Cancer Classification from DNA Microarray Data using mRMR and Artificial Neural Network

M. A. H. Akhand[1], Md. Asaduzzaman Miah[2], Mir Hussain Kabir[3], M. M. Hafizur Rahman[4]
Dept. of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh[1, 2, 3]
Dept. of Communication & Networks, College of Computer Science & Information Technology
King Faisal University, Al Hufũf, Al-Hasa, Saudi Arabia[4]

***Abstract*—Cancer is the uncontrolled growth of abnormal cells in the body and is a major death cause nowadays. It is notable that cancer treatment is much easier in the initial stage rather than it outbreaks. DNA microarray based gene expression profiling has become efficient technique for cancer identification in early stage and a number of studies are available in this regard. Existing methods used different feature selection methods to select relevant genes and then employed distinct classifiers to identify cancer. This study considered information theoretic based minimum Redundancy Maximum Relevance (mRMR) method to select important genes and then employed artificial neural network (ANN) for cancer classification. Proposed mRMR-ANN method has been tested on a suite of benchmark datasets of various cancer. Experimental results revealed the proposed method as an effective method for cancer classification when performance compared with several related exiting methods.**

Keywords—Cancer classification; gene expression data; minimum redundancy maximum relevance method; artificial neural network

I. Introduction

Cancer is the uncontrolled growth of abnormal cells in the body and is a major death cause nowadays. In a normal person without cancer, a healthy cell divides in a controlled way and produce new healthy cells. A cell with cancer grows out of control, divides and invades other tissues. All the daughter cells of a cancer cell are also cancerous. A cell changes its nature because mutation(s) have occurred in its genes. Cancer can affect anybody at any stage of life but people in older stage are more likely to be affected by cancer. The reason is that DNA may become damaged when old or may get worse due to the damage that happened in the past [1].

Cancer classification is the vital issue in DNA microarray gene expression profiling. Cancer may arise anywhere in the human body, and it names are remarked as body parts such as colon cancer, lung cancer, breast cancer. It is notable that cancer treatment is much easier in the initial stage rather than it outbreaks. DNA microarray based gene expression profiling has become an efficient technique for cancer identification in early stage. For classification, the first step is to recognize a small subset of genes which are primarily responsible for the disease [2]. And then look deep insight the selected genes for classification employing distinct classifiers.

A number of techniques have been investigated in past several years for cancer classification from DNA microarray gene expression data. A method used distinct classification method on selected genes with a particular feature selection technique. Xu et al. [3] investigated a method combining artificial neural network (ANN) and particle swarm optimization (PSO). Discrete binary PSO is employed for gene selection as well as dimensionality reduction. ANN is used to classify cancer from the selected genes. A large B-cell lymphoma dataset was considered to test the method.

Takahashi et al. [4] investigated a hybrid method of projective Adaptive Resonance Theory based ANN and boosted fuzzy classifier with SWEEP operator for cancer classification. They combined wrapper and filter approaches for gene selection. The method was tested on acute leukemia and brain tumor.

Ghorai et al. [2] investigated nonparallel plane proximal classifier (NPPC) ensemble method for cancer classification. At first, they trained a number of classifiers with mutual information criterion based selected genes. Finally, classifiers considered for the ensemble based on their performance on a validation set. The method was tested on colon and ALL/AML cancer.

Acharya et al. [5] employed Archived Multi objective Simulated Annealing (AMOSA), a multi objective optimization based clustering technique, for cancer classification. The developed technique was evaluated for three benchmark datasets: adult malignancy, brain tumor and small round blood cell tumors.

Arunkumar and Ramakrishnan [6] used extreme learning machines (ELMs) on microarray gene expression data for cancer classification. They extracted features using correlation coefficient prior to classification. The developed method was tested on several benchmark datasets: ALL/AML, CNS, Lung Cancer, Ovarian Cancer and Prostate Cancer.

Recently, Alshamlan et al. [7] investigated support vector machine (SVM) along with hybrid gene selection for cancer classification. At first, artificial bee colony (ABC), a swarm intelligence based optimization approach, was used in analyzing a microarray gene expression profile. Then, information theoretic based minimum Redundancy Maximum Relevance (mRMR) technique was combined with ABC for hybrid feature selection. Finally, SVM was used to classify cancer from features of the selected genes. They tested the algorithm on several gene expression microarray datasets including colon, leukemia, ALL/AML cancer. Rathore et al.

