International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:8, 2014

Performance Analysis of Genetic Algorithm with kNN and SVM for Feature Selection in Tumor Classification

C. Gunavathi, K. Premalatha

***Abstract*—**Tumor classification is a key area of research in the field of bioinformatics. Microarray technology is commonly used in the study of disease diagnosis using gene expression levels. The main drawback of gene expression data is that it contains thousands of genes and a very few samples. Feature selection methods are used to select the informative genes from the microarray. These methods considerably improve the classification accuracy. In the proposed method, Genetic Algorithm (GA) is used for effective feature selection. Informative genes are identified based on the T-Statistics, Signal-to-Noise Ratio (SNR) and F-Test values. The initial candidate solutions of GA are obtained from top-m informative genes. The classification accuracy of k-Nearest Neighbor (kNN) method is used as the fitness function for GA. In this work, kNN and Support Vector Machine (SVM) are used as the classifiers. The experimental results show that the proposed work is suitable for effective feature selection. With the help of the selected genes, GA-kNN method achieves 100% accuracy in 4 datasets and GA-SVM method achieves in 5 out of 10 datasets. The GA with kNN and SVM methods are demonstrated to be an accurate method for microarray based tumor classification.

***Keywords*—**F-Test, Gene Expression, Genetic Algorithm, k-Nearest-Neighbor, Microarray, Signal-to-Noise Ratio, Support Vector Machine, T-statistics, Tumor Classification.

I. Introduction

ABUNDANT methods and techniques have been proposed for tumor classification using microarray gene expression data. Rapid and modern advances in microarray gene expression technology have facilitated the simultaneous measurement of the expression levels of tens of thousands of genes in a single experiment at a rational cost. Gene expression profiling by microarray method has been came out as a capable technique for classification and diagnostic prediction of tumor.

The raw microarray data are images that are transformed into gene expression matrices. The rows in the matrix correspond to genes, and the columns represent samples or trial conditions. The number in each cell signifies the expression level of a particular gene in a particular sample or condition [1], [2]. Expression levels can be absolute or relative. If two rows are similar, it implies that the respective genes are co-regulated and possibly functionally related. By comparing samples, differentially expressed genes can be identified. The major limitation of the gene expression data is its high dimension which contains more number of genes and a very few samples. A number of gene selection methods have been introduced to select the informative genes for tumor prediction and diagnosis. Feature or Gene selection methods remove irrelevant and redundant features to improve classification accuracy.

A Simple Genetic Algorithm (SGA) is a computational concept of biological evolution that can be used to solve optimization problems [3]. The GA proposed by Holland, is a probabilistic optimal algorithm that is based on the evolutionary theories [4]. GA is based on a population of chromosomes. Successive populations of possible solutions are generated in a stochastic manner following laws similar to that of natural selection. The algorithm encodes a potential solution to a specific problem on a simple chromosome-like data structure and applies recombination operators to the structure so as to preserve significant information.

From the microarray data, the informative genes are identified based on their T-Statistics, Signal-to-Noise Ratio (SNR) and F-Test values. The initial candidate solutions of GA are obtained from top-m informative genes. The classification accuracy of k-Nearest Neighbor (kNN) method is used as the fitness function for GA.

C.Gunavathi is with the K. S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India (e-mail: sssguna@gmail.com).

K. Premalatha is with the Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India (e-mail: kpl_barath@yahoo.co.in).

II. Related Work

In this section the works related with Gene selection and tumor classification using microarray gene expression data are discussed. An evolutionary algorithm is used to identify the near-optimal set of predictive genes that classify the data [5]. Self-organizing map for clustering cancer data which composed of important gene selection step is used by [6]. Rough set concept with depended degrees is proposed by Wang and Gotoh in [7]. In this method they screened a small number of informative single gene and gene pairs on the basis of their depended degrees.

