposes and preserve the classification decision for bolstered samples for later use.
- If the sample is misclassified, there is no need to bolster it not to propagate the error. Semi-BRE updates the error count and goes for the next sample.
- After classifying all samples, semi-BRE error is calculated for each classifier in the ensemble.
- The last step in the ensemble classification is to combine the preserved predictions from all base classifiers to form the final decision of the ensemble system as a whole. Each sample is assigned to the class that takes more votes.
- Calculate the ensemble error and compare it to the base classifiers' decision, calculate the confusion matrix, plot receiver operating characteristic curve (ROC), calculate the area under ROC curve (AUC), calculate Bayesian confidence interval (BCI) upper and lower bound and plot it.

4. Results and analysis of the proposed system

In this section, results obtained by the proposed system are examined. Results of the proposed system are compared with related work results. For performance testing, total ensemble error rate, base classifiers error, AUC and BCI are calculated. Experiments are performed against three benchmark cancer datasets; namely, Leukemia, Colon and Breast cancer datasets:

- **Leukemia Dataset:** The acute leukemia data was first published in [3]. Dettling [4] preprocessed this data via logarithmic transformation to base 10 and feature selection. As a result, there were actually 3571 genes after data preprocessing. Gene expression matrix for 3571 gene expression values on 72 individuals (3571 by 72) along with a vector of class index (72 by 1) is presented. The Leukemia dataset can be found at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC151171.
- **Colon Dataset:** Colon tumor is a type of cancerous growth found in Colon tissues. Colon dataset is presented in [1]. It contains expression profiles of 2000 genes for 62 Colon samples, 40 of them are labeled as normal tissues "negative", while the other 22 samples are labeled as tumor tissues "positive". Data can be found at http://microarray.princeton.edu/oncology. Dataset preprocessing includes the logarithmic transformation to base 10 followed by normalization to zero mean and unit variance.
- **Breast cancer Dataset:** This dataset contains patients' outcome prediction for breast cancer. It includes the expression levels of 24,481 genes. The training data contains 78 samples, 34 samples of them are labeled as "relapse" which is from patients who had developed distance metastases within 5 years, and the rest 44 samples are from patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years which are labeled "non-relapse". Correspondingly, there are 12 relapse and 7 non-relapse samples in the testing data set. Dataset can be found in http://datam.i2r.a-star.edu.sg/datasets/krbd/BreastCancer/BreastCancer.html.

As shown in Table 1, results' interpretation can be broken down into three points:

- The first three base classifiers in the ensemble utilize BAHSIC feature selection algorithm [2] with user defined number of selected genes to return for classification while classification is done by 3- or 5- nearest neighbor classifiers. Values of 50, 5, and 25 were given as an input to the BAHSIC algorithm denoting the number of genes to select from genes pool. The bolstered error rate was 15.0%, 23% and 16.1%; the Bayesian credible interval was [12.5, 18.1], [20.6, 27.3] and [13.5, 19.3]; the AUC was 0.92, 0.80 and 0.91; respectively. BAHSIC algorithm yielded very poor prediction accuracy in the case of 5 genes and the more the genes the better performance achieved.
- The fourth ensemble member utilizes EVD gene selection along with a 3-nearest neighbor classifier. The number of Monte-Carlo samples is 10. Number of genes selected by EVD gene selection algorithm is 49 genes. EVD gene selection is superior over BAHSIC algorithm with approximately the same number of selected genes. The bolstered error rate was 9.5%; the Bayesian credible interval was [7.4, 12.0] and the AUC was 0.97.
- The fifth ensemble member utilizes SVDEntropy feature selection with simple ranking as a criterion for picking genes along with 3-nearest neighbor classifier. The number of Monte-Carlo samples remained 10. The bolstered error rate was 8.1%; the Bayesian credible interval was [6.3, 10.6]; the AUC was 0.97;

Table 1
Summarize the number of genes returned after implementing feature selection algorithms versus error rate (%).

