Received: March 20, 2017


International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# Biomarker Selection from Gene Expression Data for Tumour Categorization Using Bat Algorithm

**Gunavathi Chellamuthu[1]*, Premalatha Kandasamy[2], Sivasubramanian Kanagaraj[3]**

[1]*School of Information Technology and Engineering, Vellore Institute of Technology University, Vellore, India*
[2]*Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India*
[3]*Department of Electronics and Communication Engineering, KS Rangasamy College of Technology, Tiruchengode, India*
* Corresponding author's Email: sssguna@gmail.com

**Abstract:** Microarray technology is commonly used in the study of disease diagnosis using gene expression levels. The classification of cancer is a foremost area of research in the field of bioinformatics. Microarray technology enables the researcher to investigate the expression levels of thousands of genes in a single experiment and gives the measurements of their differential expression. The main drawback of gene expression data is that it contains thousands of genes and a very few samples. Feature or gene selection methods are used to successfully extract the genes that directly involved in the classification and to eliminate irrelevant genes. These methods considerably improve the classification accuracy. The proposed method applies bat algorithm (BA) for feature selection in tumour classification. Initially, the top-10 genes are selected by T-Statistics, signal-to-noise ratio (SNR) and F-Test. The classifier accuracy of k-nearest neighbour (kNN) technique is used as the fitness function for BA. The simulated results are demonstrated and analyzed with 10 different cancer gene expression dataset. For Lung Cancer Michigan and Lung Harvard2 datasets the proposed method gives 100% classification accuracy with minimum number of genes. For 5 other datasets, the proposed method gives more than 90% of classification accuracy. The results show the suitability of the proposed algorithm for feature selection in cancer classification.

**Keywords:** Microarray gene expression, Tumour categorization, Feature selection, Bat algorithm, K-nearest-neighbour.

## 1. Introduction

Cancer is featured by an irregular, uncontrollable growth that may destroy and attack neighbouring healthy body tissues or somewhere else in the body. Gene expression profiling by microarray method has been emerged as an efficient technique for classification and diagnostic prediction of cancer. Cancer classification refers to the process of constructing a model on the microarray dataset and then distinguishing one type of samples from other types with this induced model.

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 experimental conditions. The numbers in each cell denotes the expression level of particular gene in a particular sample or condition [1, 2]. Expression levels can be absolute or relative. They are used to simultaneously monitor and study the expression levels of thousands of genes, relationship between genes, their functions and classifying genes or samples. 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 in gene expression data is its high dimensionality. It contains more number of genes and a very few samples. Feature or gene selection methods are needed to find the important genes that are reason for cancer. Feature selection

methods remove irrelevant and redundant features to improve classification accuracy. A number of gene selection methods have been introduced to select informative genes for cancer prediction and diagnosis. Most commonly used gene selection methods are Relief-F, Minimal-Redundancy-Maximal Relevance (MRMR), T-statistic, Information Gain and Chi-square statistic [3]. Feature selection methods can be categorized into filter, wrapper, and embedded or hybrid [4]. T-statistics, Signal-to-Noise Ratio and F-Test are the feature selection measures used in the proposed work to find the top-10 significant or informative genes.

Optimization is the act of achieving the best possible result under given conditions. The objective of an optimization algorithm is to minimize or maximize the objective function. Bat algorithm (BA) is a novel meta-heuristic optimization algorithm based on the echolocation behavior of microbats with varying pulse rate of emission and loudness. BA is very simple to understand. It has only few parameters to adjust. Its convergence speed is fast, and it is easy to implement.

The proposed approach is a hybrid system that uses the BA for feature selection to classify the given samples and the fitness function of BA is measured by the kNN technique. This simple model based on statistical measures and optimization technique performs two level of feature selection to get the most informative genes for the classification process. The paper is organized as follows: Section 2 describes about gene selection methods such as T-statistics, Signal-to-Noise Ratio and F-Test. Section 3 explains about k-Nearest Neighbour Classification algorithm. Section 4 gives the details about BA. Section 5 explains about tumour categorization with BA. Section 6 presents the experimental results obtained from the proposed method.