A Swarm Intelligence feature selection algorithm is proposed based on the initialization and update of only a subset of particles in the swarm [8]. Gene Doublets concept is introduced based on the gene pair combinations [9]. A new Ensemble Gene Selection method is applied to choose multiple gene subsets for classification purpose, where the significant degree of gene is measured by conditional mutual information or its normalized form [10].

A hybrid method, which consists of correlation-based feature selection and the Taguchi chaotic binary PSO, is used for important gene selection [11]. Hyper-Box Enclosure (HBE) method based on mixed integer programming for the classification of some cancer types with a minimal set of predictor genes is used by [12]. The use of single gene is

explored to construct classification model by Wang and Simon [13]. This method first identified the genes with the most powerful univariate class discrimination ability and constructed simple classification rules for class prediction using the single gene.

An efficient feature selection approach based on statistically defined effective range of features for every class termed as Effective Range based Gene Selection (ERGS) is proposed by [14]. BioMarker Identifier (BMI), which identified features with the ability to distinguish between two data groups of interest, is suggested in [15]. Margin Influence Analysis (MIA) is an approach designed to work with SVM for selecting informative genes [16]. A model for feature selection using Signal-to-Noise Ratio (SNR) ranking is proposed by [17].

Semi-Supervised Local Fisher discriminant (iSELF) analysis for gene expression data classification is introduced in [18]. A method that relaxed the maximum accuracy criterion to select the combination of attribute selection and classification algorithm is introduced by [19]. A comparative analysis of swarm intelligence techniques for feature selection in cancer classification is used in [20]. A feature selection algorithm which divides the genes into subsets to find the informative genes is proposed by [21].

III. Statistical Measures

Feature selection methods are used to rank the informative genes from the microarray data. The statistical methods used for feature selection in this research work are discussed here.

A. T-Statistics

Genes, who have considerably different expressions involving normal and tumor tissues, are candidates for selection. A simple T-statistic measure given in (1) is used to find the degree of gene expression difference, between normal and tumor tissues [22]. The top-m genes with the largest T-statistic are selected for inclusion in the discriminant analysis.

$$t = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{v1}{n1} + \frac{v2}{n2}}} \quad (1)$$

Here

$\overline{x1}$ - Mean of Normal samples
$\overline{x2}$ - Mean of Tumor samples
n1 - Normal Sample size
n2 - Tumor Sample size
$v1$ - variance of Normal samples
$v2$ - variance of Tumor samples

B. Signal-to-Noise ratio

SNR is the amount of biological signal relative to the amount of noise, which is a measure of the biological signal-to-noise ratio. An important measure used to find the significance of genes is the Pearson Correlation Co-efficient. According to Golub it is changed to emphasize the 'Signal-to-Noise Ratio' in using a gene as a predictor [1]. This predictor is shaped with the purpose of finding the Prediction Strength of a particular gene [23]. The Signal-to-Noise ratio PS of a gene 'g' is calculated by (2).

$$PS(g) = \frac{\overline{x1} - \overline{x2}}{s1 - s2} \quad (2)$$

Here

$\overline{x1}$ – Mean of Normal samples
$\overline{x2}$ - Mean of Tumor samples
$s1$ – Standard Deviation of Normal samples
$s2$ - Standard Deviation of Tumor samples

This value is used to reveal the difference between the classes relative to the standard deviation within the classes. Large values of PS (g) indicate a strong correlation between the gene expression and the class distinction, while the sign of PS (g) being positive or negative corresponds to *g* being more highly expressed in class 1 or class 2. Genes with large SNR value are informative and are selected for tumor classification.

C. F-Test

F-Test is the ratio of the variances of the given two set of values which is used to test if the standard deviations of two populations are equal or if the standard deviation from one population is less than that of another population. In this work two-tailed F-Test value is used to check the variances of normal Samples and tumor Samples. Formula to calculate the F-Test value of a gene is given in (3). Top-m genes with the smallest F-Test value are selected for inclusion in the further analysis.

$$F = \frac{v1}{v2} \quad (3)$$

Here

$v1$ - Variance of Normal Samples
$v2$ - Variance of Tumor Samples

IV. Genetic Algorithm

A Simple Genetic Algorithm (SGA) is a computational concept of biological evolution that can be used to solve some optimization problems [3]. These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to these structures so as to preserve significant information. An implementation of a genetic algorithm begins with a random population of chromosomes. One can evaluates these data structures and apply reproduction operators in such a way that the chromosomes which provide a better solution to the objective problem are given more chances to reproduce themselves than those chromosomes which gives poorer solutions. The goodness of a solution is typically defined with respect to the current population.