	As-is system					Proposed system				
Feature selection	BAHSIC			EVD	SVD	BAHSIC			EVD	SVD
Base classifier	BC1	BC2	BC3	BC4	BC5	BC1	BC2	BC3	BC4	BC5
Genes (Colon DS)	50	5	25	49	316	50	5	25	49	240
Error (Colon DS)	13.40	31.10	16.60	9.20	8.90	12.74	27.58	18.87	10.32	9.03
Genes(Leukemia)	50	5	25	61	187	50	5	25	224	187
Error (Leukemia)	4.86	30.42	7.78	1.67	1.67	7.22	27.36	8.47	1.39	2.64
Genes (Breast DS)	50	5	25	5127	1236	50	5	25	5127	1236
Error (Breast DS)	1.90	31.03	2.59	0.00	1.72	0.86	37.76	2.76	0.00	1.72

316 genes were selected. That is, classification performance turned out to be similar to that in EVD gene selection algorithm, but it was achieved with six times extra genes.

Majority voting is used to combine classifiers' predictions made by all five ensemble members into one decision. Errors attained by ensemble members on a single run were 11.77%, 30.16%, 17.10%, 8.71% and 7.90%, while the ensemble error was only 6.5%; the BCI for the ensemble error is [4.90 8.80]; The area under curve was 0.97; the number of genes selected by the five base classifiers was equal to 50, 5, 25, 49, and 316, respectively. The performance of the ensemble outperforms the performance of the best ensemble member classifier.

Fig. 4 shows results from the as-is system and the proposed system when applied to the first benchmark dataset namely, Leukemia cancer dataset. As shown in Fig. 4(a) ROC Curve of as-is ensemble where AUC = 0.98; 4(b) Bayesian Posterior Distribution as-is ensemble BCI = [0.90, 2.60] and ensemble error = 1.39%; 4(c) ROC Curve of the proposed ensemble system where AUC increase to AUC = 1.00; 4(d) Baysian posterior distribution graph shows that the BCI curve shifted to the left which implies better classification accuracy BCI = [0.10, 0.60]; errors attained by the five base classifiers were 7.22%, 27.36%, 1.39%, 2.64% and 1.94%, respectively. While the total ensemble error for the proposed system was only 0.00%.

Fig. 5 shows results from the as-is system and the proposed one when applied to the second benchmark dataset namely, Colon cancer dataset. Fig. 5(a) ROC Curve of as-is ensemble where AUC = 0.97; 5(b) Bayesian Posterior Distribution as-is ensemble BCI = [4.90, 8.80] and ensemble error = 6.5%; 5(c) ROC Curve of the proposed ensemble system where AUC increase to AUC = 0.99; 5(d) Baysian posterior distribution graph shows that the BCI curve shifted to the left which implies better classification accuracy BCI = [2.70, 5.80]; errors attained by base classifiers on a single run were equal to 12.74, 27.58, 18.87, 10.32 and 9.03, respecively. While the proposed ensemble system's error was decreased to 3.87%.

Fig. 6 shows results from the as-is system and the proposed one when applied to the third benchmark dataset namely, Breast cancer dataset. Fig. 6(a) ROC Curve of as-is ensemble where AUC = 1.00; 6(b) Bayesian Posterior Distribution as-is ensemble BCI = [0.30, 1.60]; 6(c) ROC Curve of the proposed ensemble system where AUC = 1.00; 6(d) Bayesian posterior distribution graph shows that the BCI curve shifted to the left which implies better