## 2. Gene selection methods

### 2.1 T-statistics

Genes who have significantly different expressions between normal and tumour tissues or between subtypes of tumour tissues are also candidates for selection. A simple T-statistic can be used to measure the degree of gene expression difference between normal and tumour tissues [5]. The top-10 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 Tumour samples
$n1$ - Normal Sample size
$n2$ - Tumour Sample size
$v1$ - variance of Normal samples
$v2$ - variance of Tumour samples

### 2.2 Signal-to-noise ratio

A significant measure used in finding the importance of genes is the Pearson Correlation Coefficient. It is modified as follows to emphasize the 'Signal-to-Noise Ratio' in using a gene as a predictor [1]. This predictor is created with the purpose of finding the Prediction Strength of a particular Gene [6]. The Signal-to-Noise ratio *PS* of a gene *'g'* is defined as

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

Here

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

This value is used to reflect 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 tumour classification. Top-10 genes with the largest SNR value are selected for inclusion in the discriminant analysis.

### 2.3 F-Test

F-Test is generally defined as the ratio of the variances of the given two set of values. The F-test 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. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the standard deviations are not equal. The one-tailed version only tests in one direction that is the standard deviation from the first population is either greater than or less than (but not both) the

second population standard deviation. Top-10 genes with the smallest F-Test value are selected for inclusion in the discriminant analysis.

$$F = \frac{v1}{v2} \tag{3}$$

Here

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

## 3. K-nearest neighbour algorithm

The k-Nearest Neighbour algorithm is one of the simplest of all machine learning algorithms. It is one of the Lazy learners in which the learner waits until the last moment before constructing any model for the purpose of classifying a given test tuple. When given a training sample, a lazy learner simply stores it and waits until it is given a test tuple. It is a method for classifying objects based on closest training examples in the feature space. Here a sample is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k-Nearest Neighbours (k is a positive integer, typically small) measured by a distance function. If k = 1, then the object is simply assigned to the class of its nearest.

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 this work Euclidean distance between two samples is used as the distance measure.

## 4. Bat algorithm

Bat Algorithm (BA) is a novel meta-heuristic optimization algorithm firstly proposed in Xin She Yang in 2010 [7]. Microbats are insectivores. Bats use echolocation to locate and catch their prey. Bat echolocation is a perceptual system where ultrasonic sounds are emitted specifically to produce echoes. By using the time delay between the outgoing pulse and the returning echoes the brain and auditory nervous system of the bat produces a detailed three dimensional image of the surroundings. From this, bats can detect, localize and even classify their prey in complete darkness. When bats fly, they produce a constant stream of high-pitched sounds that can be heard only by them. When the sound waves produced by these bats hit an insect or other animal, the echoes bounce back to the bats, and guide them to the source [8]. Their pulses vary in properties and can be correlated with their hunting strategies, depending on the species.

The rules for Bat algorithm are:

1. All bats utilize echolocation to sense distance. They can also distinguish the difference between food/prey and background barriers in some supernatural way
2. Bats fly randomly with velocity $v_i$ at position $X_i$ with a fixed frequency $f_{min}$, varying wavelength $\lambda$ and loudness $A_0$ to search for prey. They can automatically fine-tune the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission $R \in [0, 1]$, based on the proximity of their target
3. The loudness can vary in many ways. Here, it is assumed that the loudness varies from a large (positive) value $A_0$ to a minimum constant value $A_{min}$.

In the implementations, virtual bats are used naturally. Consider, in an n-dimensional search space the positions $X_i$ and velocities $v_i$ of the bats are to be updated. The pulse frequency $f_i$ is calculated by using Eq. (4). At time $t$, the new solutions $X_i(t)$ and velocities $v_i(t)$ are calculated by using Eqs. (5) and (6).

$$f_i = f_{min} + (f_{max} - f_{min}) \times S_1 \tag{4}$$

$$X_i(t) = X_i(t-1) + v_i(t) \tag{5}$$

$$v_i(t) = v_i(t-1) + (X_i(t-1) - X^*) f_i \tag{6}$$

Here, $S_1$ is a random number, $S_1 \in [0,1]$ which is drawn from a uniform distribution. The frequency $f$ ranges from $f_{min}$ to $f_{max}$. Initially all the bats are randomly assigned with a frequency that is drawn uniformly from $[f_{min}, f_{max}]$. In this work, the value of $f_{min}$ is 0 and $f_{max}$ is 2. $X^*$ is the current global best location among all the n bats/solutions.