A. Simple Genetic Algorithm

Step 1. Generate random population of N Chromosomes. These Chromosomes are potential solution for the given problem.
Step 2. Assess the fitness function f(x) of each chromosome x in the random population.
Step 3. Create a new population of Chromosomes by repeating the following steps until the new population is complete
3.1 Select two parent chromosomes from the current population according to their fitness value (the better fitness, the bigger chance to be selected)
3.2 With a crossover probability crossover the parents to form a new offspring (children). If crossover operation was not done, offspring is an exact copy of parents.
3.3 With a mutation probability mutate new offspring at each locus (position in chromosome).
3.4 Place new offspring in a new population
Step 4. Use newly generated population for the further runs
Step 5. If the end condition is satisfied, stop the process and return the best solution in current population
Step 6. Go to step 2

B. GA Operators

1. Selection

Chromosomes are selected from the current population to be parents to crossover. Based on Darwin's evolution theory the best ones should survive and create new offspring. There are many methods to select the best chromosomes. They are Roulette Wheel selection, Boltzman selection, Tournament selection, Rank selection, Steady state selection and some other selection methods.

2. Crossover

Crossover is a genetic operator that combines (mates) two chromosomes (parents) to produce a new chromosome (offspring). The idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. Crossover occurs during evolution according to a user-definable crossover probability. Single point crossover, Two point crossover, Uniform crossover and Arithmetic crossover are the types of crossover method.

3. Mutation

After crossover is performed, mutation takes place. This is to prevent falling all solutions in the population into a local optimum of solved problem. Mutation changes randomly the new offspring. For binary encoding a few randomly chosen bits are changed from 1 to 0 or 0 to 1. Here the selected bits are inverted.

4. Evaluation

After producing offspring they must be inserted into the population. This is especially important, if less offspring are produced than the size of the original population. Another case is, when not all offspring are to be used at each generation or if more offspring are generated than needed. A reinsertion scheme determines which individuals should be inserted into the new population and which individuals of the population will be replaced by offspring. The used selection algorithm determines the reinsertion scheme. The elitist combined with fitness-based reinsertion prevents this losing of information and is the recommended method. At each generation, a given number of the least fit parent is replaced by the same number of the fit offspring.

V. Classification Algorithms

A. k-Nearest Neighbor Classifier

The kNN is an instance-based classifier which works on the assumption that classification of unknown instances can be identified by relating the unknown to the known instances according to some distance or similarity measure. The two instances far apart in the instance space defined by the appropriate distance function are less likely to belong to the same class than two closely situated instances.

The kNN algorithm does not abstract any information from the training data during the learning phase. The process of generalization is postponed until the time of classification. Classification using a kNN classifier is done by locating the nearest neighbor in instance space and labeling the unknown instance with the same class label as that of the known neighbor. This approach is often referred to as a nearest neighbor classifier. The high degree of local sensitivity makes nearest neighbor classifiers highly prone to noise in the training data. The robust models can be achieved by identifying k, where $k > 1$, neighbors and the majority vote decide the outcome of the class labeling. If $k=1$, then the object is simply assigned to the class of its nearest neighbor. A higher value of k results in a smoother, less locally sensitive function.

To find the closeness normally some distance measures are used. Sometimes one minus correlation value is also taken as a distance metric. For continuous variables the following three distance measures are used. They are Euclidean distance, Manhattan distance and Minkowski distance. In the instance of categorical variables the Hamming distance must be used. In this work Euclidean distance is used as the distance measure.