[8] investigated gene expressions based colon classification (GECC) using different feature selection methods including mRMR and ensemble of SVMs. A modified version of SVM, called Transductive SVM, is also investigated for cancer classification by Maulik et al. [9].

This study investigates ANN based cancer classification on selected genes from gene expression data. First, mRMR has been employed for gene selection and then ANN is used for cancer classification. Proposed mRMR-ANN has been tested on a suite of benchmark datasets of various cancers and outperformed existing methods while compared with those methods.

The outline of the remaining paper is as follows. Section II explains the proposed cancer classification from DNA microarray data using mRMR and ANN. Section III is for experimental studies which presents outcomes of the proposed method in solving benchmark datasets as well as compares with other related methods. At last, Section IV gives a brief conclusion of the paper.

II. Proposed Cancer Classification from DNA Microarray Data

There are three major steps in the proposed mRMR-ANN method: data preprocessing, gene selection by mRMR and finally classification with ANN. Fig. 1 shows the major steps of the proposed method and following subsections briefly describes the steps.

A. *Preprocessing of Microarray Gene Expression Data*

The presence of noise in the microarray gene expression data is common. Data can also be missing in some cases due to various stages of preparation. Together with small sample size, the classification task is challenging. That is why preprocessing is performed on the data owing to achieve better classification accuracy. On the other hand, data are normalized to transform all the data in same range which is essential for proper operation of classifiers.

In this study, K neighbor method is used to fill the missing data which is the extension of the nearest neighbor method. The method defines the missing value based on the values of K nearest neighbors from the testing sample [10]. Euclidean distance is commonly used to measure the distance between the data samples. Eq. (1) shows the distance between the two data points x_0 and x_i in the p dimensional space.

$$\text{Dist}(x_o, x_i) = (|x_{o1} - x_{i1}|^2 + |x_{o2} - x_{i2}|^2 + \cdots + |x_{op} - x_{ip}|^2)^{\frac{1}{2}} \quad (1)$$

Data normalization is a process in which data attributes within a data model are organized to increase the cohesion of entity types. The unity based normalization has been incorporated in this study where values of a particular attribute are transformed between 0 and 1. Eq. (2) is for the transformation.

$$X(t)_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}, \quad (2)$$

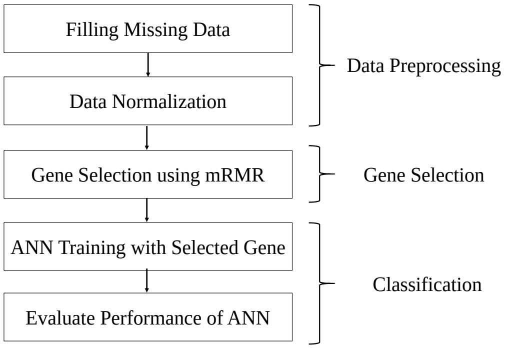

Fig. 1. Steps of Proposed mRMR-ANN Method for Cancer Classification.

where X_i and $X(t)_i$ are actual and transformed data points, respectively; X_{min} is the minima among all the data points; and X_{max} is the maxima among all the data points.

B. *Gene Selection using Minimum Redundancy and Maximum Relevance (mRMR)*

Selection of a small subset of appropriate genes from a lot of genes in microarray data is essential for precise cancer classification [11]. Although any feature selection method may be useful for this purpose, a conventional method typically ranks genes according to their differential expressions and picks the top-ranked genes for classification task. On the other hand, feature sets obtained through the minimum redundancy–maximum relevance framework might perform better classification than any rank based approach. In this regard, information theoretic based mRMR feature selection has been considered in this study which is frequently used to identify important and relevant features from the given data.

mRMR selects a feature subset that best characterizes the statistical property of a target classification variable, subject to the constraint that the selected features are marginally as similar to the classification variable as possible, but mutually as dissimilar to each other as possible. Eq. (3) shows the relevance of a feature set *S* for the class *c* which is defined by the average value of all mutual information (MI) values between the individual feature f_i and the class *c*.