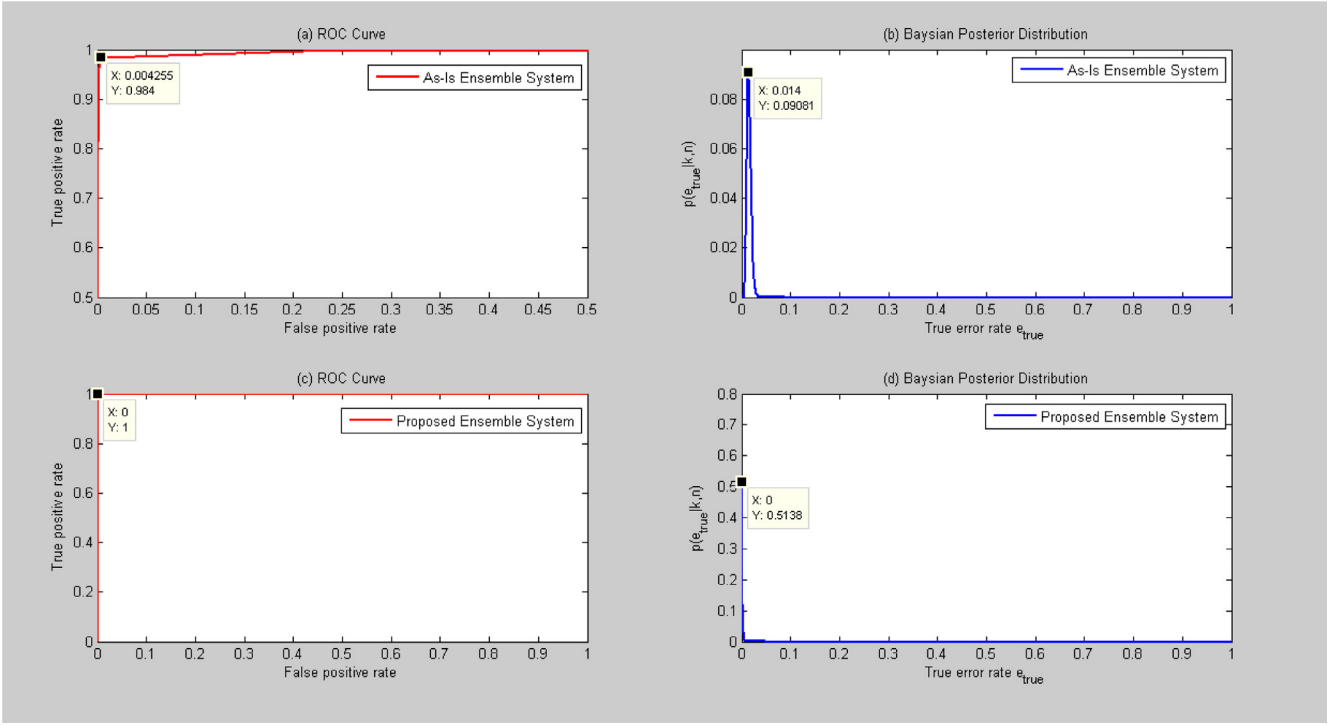

Figure 4. Comparison between the performance of the as-Is ensemble system and the proposed system when applied over Leukemia dataset in terms of (a) ROC curve and (b) Bayesian posterior distribution.