B. Support Vector Machine

Support Vector Machine, a technique derived from statistical learning theory, is used to classify data points by assigning them to one of two disjoint half spaces [24]. They are able to classify non-linear relationships in the data through the use of kernel functions specific to the datasets. SVM is widely used by many researchers in classification of cancer samples using microarray gene expression profiling. It uses a nonlinear mapping to transform the original training data into higher dimension. Within that new dimension it searches for the linear optimal separating hyperplane or a decision boundary separating the data of one class from another. The data from two different classes can always be separated by a

hyperplane. The SVM finds the hyperplane with the help of support vectors and margins. SVM methods are much less prone to over fitting of data.

VI. Feature Selection Based on GA

A. Chromosome Representation

The chromosome should contain the information about the solution which it represents. The most used way of encoding is a binary string, but the above optimization technique is well suited for continuous optimization problem. The random values are generated for gene position. The genes are considered when the value in its position is greater than 0.5, otherwise it is ignored. Fig. 1 shows the candidate solution representation.

g_1	g_2	g_3	g_4	…	g_{n-1}	g_m
0.25	0.56	0.12	0.98	---	0.43	0.112

Fig. 1 Chromosome representation

B. Fitness Function

The accuracy of kNN classifier is used as the fitness function for GA [25], [26]. The fitness function *fitness(x)* is defined as in (4).

$$fitness(x) = Accuracy(x) \tag{4}$$

Accuracy(x) is the test accuracy of testing data *x* in the kNN classifier which is built with the feature subset selection of training data. The classification accuracy of kNN is given by (5).

$$Accuracy(x) = (c/t) \times 100 \tag{5}$$

where

c - Samples that are classified correctly in test data by kNN technique

t - Total number of Samples in test data

C. k-fold Cross Validation

k-fold cross-validation is used for the result to be more valuable. In *k*-fold cross-validation, the original sample is divided into random *k*-subsamples; one among them is kept as the validation data for testing. The remaining *k*-1 sub-samples are used for training. The cross-validation process is repeated for *k*-times (the folds), with each of the *k* sub-samples used exactly once as the validation data. The average of *k* results from the folds gives the test accuracy of the algorithm. In order to achieve a reliable performance of the classifier, the 5-fold cross-validation method is used in this proposed method.

VII. Experimental Setup and Results Discussion

In order to assess the performance of the proposed method, ten datasets were analyzed. The datasets are collected from Kent Ridge Biomedical Data Repository. The details are given in Table I. In the columns (Class1 and Class2) of Table I, the number within the bracket denotes the number of samples. The Parameters and their values of GA are shown in Table II.

From the microarray data, informative genes are identified based on T-statistics, Signal-to-Noise Ratio and F-Test values. The initial candidate solutions of the GA are obtained from the top-m informative genes. The m values are taken as 10, 50 and 100. The selected genes are used for further classification. The classification accuracy of kNN is used as the fitness function for GA.

TABLE I
Cancer Microarray Gene Expression Datasets

Dataset Name	Number of Genes	Class1	Class2	Total Samples
CNS	7129	Survivors (21)	Failures (39)	60
DLBCL Harvard	7129	DLBCL (58)	FL (19)	77
DLBCL Outcome	7129	Cured (32)	Fatal (26)	58
Lung Cancer Michigan	7129	Tumor (86)	Normal (10)	96
Ovarian Cancer	15154	Normal (91)	Cancer (162)	253
Prostate Outcome	12600	Non-Relapse (13)	Relapse (8)	21
AML-ALL	7129	ALL (47)	AML (25)	72
Colon Tumor	2000	Tumor (40)	Healthy (22)	62
Lung Harvard2	12533	ADCA (150)	Mesothelioma (31)	181
Prostate	12600	Normal (59)	Tumor (77)	136

TABLE II
GA Parameters and Their Values

Parameter	Value
Chromosome size	10, 50 and 100
Population size	50
Maximum no. of Generations	200
Selection Method	Roulette Wheel
Selection rate	50 %
Crossover Type	Single point
Crossover Probability	0.6
Mutation Type	Flip bit
Mutation Probability	0.1
Distance Measure in kNN	Euclidean distance
k-value is kNN	5