$$\text{D(S, c)} = \frac{1}{|S|}\sum_{f_i \in S} I(f_i, c) \quad (3)$$

In the equation *I(.)* indicates MI function. Eq. (4) shows the redundancy of all features in the set *S* which is the average value of all MI values between the features f_i and $f_{j.}$

$$R(S) = \frac{1}{|S|^2}\sum_{f_i, f_j \in S} I(f_i, f_j) \quad (4)$$

Finally, the mRMR criterion is shown in Eq. (5) which is a combination of two measures given in Eq. (3) and Eq. (4).

$$\text{mRMR} = \max\left[\frac{1}{|S|}\sum_{f_i \in S} I(f_i, c) - \frac{1}{|S|^2}\sum_{f_i, f_j \in S} I(f_i, f_j)\right] \quad (5)$$

In the equation *max(.)* function is for combination of relevance and redundancy estimation. At a glance, mRMR is a filter method maintaining trade-off between relevancy and redundancy. It ranks features according to criteria and provides top ranked user defined number of features. A demonstration of gene selection through mRMR is shown in Fig. 2. Among number of *F* genes, number of *K* (user defined number) genes has been selected using mRMR which are most informative for classification.

C. Classification using Artificial Neural Network (ANN)

ANN is designed with the goal of building intelligent machines to solve complex perceptual problems by mimicking special features biological neurons in the human brain [12]. The key elements of ANN are artificial neurons which have the information processing capability. Neurons are connected with other neurons though synaptic weights with a particular fashion for a particular task, such as data classification in this study. The synaptic weight values of an ANN are adjusted through a learning process to perform the specific task.

Fig. 3 displays the structure of ANN that employed in this study for cancer classification. The ANN has three different layers with feed forward architecture. The input layer is a set of input units (i.e., neurons), which receive the elements of feature vectors. Each input neuron is connected to the neurons of the hidden layer through different weight values. The hidden neurons are also fully connected to neurons of the output layer through another set of weight values. The output layer generates the response of ANN for a pattern placed to the input layer. The information given to the network is propagated from the input layer to the output layer through the hidden layer. And the weights $W_1, W_2, \ldots W_n$ determine the influence of nodes of a layer in making decision to the next layer, i.e., hidden or output layer. For input vector $I = [I1, I2, \ldots In\ ]^T$, each input is multiplied by the associated weight for summed input of a hidden layer neuron as of Eq. (6). The positive weights excite and the negative weights inhibit the node output.

$$Ih = I^T . W = I_1 W_1 + I_2 W_2 + \cdots + I_n W_n = \sum_{i=0}^{n} I_i W_i \qquad (6)$$

Finally, the output of the hidden neuron is calculated through activation function as of Eq. (7) where φ is the magnitude offset or bias term.

$$Oh = f(Ih - \varphi) \qquad (7)$$

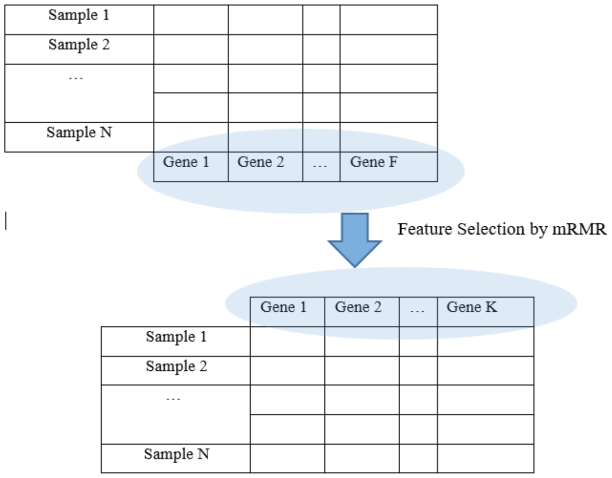

Fig. 2. Overview of Gene Selection using mRMR.

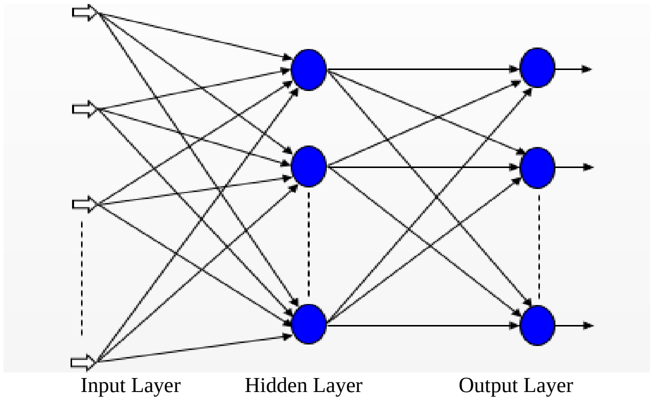

Fig. 3. Structure of Artificial Neural Network used in mRMR-ANN.