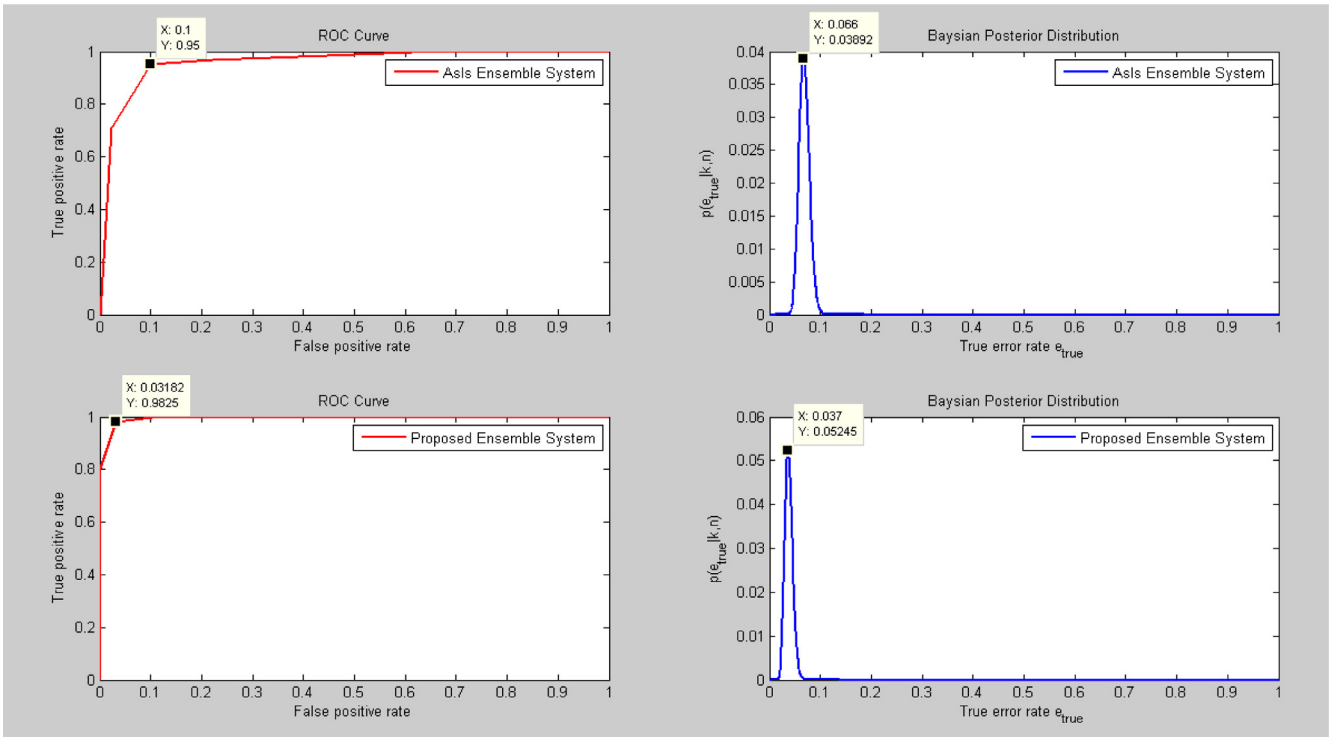

Figure 5. Comparison between the performance of the as-Is ensemble system and the proposed system when applied over Colon dataset in terms of (a) ROC curve and (b) Bayesian posterior distribution.