The kNN with 5-fold cross validation method gives the classification accuracy as output. The GA is configured to have 50 individuals and was run for 200 generations in each trial. In conventional GA method, Crossover probability and Mutation probability is taken as 0.6 and 0.1. Figs. (2)-(7) show the results obtained from GA with kNN and SVM based methods. It gives the maximum average classification accuracy with minimum number of genes with top-m genes when applied different measures like T-statistics, Signal-to-Noise ratio and F-Test.

From the results it is inferred that the m value does not influence the accuracy of the classifier. So the value of m should be identified through empirical analysis. The experimental results show that the GA-kNN method gives 100% average accuracy for DLBCL Harvard, Lung cancer Michigan, Ovarian Cancer and Lung Harvard2 datasets and GA-SVM method gives 100% average accuracy for DLBCL

Harvard, Lung cancer Michigan, Ovarian Cancer, AML-ALL and Lung Harvard2 datasets. For most of the datasets examined, the classification accuracies obtained by the proposed feature selection method outperformed or matched with existing methods.

Tables III to XII compare the results obtained by the proposed method with other existing methods. For the datasets other than CNS, Prostate outcome and Colon Tumor the proposed method gives better accuracy.

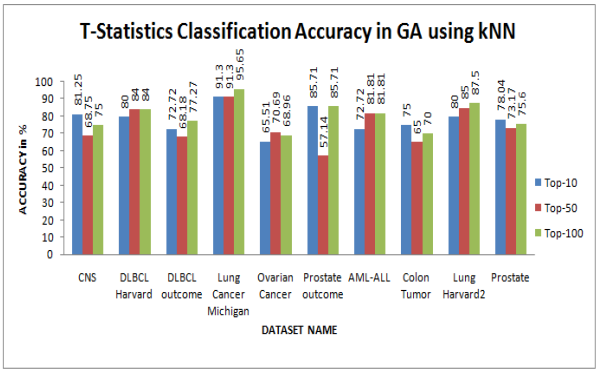

Fig. 2 Classification Accuracy of T-Statistics-GA-kNN method

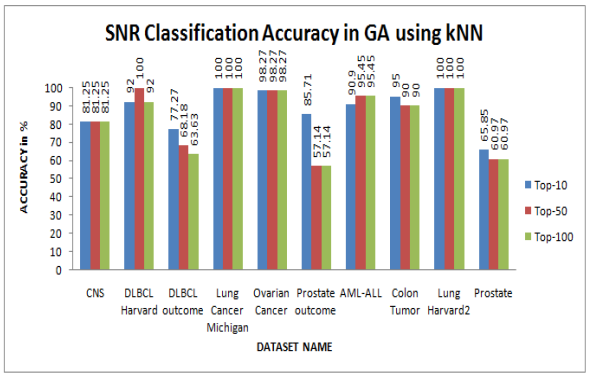

Fig. 3 Classification Accuracy of SNR-GA-kNN method

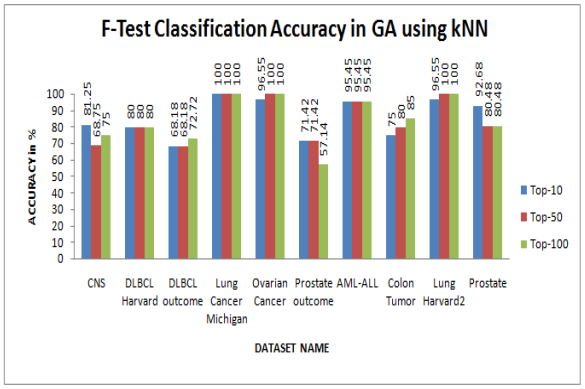

Fig. 4 Classification Accuracy of F-Test-GA-kNN method

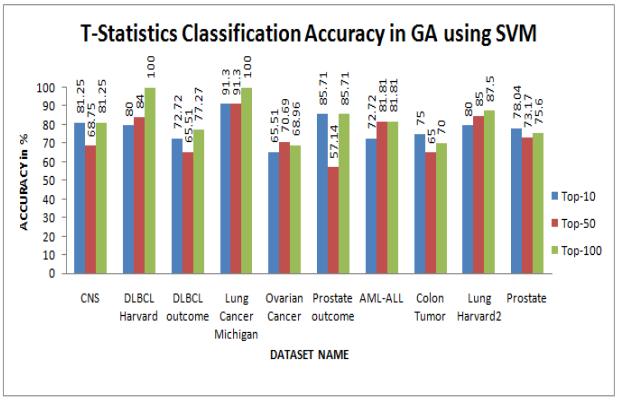