For the local search, once a solution $X_i$ is chosen from the current best solutions, a new solution $X_{i+1}$ for each bat is generated locally using random walk.

$$X_{i+1} = X_i + \eta A^t \tag{7}$$

Here, $\eta$ is a random number between $[-1, 1]$ and $A^t$ is the average loudness of all bats at time step $t$.

As the iterations proceed, the loudness $A_i$ and rate of pulse emission $R_i$ have to be updated. The bats adjust their pulse emission rate and loudness depending on the closeness of the prey/target. Normally the loudness decreases once a bat has found its prey. The rate of pulse emission $R_i$ increases as iteration increases. The loudness and rate of pulse

emission are updated according to the Equations (8) and (9).

$$A_i(t+1) = A_i(t).S_2 \quad (8)$$

$$R_i(t+1) = R_i(0)[1-\exp(-\gamma t)] \quad (9)$$

Here, $S_2$ and $\gamma$ are constants, given that $0 < S_2 < 1$ and $\gamma > 0$. Also $A_i(t) \rightarrow 0$, $R_i(t) \rightarrow R_i(0)$ as $t \rightarrow \infty$. The BA proposed by Yang in 2010 is shown in Fig. 1.

**Algorithm: BA**
Initialize the bat population and their velocity
Define pulse frequency $f_i$ at $X_i$
Initialize pulse rates $R_i$ and the loudness $A_i$
while (t <max number of iterations)
  Generate new solutions by adjusting frequency, and updating
  velocities and positions (using equations 4, 5 & 6)
  if (rand > $R_i$)
    Select a solution among the best solution
    Generate a local solution around the selected best solution
    (using equation 7)
  end if
  Generate a new solution by flying randomly
  if (rand < $A_i$ & $f(X_i) < f(X^*)$)
    Accept new solutions
    Reduce $A_i$ and increase $R_i$ (using 8 & 9)
  end if
  Rank the bats and find the current best $X^*$
end while

Figure.1 Pseudo code of the BA

## 5. Cancer classification using BA

The proposed approach is based on BA with kNN on the selected genes (individuals).

### 5.1 Bat representation

The bat should contain information about the solution which it represents. The most used way of encoding is a binary string. In the bat representation binary code '1' or '0' is used to mark whether a gene is selected or not. So each bat in the population is encoded by a string like '0101010101'. Finally, the gene subsets are obtained by choosing the genes that are marked by '1'.

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

Figure.2 Bat representation

### 5.2 Fitness function

The fitness function *f(x)* of a bat is measured by kNN technique [9]. The accuracy of kNN classifier is used as the fitness function. The fitness function *f(x)* is defined as

$$Fitness(x) = Accuracy(x) \quad (10)$$

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

$$Accuracy(x) = (c / t) \times 100 \quad (11)$$

Here
*c* - Samples that are classified correctly in test data by kNN technique
*t* - Total number of Samples in test data

## 6. Experimental results

The proposed method uses T-statistics, Signal-to-Noise Ratio and F-Test to select top-10 genes. These genes alone used for further classification. Bat algorithm is applied on the selected genes. For classification purpose the given dataset is divided into training and test samples. Initially the system is trained with training samples. Then the proposed method is tested on test samples. The classification accuracy of kNN is used as a fitness function for BA. The kNN with 5-fold cross validation method gives the classification accuracy as output. The BA was configured to have 10 bats and was run for 100 iterations in each trial. The parameter pulse rate of BA is considered as 0.5 and the loudness value is taken as 0.25 [7]. The values of S1 and S2 are the random numbers in the range 0 to 1. The value of $\gamma$ is a constant.

In order to assess the performance of the proposed method, 10 datasets were analyzed. These datasets were collected from Kent Ridge Biomedical Data Repository [10]. The details about the datasets are given in Table1. Table 2 gives the Parameters and their values used in this method.