The output of the ANN is taken from the output layer and determines similar fashion of hidden layer using Eq. (6) and Eq. (7). For classification task, ANN needs to be trained to produce the desired input output mapping. In training, at first error is determined comparing actual output with desired output for training patterns and weight values are updated so that error is minimized. For *m* number of training data the squared error (*E*) can be given as:

$$E = \sum_{i=0}^{m} (t_i - o_i)^2 \qquad (8)$$

where *t* is the target/desired output and *o* is the calculated output from training data. The most popular training algorithm Back-Propagation (BP) is employed in this study for training the ANN [12].

III. Experimental Studies

This section experimentally investigates the efficacy of proposed mRMR-ANN method for cancer classification. A set of benchmark problems were chosen as a test bed and the performance of the mRMR-ANN compared with other popular methods. The experimental methodology were chosen carefully for fair comparison. An experimental analysis has also been presented for better understanding of the way of performance improvement in proposed method for cancer classification.

A. Benchmark Datasets

Four well studied microarray datasets have been considered for this study and those are Colon1, Colon2, ALL/AML and MLL. Colon1 dataset contains 62 samples and 2000 genes. There are 40 tumors and 22 normal samples. It is a binary class problem. Each of training and testing set contains 31 samples. Colon2 (i.e., Notterman Colon) dataset consists of 36 samples having 7,457 genes in each; out of which 18 samples are normal, and remaining 18 samples are malignant. ALL/AML dataset contains 72 samples of 7,129 genes; in which 47 samples are for acute lymphoblastic leukemia (ALL) and 25 samples are for acute myeloid leukemia (AML). Number of training and testing samples are 38 and 34, respectively. The mixed-lineage leukemia's (MLL) is a multiclass problem and dataset contains 72 samples of 12,533 genes. The subtypes are ALL (24 samples), AML (28 samples), and MLL (20 samples). Number of training and testing samples are 57 and 15, respectively. Table I shows summary of datasets with source reference.

TABLE. I. DESCRIPTION BENCHMARK MICROARRAY DATASETS

Dataset	Number of Classes	Total Genes	Total Samples
Colon1[13]	2	2,000	62
Colon2[14]	2	7,457	36
ALL/AML[13]	2	7,129	72
MLL[13]	3	12,582	72

B. Experimental Setup

Proposed mRMR-ANN has been implemented in Matlab R2015a. We have managed mRMR toolbox and its prerequisite of Mutual Information (MI) toolbox from MathWorks site. MI toolbox is kept in the same directory of the mRMR package. It is worth mentionable that Visual Studio is also necessary for functioning mRMR package because of many .cpp files in the package. We have run “makeosmex.m” file to execute .cpp extension file which are needed for computing MI. It was necessary to change log (2) to log (2.0) of all of the .cpp files. Inputs of the mRMR toolbox are *d* for all the features of dataset, *f* for the class labels of the dataset and *K* for the number of features to be selected. Output of the mRMR is the index number of the features in the dataset in the order of the relevance value. That means the most relevant feature is at first and the lowest relevant feature at the last position.

In ANN, sigmoid is used as activation function in Eq. (7) for hidden and output layers. Number of neurons at the input layer was number of features selected using mRMR. Number of neurons at the output layer was set according to the number of class of the problem. As an example, there will be three neurons in output layer to classify three cancer subtypes of MLL. On the other hand, number of neurons at the hidden layer is a user defined parameter and varied for better classification accuracy.

In order to measure the statistical significance of the proposed scheme, experiment has been repeated for 20 times on each setting for a particular problem. The average testing accuracy and their standard deviations are reported for each dataset. The experiments have been conducted on a PC with Windows 7 OS having system configuration Intel(R) Core(TM) i5-2520U CPU @ 2.5GHz with 8GB of RAM.