Fig. 5 Classification Accuracy of T-Statistics-GA-SVM method

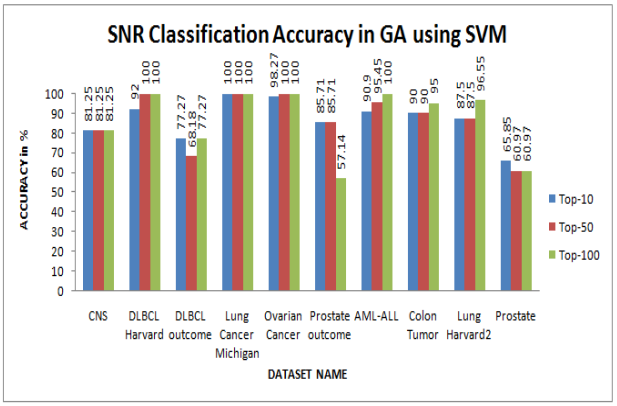

Fig. 6 Classification Accuracy of SNR-GA-SVM method

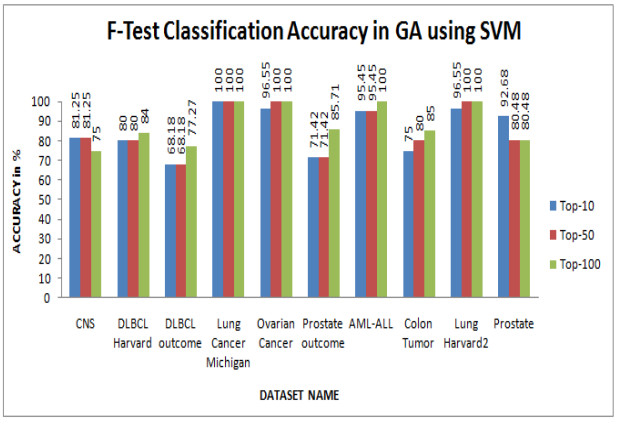

Fig. 7 Classification Accuracy of F-Test-GA-SVM method

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR CNS

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	81.25
This work	GA + kNN	81.25
Alonso et al. [19]	Combination of attribute selection and classification algorithm	75.49
Liu et al. [10]	EGS - Ensemble Gene Selection Method	98.33

TABLE IV
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR DLBCL HARVARD

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	100
This work	GA + kNN	100
Huang et al. [18]	iSELF- improved semi-supervised local fisher discriminant analysis	94.67
Alonso et al. [19]	Combination of attribute selection and classification algorithm	100
Dagliyan et al. [12]	HBE - Hyper-Box Enclosure Method	96.1
Chopra et al. [9]	Based on Gene doublets	98.1
Martinez et al. [8]	Swarm intelligence feature selection algorithm	100

TABLE V
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR DLBCL OUTCOME

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	77.27
This work	GA + kNN	77.27
Alonso et al. [19]	Combination of attribute selection and classification algorithm	67.84
Wang and Simon [13]	Univariate class discrimination with single gene	74

TABLE VI
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR LUNG CANCER MICHIGAN

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	100
This work	GA + kNN	100
Alonso et al. [19]	Combination of attribute selection and classification algorithm	100
Liu et al. [10]	EGS - Ensemble Gene Selection Method	89.58