Table1. Details of Datasets used in this method

| 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 | Tumour (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 Tumour | 2000 | Tumour (40) | Healthy (22) | 62 |
| Lung Harvard2 | 12533 | ADCA (150) | Mesothelioma (31) | 181 |
| Prostate | 12600 | Normal (59) | Tumour (77) | 136 |

Table 2. Parameters and values

| **Parameter** | **Value** |
|---|---|
| Loudness ($A$) | 0.25 |
| Pulse rate ($R$) | 0.5 |
| $\gamma$ | 0.5 |
| $S_1$ | rand(0,1) |
| $S_2$ | rand(0,1) |
| Number of bats | 10 |
| Number of iterations | 100 |
| Distance Measure in kNN | Euclidean distance |
| k-value is kNN | 3 |

Table 3. Experimental results

| S.No. | Dataset | T-statistics | | SNR | | F-Test | |
|---|---|---|---|---|---|---|---|
| | | Number of gene(s) | Accuracy | Number of gene(s) | Accuracy | Number of gene(s) | Accuracy |
| 1 | CNS | 1 | 81.25% | 3 | 81.25 % | 3 | 81.25 % |
| 2 | DLBCL Harvard | 3 | 78% | 6 | 92% | 5 | 80% |
| 3 | DLBCL Outcome | 3 | 72.72 % | 1 | 77.27 % | 3 | 68.18 % |
| 4 | Lung Cancer Michigan | 6 | 91.3 % | 4 | 100 % | 2 | 100 % |
| 5 | Ovarian Cancer | 4 | 65.51 % | 1 | 98.27 % | 5 | 96.55 % |
| 6 | Prostate Outcome | 2 | 85.71 % | 1 | 85.71 % | 1 | 71.42 % |
| 7 | AML-ALL | 3 | 72.72 % | 5 | 90.9 % | 2 | 95.45 % |
| 8 | Colon Tumour | 3 | 75 % | 5 | 95 % | 1 | 75 % |
| 9 | Lung Harvard2 | 4 | 80 % | 4 | 100 % | 4 | 96.55 % |
| 10 | Prostate | 1 | 78.04 % | 4 | 65.85 % | 3 | 92.68 % |

Table 4. Maximum accuracy with minimum genes

| S.No. | Dataset Name | Maximum Accuracy (with Minimum Genes) | | Gene selection method |
|---|---|---|---|---|
| | | Number of gene(s) | Accuracy | |
| 1 | CNS | 1 | 81.25 % | T-statistics |
| 2 | DLBCL Harvard | 6 | 92 % | SNR |
| 3 | DLBCL outcome | 1 | 77.27 % | SNR |
| 4 | Lung Cancer Michigan | 2 | 100 % | F-Test |
| 5 | Ovarian Cancer | 1 | 98.27 % | SNR |
| 6 | Prostate outcome | 1 | 85.71 % | SNR |
| 7 | AML-ALL | 2 | 95.45 % | F-Test |
| 8 | Colon Tumour | 5 | 95 % | SNR |
| 9 | Lung Harvard2 | 4 | 100 % | SNR |
| 10 | Prostate | 3 | 92.68 % | F-Test |