C. Experimental Results and Performance Comparison

This section first presents experimental results of proposed mRMR-ANN on each individual dataset for different settings and then compares outcome of it with prominent exiting methods. The number of selected genes (GN) through mRMR varied from 100 to 500. On the other hand, the number of hidden neuron (HN) of ANN varied from 20 to 200. ANN was trained over 500 iterations. The feature selection using mRMR and training with ANN were performed with training set and test set was reserved to check system performance on unseen data. Tables II to V show testing set classification accuracy (TSCA) for Colon1, Colon2, ALL/AML and MLL datasets for each individual setting. The results presented in the tables are the outcome of 20 independent runs of ANN for each setting.

TABLE. II. TEST SET CLASSIFICATION ACCURACY IN PERCENTAGE (%) FOR COLON1 CANCER

Num. of GN	Num. of HN	Best TSCA	Worst TSCA	Avg. TSCA	SD of Avg.
100	20	83.9	83.9	83.9	0
	50	83.9	83.9	83.9	0
	100	83.9	83.9	83.9	0
	200	83.9	83.9	83.9	0
200	20	83.9	83.9	83.9	0
	50	83.9	83.9	83.9	0
	100	83.9	83.9	83.9	0
	200	83.9	83.9	83.9	0
300	20	83.9	83.9	83.9	0
	50	83.9	83.9	83.9	0
	100	83.9	83.9	83.9	0
	200	83.9	83.9	83.9	0
400	20	83.9	83.9	83.9	0
	50	83.9	83.9	83.9	0
	100	83.9	83.9	83.9	0
	200	83.9	83.9	83.9	0
500	20	83.9	83.9	83.9	0
	50	83.9	83.9	83.9	0
	100	83.9	83.9	83.9	0
	200	**87.1**	**87.1**	**87.1**	**0**

TABLE. III. TEST SET CLASSIFICATION ACCURACY IN PERCENTAGE (%) FOR COLON2 CANCER

Num. of GN	Num. of HN	Best TSCA	Worst TSCA	Avg. TSCA	SD of Avg.
100	20	100	94.4	99.7	0.01
	50	88.9	88.9	88.9	0
	100	88.9	88.9	88.9	0
	200	94.4	88.9	89.2	0.01
200	20	100	100	100	0
	50	100	100	100	0
	100	100	100	100	0
	200	100	94.4	99.7	0.01
300	20	94.4	94.4	94.4	0
	50	100	94.4	99.2	0.02
	100	94.4	94.4	94.4	0
	200	100	100	100	0
400	20	100	94.4	99.7	0.01
	50	94.4	94.4	94.4	0
	100	100	100	100	0
	200	94.4	94.4	94.4	0
500	20	94.4	94.4	94.4	0
	50	100	100	100	0
	100	94.4	94.4	94.4	0
	200	**100**	**100**	**100**	**0**

TABLE. IV. TEST SET CLASSIFICATION ACCURACY IN PERCENTAGE (%) FOR ALL/AML CANCER

Num. of GN	Num. of HN	Best TSCA	Worst TSCA	Avg. TSCA	SD of Avg.
100	20	91.2	85.3	85.9	0.018
	50	88.2	88.2	88.2	0.000
	100	85.3	82.4	85.1	0.007
	200	91.2	85.3	87.5	0.021
200	20	94.1	91.2	91.5	0.009
	50	94.1	91.2	91.3	0.007
	100	97.1	88.2	91.1	0.017
	200	94.1	88.2	89.3	0.017
300	20	94.1	76.5	85.4	0.031
	50	97.1	94.1	94.3	0.007
	100	97.1	88.2	95.1	0.024
	200	91.2	82.4	90.3	0.027
400	20	97.1	82.4	91.3	0.031
	50	76.5	50	60.4	0.053
	100	97.1	85.3	92.8	0.029
	200	76.1	85.3	91.2	0.021
500	20	94.1	70.6	89.9	0.054
	50	79.4	61.8	64.6	0.053
	100	85.3	73.5	76.5	0.029
	200	**97.1**	**64.7**	**92.4**	**0.079**

It is observed from Table II for Colon1 cancer that GN values from 100 to 400 the system shows invariant result and for GN=500 and HN=200 system shows best TSCA of 87.1%. On the other hand, result varies for Colon2 cancer due to parameter setting as it is seen in Table III. The reasons for such observation presume that Colon1 contains less number of GN and Colon2 contains less number of samples with respect to other datasets. Among the four datasets, mRMR-ANN achieved 100 % TSCA for Colon2 and MLL datasets. It is observed from the tables that mRMR-ANN performed relatively better for larger values of GN and HN in comparison of smaller values of those. As an example, for ALL/AML cancer (Table IV) average TSCA was 85.9% for GN=100 and HN=20; for the same problem average TSCA was 95.1% for GN=300 and HN=100. It is logical for worse TSCA with less GN because some genes may be missed by mRMR looking training data. Finally, proposed mRMR-ANN seems to be a suitable method for cancer classification showing good TSCA.