TABLE VII
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR OVARIAN CANCER

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	100
This work	GA + kNN	100
Alonso et al. [19]	Combination of attribute selection and classification algorithm	100

TABLE VIII
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR PROSTATE OUTCOME

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	85.71
This work	GA + kNN	85.71
Dagliyan et al. [12]	HBE - Hyper-Box Enclosure Method	95.24

TABLE IX
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR AML-ALL

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	100
This work	GA + kNN	95.45
Alonso et al. [19]	Combination of attribute selection and classification algorithm	100
Chandra and Gupta [14]	Effective Range based Gene Selection	98.61
Dagliyan et al. [12]	HBE - Hyper-Box Enclosure Method	100
Martinez et al. [8]	Swarm intelligence feature selection algorithm	100
Liu et al. [10]	EGS - Ensemble Gene Selection Method	100
Chopra et al. [9]	Based on Gene doublets	100
Wang and Gotoh [7]	Rough sets	100
Vanichayobon et al. [6]	Gene selection step and clustering cancer data by using self-organizing map	100
Umpai and Stuart [5]	Evolutionary algorithm	98.24

TABLE X
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR COLON TUMOR

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	95
This work	GA + kNN	95
Alonso et al. [19]	Combination of attribute selection and classification algorithm	88.41
Chandra and Gupta [14]	Effective Range based Gene Selection	83.87
Li et al. [16]	Margin Influence Analysis with SVM	100
Chopra et al. [9]	Based on Gene doublets	91.1

TABLE XI
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR LUNG HARVARD2

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	100
This work	GA + kNN	100
Alonso et al. [19]	Combination of attribute selection and classification algorithm	99.63
Chandra and Gupta [14]	Effective Range based Gene Selection	100
Wang and Simon [13]	Univariate class discrimination with single gene	99
Chopra et al. [9]	Based on Gene doublets	100
Wang and Gotoh [7]	Rough sets	97.32
Vanichayobon et al. [6]	Gene selection step and clustering cancer data by using self-organizing map	100

TABLE XII
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR PROSTATE

Reference	Methodology	Average Classification Accuracy in percentage
This work	GA + SVM	92.68
This work	GA + kNN	92.68
Wang and Gotoh [7]	Rough sets	91.18

VIII. CONCLUSION

Tumor classification using gene expression data is an important task for addressing the problem of tumor prediction and diagnosis. For an effective and precise classification, investigations of feature selection methods are essential. T-statistics, Signal-to-Noise Ratio and F-Test are the feature selection methods used to select the important genes. Genetic Algorithm with kNN and SVM Classifier method are applied on the top-m genes in this research work. Here the classification accuracy of kNN is considered as the fitness function for GA. The kNN classifier is one of the most famous neighborhood classifier in pattern recognition. SVM is commonly used in the domain of cancer studies, protein identification and especially in Microarray data. Here 5-fold cross-validation is applied to avoid the over fitting of the data. The performance of this hybrid method is tested with ten different cancer datasets. Conventional GA method gives 100% classification accuracy for 4 datasets with kNN and 5 datasets out of 10 datasets with SVM. These simple models based on statistical measures and Genetic Algorithm performs two level of feature selection to get the most informative genes for classification process. This method can be successfully applied to the gene expression data of any type of cancer, because it was successfully demonstrated with ten different Cancer Datasets in this research work.

REFERENCES

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, 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.

[2] E. Domany, "Cluster analysis of gene expression data," *J Stat Phys*, vol. 110, pp. 1117-1139, 2003.

[3] D.E. Goldberg, *Genetic Algorithms-in Search, Optimization and Machine Learning*. London: Addison-Wesley Publishing Company Inc, 1989.

[4] J. Holland, *Adaption in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, 1975.

[5] T. Umpai and A. Stuart, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes," *BMC Bioinformatics*, vol. 6, no. 148, 2005.