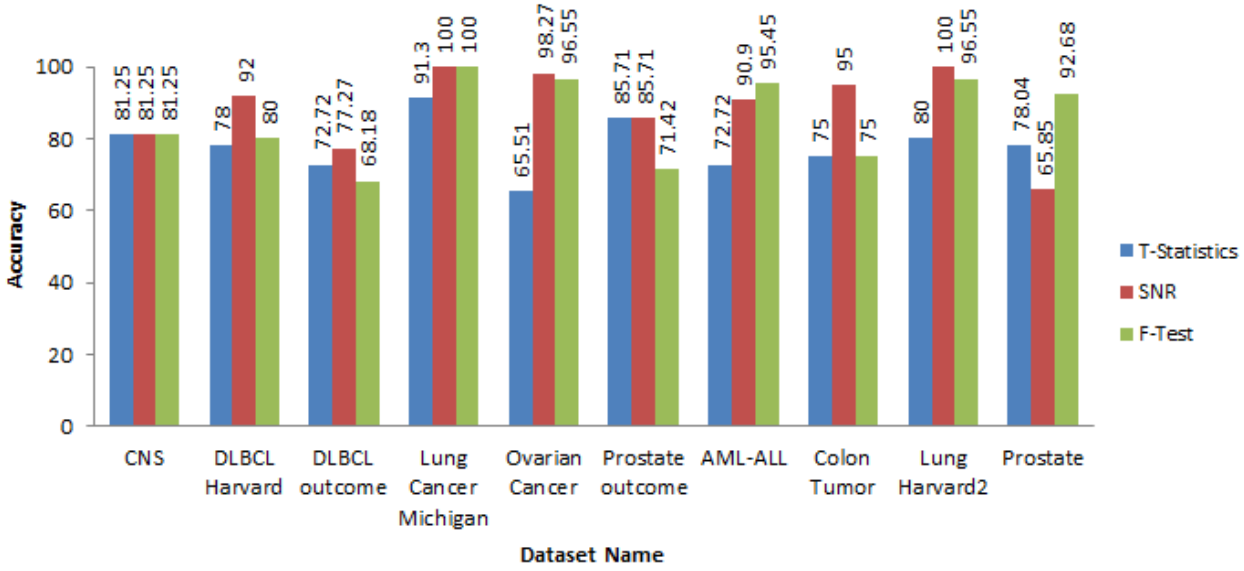


Figure.3 Classification accuracy

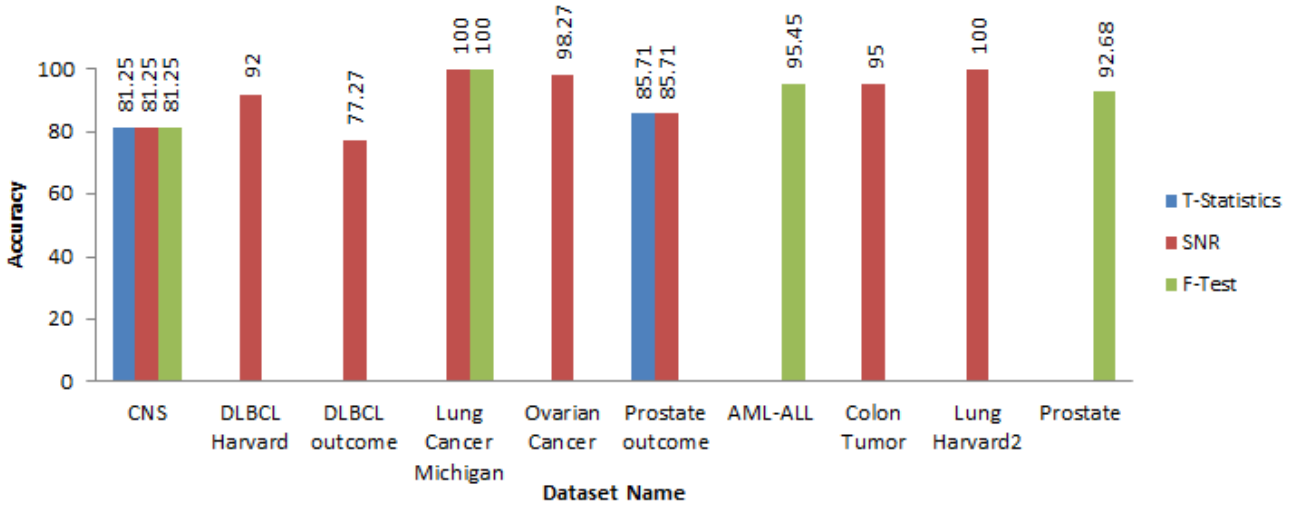


Figure.4 Maximum accuracy

Table 5. Comparison with other methods

| Reference / Dataset | CNS | DLBCL Harvard | DLBCL outcome | Lung Cancer Michigan | Ovarian Cancer | Prostate outcome | AML-ALL | Colon Tumor | Lung Harvard2 | Prostate |
|---|---|---|---|---|---|---|---|---|---|---|
| [11] | - | - | - | - | - | - | - | 90.68 | - | - |
| [12] | 75.49 | 100 | 67.84 | 100 | 100 | - | 100 | 88.41 | 99.63 | - |
| [13] | - | - | - | - | - | - | 100 | 100 | - | - |
| [14] | - | - | 74.00 | - | - | - | - | - | 99.00 | - |
| [3] | - | - | - | - | - | - | 98.61 | 83.87 | 100 | - |
| [15] | 98.33 | - | - | 89.58 | - | - | 100 | - | - | - |
| BA | 81.25 | 92 | 77.27 | 100 | 98.27 | 85.71 | 95.45 | 95 | 100 | 92.68 |

Table 3 shows the results obtained from the proposed method. It gives the Classification accuracy with minimum number of genes with top-10 genes when applied different measures like Signal-to-Noise ratio, T-statistics and F-Test. Table 4 represents the corresponding measure which gives the maximum accuracy with minimum number of genes among top-10 genes.