Table VI shows the numerical comparative results of the mRMR-ANN and other existing related methods. The result presented for the proposed method is the best TSCA values from Tables II to V. On the other hand, the results of other methods are the reported results in referred papers. An existing method tested on one or two cancer datasets and therefore others are marked as '-' meaning that results are not available. NPPC method achieved TSCA of 84.02% and 96.46% for Colon1 and ALL/AML datasets. For both the datasets, mRMR-ANN outperformed NPPC showing TSCA of 87.10% and 97.10%, respectively. ELM was also tested for ALL/AML dataset but achieved worst performance i.e., 93.10%. Recent method GECC tested for Colon2 dataset and achieved TSCA 97.22%; and TSVM tested for MLL dataset and achieved TSCA 88.80%. On the other hand, for both Colon2 and MLL datasets proposed mRMR-ANN achieved 100% classification accuracy. Therefore, proposed method is found relatively better than existing methods for cancer classification.

The experimental results presented in Tables II to V were for fixed number of training iteration of ANN with different GN and HN values. Therefore, it is interesting to observe effect of training iteration on the performance of proposed mRMR-ANN and Fig. 4 shows the graphical representation of TSCA with respect to iteration for the four datasets. Training iteration was varied from 10 to 1000 while number of mRMR selected genes and hidden neurons of ANN were fixed at 100. From the figure it is observed that TSCA was very low for less number of iteration and improved gradually. As an example, for Colon1 dataset, TSCA was 48.39% at 10 iteration and improved up to 83.87% at 100 iteration. It is notable from the figure that after 400 iteration no improvement has been observed for any dataset. It indicates the experimental results presented for 500 fixed iteration is appropriate.

TABLE. V. TEST SET CLASSIFICATION ACCURACY IN PERCENTAGE (%) FOR MLL CANCER

Num. of GN	Num. of HN	Best TSCA	Worst TSCA	Avg. TSCA	SD of Avg.
100	20	86.7	86.7	86.7	0
	50	86.7	86.7	86.7	0
	100	86.7	86.7	86.7	0
	200	86.7	86.7	86.7	0
200	20	86.7	86.7	86.7	0
	50	86.7	86.7	86.7	0
	100	93.3	93.3	93.3	0
	200	86.7	86.7	86.7	0
300	20	86.7	86.7	86.7	0
	50	86.7	86.7	86.7	0
	100	93.3	93.3	93.3	0
	200	86.7	86.7	86.7	0
400	20	86.7	86.7	86.7	0
	50	**100**	**100**	**100**	**0**
	100	86.7	86.7	86.7	0
	200	86.7	86.7	86.7	0
500	20	93.3	93.3	93.3	0
	50	93.3	93.3	93.3	0
	100	93.3	93.3	93.3	0
	200	93.3	93.3	93.3	0

TABLE. VI. CLASSIFICATION ACCURACY (TSCA IN %) COMPARISON OF MRMR-ANN WITH PROMINENT EXISTING METHODS

Dataset	NPPC [2]	ELM [6]	GECC [8]	TSVM [9]	Proposed mRMR-ANN
Colon1	84.02	-	-	-	87.10
Colon2	-	-	97.22	-	100
ALL/AML	96.46	93.10	-	-	97.10
MLL	-	-	-	88.80	100

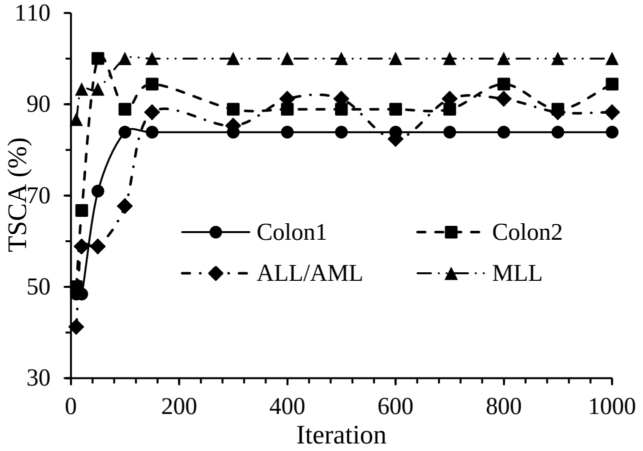

Fig. 4. Test Set Classification Accuracy (TSCA) vs. Iteration.