[6] S. Vanichayobon, W. Siriphan, and W. Wiphada, "Microarray Gene Selection Using Self-Organizing Map," in *Proceedings of the seventh WSEAS International Conference on Simulation, Modelling and Optimization*, Beijing, China, 2007.

[7] X. Wang and O. Gotoh, "Accurate molecular classification of cancer using simple rules," *BMC Medical Genomics*, vol. 2, no. 64, 2009.

[8] E. Martinez, M.A. Mario, and T. Victor, "Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm," *Computational Biology and Chemistry*, vol. 34, pp. 244 – 250, 2010.

[9] P. Chopra, J. Lee, J. Kang, and S. Lee, "Improving Cancer Classification Accuracy Using Gene Pairs," *PLoS ONE*, vol. 5, no. 12, 2010.

[10] H. Liu, L. Lei, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recognition*, vol. 43, pp. 2763 – 2772, 2010.

[11] C. Li-Yeh, Y. Cheng-San, W. Kuo-Chuan, and Y. Cheng-Hong, "Gene selection and classification using Taguchi chaotic binary particle swarm optimization," *Expert Systems with Applications*, vol. 38, pp. 13367 – 13377, 2011.

[12] O. Dagliyan, F. Uney-Yuksektepe, I.H. Kavakli, and M. Turkay, "Optimization Based Tumor Classification from Microarray Gene Expression Data," *PLoS ONE*, vol. 6, no. 2, 2011.

[13] X. Wang and R. Simon," Microarray-based cancer prediction using single Genes," *BMC Bioinformatics*, vol. 12, no. 391, 2011.

[14] B. Chandra and M. Gupta, "An efficient statistical feature selection for classification of gene expression data," *Journal of Biomedical Informatics*, vol. 44, pp. 529 – 535, 2011.

[15] I.H. Lee, H.L. Gerald, and V. Mahesh, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer," *Journal of Clinical Bioinformatics*, vol. 1, no. 11, 2011.
[16] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, B.B. Tan, B.C. Deng, and C.C. Lin, "Recipe for Uncovering Predictive Genes Using Support Vector Machines Based on Model Population Analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1633 – 1641, 2011.
[17] D. Mishra and B. Sahu, "Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach," *International Journal of Scientific & Engineering Research*, vol. 2, no. 4, 2011.
[18] H. Huang, J. Li, and J. Liu, "Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis," *Expert Systems with Applications*, vol. 39, pp. 2314 – 2320, 2012.
[19] G.C.J. Alonso, I.Q. Moro-Sancho, A. Simon-Hurtado, and R. Varela-Arrabal, "Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods," *Expert Systems with Applications*, vol. 39, pp. 7270 –7280, 2012.
[20] C.Gunavathi and K.Premalatha, "A Comparative Analysis of Swarm Intelligence Techniques for Feature Selection in Cancer Classification" *The Scientific World Journal*, vol. 2014, Article ID 693831, http://dx.doi.org/10.1155/2014/693831.
[21] A. Sharma, I. Seiya, and M. Satoru, "A Top-R Feature Selection Algorithm For Microarray Gene Expression Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 3, pp. 754 – 764, 2012.
[22] K. Yendrapalli, R. Basnet, S. Mukkamala, and A.H. Sung, "Gene Selection for Tumor Classification Using Microarray Gene Expression Data," in *Proceedings of the World Congress on Engineering*, vol. I, 2007.
[23] X. Momiao, W. Li, J. Zhao, J. Li, and B. Eric, "Feature (Gene) Selection in Gene Expression-Based Tumor Classification," *Journal of Molecular Genetics and Metabolism*, vol. 73, pp. 239–247, 2001.
[24] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp.273–297, 1995.
[25] M.S. Mohamed, D. Safaai, and R.O. Muhammad, "Genetic Algorithms wrapper approach to select informative genes for gene expression microarray classification using support vector machines," in *InCoB'04: Proceedings of Third International Conference on Bioinformatics*, Auckland, New Zealand, 2004.
[26] N.S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175-185, 1992.