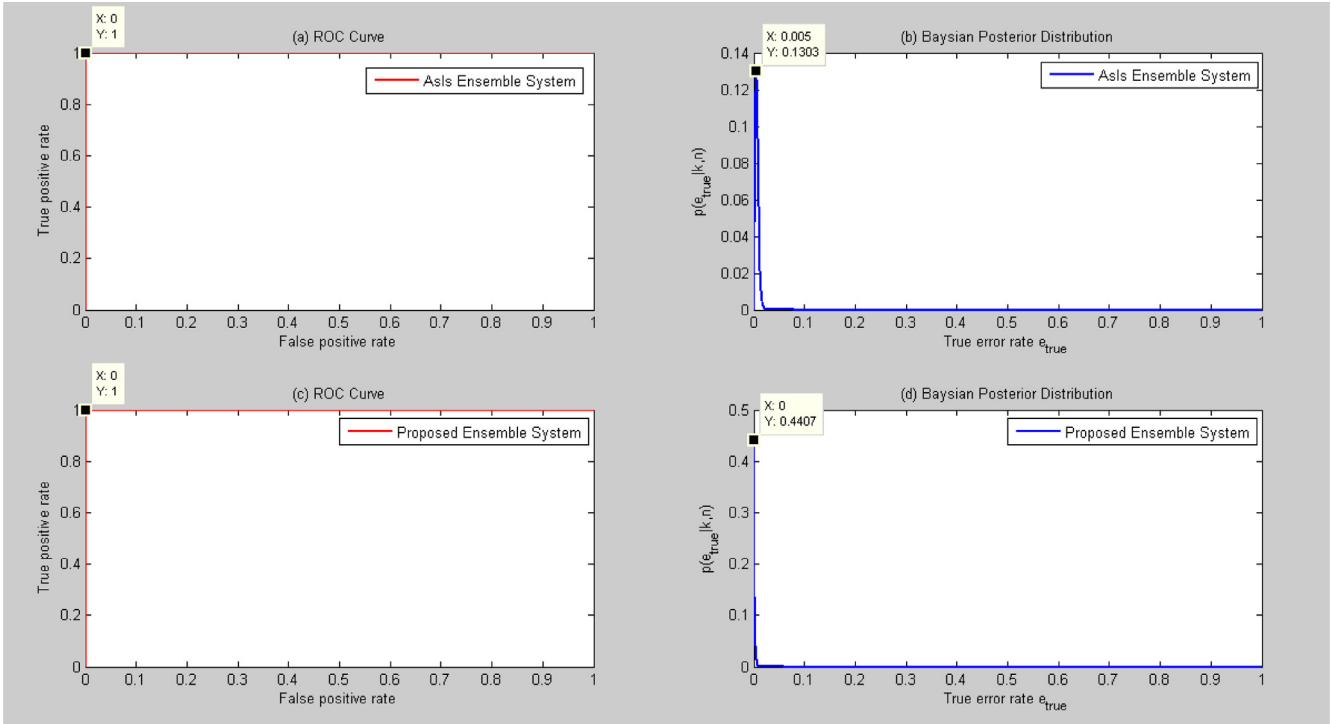

Figure 6. Comparison between the performance of the as-Is ensemble system and the proposed system when applied over Breast dataset in terms of (a) ROC curve and (b) Bayesian posterior distribution.

classification accuracy BCI = [0.10, 0.70]; errors attained by base classifiers on a single run were equal to 1.21, 37.76, 2.76, 0.00, and 1.72, respectively; while, the proposed ensemble system's error was decreased to 0.00%.

From the above analysis, one can conclude that results of the proposed system indicate enhancements in the performance characteristics when applied against Colon and Leukemia cancers. For the breast cancer, both the proposed and existing systems recorded

Table 2
Summarization of the performance characteristics of the as-is ensemble and the proposed ensemble system.

Performance characteristics	As-is system			Proposed system		
	Colon	Leukemia	Breast	Colon	Leukemia	Breast
Ensemble Error	6.5%	2.6%	0.52%	3.87%	0.14%	0.00%
BCI	[4.90, 8.80]	[1.10, 3.00]	[0.30, 1.60]	[2.70, 5.80]	[0.20, 0.90]	[0.10, 0.70]
AUC	0.97	0.95	1.00	0.99	1.00	1.00

impressive results as. Table 2 summarizes the ensemble classification error, BCI and AUC metrics for Colon, Leukemia and Breast cancer datasets.

5. Conclusion

In this report, a new ensemble system for Cancer classification based on gene expression profiles was presented. The approach followed lead to a fast and adequate system that outperforms the ensemble system suggested by [1]. It also overcame the three addressed drawbacks; namely, enhancing result accuracy, covering more cancer types, and mitigating the effect of over-fitting.in this work, K-NN classifier has been used as a base member of the ensemble. In future work, more different classifiers can be used as base members. Moreover, this system can be applied to other benchmarks especially multiclass datasets.

Acknowledgments

This research would not have been accomplished without the valuable assistance and support from those who have contributed to the preparation and completion of this paper.

I would like to express my profound thanks to my supervisors for their enthusiastic guidance, continuous support and kind encouragement throughout the whole period of my study.

References

[1] Okun O. Feature selection and ensemble methods for bioinformatics: algorithmic classification and implementations. Med Inform Sci Ref 2011.
[2] Song L, Smola A, Gretton A, Borgwardt KM, Bedo J. Supervised feature selection via dependence estimation. In: Proceedings of the 24th international conference on machine learning. ACM; 2007. p. 823–30. 2007, June.
[3] Li W, Sun F, Grosse I. Extreme value distribution based gene selection criteria for discriminant microarray data analysis using logistic regression. J Comput Biol 2004;11(2–3):215–26.
[4] Varshavsky R, Gottlieb A, Linial M, Horn D. Novel unsupervised feature filtering of biological data. Bioinformatics 2006;22(14):e507–13.
[5] Dougherty ER, Sima C, Hanczar B, Braga-Neto UM. Performance of error estimators for classification. Curr Bioinform 2010;5(1):53–67.
[6] Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004;20(3):374–80.
[7] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46(1–3):389–422.
[8] Vu TT, Braga-Neto UM. Is bagging effective in the classification of small-sample genomic and proteomic data? EURASIP J Bioinf Syst Biol 2009;2009(1):1.
[9] Kuncheva LI. Combining pattern classifiers: methods and algorithms. John Wiley & Sons; 2004.
[10] Segal MR. Machine learning benchmarks and random forest regression. Center Bioinform Molec Biostat 2004.
[11] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 1999;96(12):6745–50.
[12] Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004;20(3):374–80.