IV. Conclusions

Cancer treatment is much easier in the initial stage and DNA microarray based gene expression profiling has become an efficient technique for it. The gene expression is a very high dimensional data but relatively small number of genes are responsible for cancer; and therefore, classification looking on a subset of genes' expression (selecting through a suitable feature selection method) is a common way. In this study, information theoretic based mRMR has been considered for selecting cancer related genes and then ANN has been employed for classification. The proposed mRMR-ANN method first normalizes the gene expression data to employ mRMR and found effective to achieve better result. Although a few methods used mRMR in cancer classification, ANN with mRMR of this study has outperformed other methods while tested on several benchmark cancer datasets.

There are several future potential directions that follow from this study. Cancer classification is a sensitive task and its high accuracy is necessary. Proposed mRMR-ANN has shown to classify all the test samples correctly for two problems and it is remained an open challenge for other problems. This study considered maximum 500 genes in classification, more genes might give better outcome. On the other hand, the use of ensemble of ANNs instead of single ANN might be a good choice to improve classification performance.

References

[1] NCI, 2016. What Is Cancer? National Cancer Institute (NCI). Available: http://www.cancer.gov/about-cancer/understanding/what-is-cancer, Accessed: May 30, 2016.

[2] Ghorai, S., A. Mukherjee, S. Sengupta and P. K. Dutta, "Cancer Classification from gene expression data by NPPC Ensemble," Journal of IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8: 659-671, 2011.

[3] Xu, R., X. Cai, and D. Wunsch, "Gene Expression Data for DLBCL Cancer Survival Prediction with a Combination of Machine Learning Technologies," In Proceedings of the IEEE International Conference on Medicine and Biology, pp: 894-897, 2005.

[4] Takahashia, H., Y. Murasea, T. Kobayashid, and H. Hondaa, "New cancer diagnosis modeling using boosting and projective adaptive resonance theory with improved reliable index," Biochemical Engineering Journal, 33: 100–109, 2007.

[5] Acharya, S., S. Saha and Y. Thadisina, 2007. Multiobjective Simulated Annealing based Clustering of Tissue Samples for Cancer Diagnosis. IEEE Journal of Biomedical and Health Informatics, 20: 691-698.

[6] Arunkumar, C., and S. Ramakrishnan, "Binary Classification of Cancer Microarray Gene Expression Data using Extreme Learning Machines," In Proceedings of the 2014 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp: 1-4, 2014.

[7] Alshamlan, H., G. Badr and Y. Alohali1, "mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling," BioMed Research International, Article ID 604910, 15 pages, 2015. Doi:10.1155/2015/604910.

[8] Rathore, S., M. Hussain and A. Khan, "GECC: Gene Expression Based Ensemble Classification of Colon Samples," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(6): 1131 – 1145, 2014.

[9] Maulik, U., A. Mukhopadhyay and D. Chakraborty, "Gene-Expression-Based Cancer Subtypes Prediction through Feature Selection and Transductive SVM," IEEE Transaction on Biomedical Engineering, 60(4): 1111–1117, 2013.

[10] Zhang, C., J. Kai, H.C. Feng and T. Yang, "The Nearest Neighbor Algorithm of Filling Missing Data Based on Cluster Analysis," Applied Mechanics and Materials, 347-350: 2324-2328, 2013.

[11] Peng, H. and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, 27: 1226–1238, 2005.

[12] S. Haykin, Neural Networks—A Comprehensive Foundation. Prentice Hall, 2nd edition, 1999.

[13] Li, J., and H. Liu, "Kent Ridge Bio-medical Dataset" Available: http://datam.i2r.astar.edu.sg/datastes/krbd/, Accessed: 20 April, 2016.

[14] Notterman, D., U. Alon, A. J. Sierk and A. J. Levine, "Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays," The Journal Cancer Research, 61(7): 3124–3130, 2001.