Figure 3 depicts the classification accuracy obtained from different measures when applying BA and kNN on top-10 genes. Figure 4 depicts the maximum accuracy obtained for different cancer types when applying BA and kNN on top-10 genes.

Table 5 displays the results obtained from BA based feature selection method for each of the dataset and the results are compared with other existing methods in the literature. BA has a capability of automatically zooming into an area where favourable solutions have been found. This zooming is supplemented by the automatic switch from explorative moves to local intensive exploitation. BA has guaranteed global convergence properties under the right condition, and it can also solve large-scale problems effectively. From the results it is observed that the performance of the BA based feature selection method is comparable with other works. It gives maximum classification accuracy with minimum number of genes.

## 7. Conclusion

Cancer classification using gene expression data is an important task for addressing the problem of cancer diagnosis and drug discovery. T-statistics, Signal-to-Noise Ratio and F-Test are the feature selection methods used to select the important genes. Bat Algorithm with kNN Classifier method is applied on those top genes in this research work. Here the classification accuracy of kNN is considered as the fitness function for the BA. The kNN classifier is one of the most famous neighbourhood classifier in pattern recognition. The kNN with 5-fold cross-validation is applied to avoid the over fitting of the data. The performance of hybrid method is tested with ten different cancer datasets. For Lung Cancer Michigan and Lung Harvard2 datasets the proposed method gives 100% classification accuracy with minimum number of genes. For DLBCL Harvard, Ovarian Cancer, AML-ALL, Colon Tumour and Prostate datasets, the proposed method gives more than 90% of classification accuracy. The results prove that only the informative gene selection leads to improve the classification accuracy. The above method can be 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. In this proposed work, only binary-class cancer gene expression datasets are considered.

Further research may focus on datasets with multiple-class labels. Other statistical measures such as information gain, chi-square test can also be considered for gene ranking. Hybrid approaches of optimization may be implemented with an improved solution which can be suggested to avoid premature convergence.

## 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", *Journal of Statistical Physics*, Vol.110, No.3-6, pp.1117-1139, 2003.
[3] B. Chandra and M. Gupta, "An efficient statistical feature selection approach for classification of gene expression data", *Journal of Biomedical Informatics*, Vol.44, No.4, pp.529–535, 2011.
[4] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, Vol.23, No.19, pp. 2507-2517, 2007.
[5] K. Yendrapalli, R. Basnet, S. Mukkamala, and AH. Sung, "Gene Selection for Tumor Classification Using Microarray Gene Expression Data", In: *Proc. of the World Congress on Engineering*, London, UK, Vol.I, pp.290-295, 2007.
[6] M. Xiong, W. Li, J. Zhao, L. Jin, and E. Boerwinkle, "Feature (Gene) Selection in Gene Expression-Based Tumor Classification", *Journal of Molecular Genetics and Metabolism*, Vol.73, No.3, pp.239–247, 2001.
[7] X.S. Yang, "A New Metaheuristic Bat-Inspired Algorithm, Nature Inspired Cooperative Strategies for Optimization", Eds. J. R. Gonzalez et al, *Studies in Computational Intelligence*, Springer Berlin, Springer, Vol.284, pp.65-74, 2010.
[8] P. Lauber, *Bats: Wings in the Night*, Random House, New York, 1968.
[9] MS. Mohamed, S. Deris, and M.R. Othman, "Genetic Algorithms wrapper approach to select informative genes for gene expression microarray classification using support vector machines", In: *Proc. of Third International Conf. on Bioinformatics*, Auckland, New Zealand, 2004.
[10] Kent Ridge Biomedical Data Repository 2002, Available from: <http://datam.i2r.a-star.edu.sg/datasets/krbd/> [15 February 2013].
[11] J.M. Arevalillo and H. Navarro, "Exploring correlations in gene expression microarray data for maximum predictive-minimum redundancy biomarker selection and classification", *Computers in Biology and Medicine*, Vol.43, No.10, pp.1437-1443, 2013.
[12] G.C.J. Alonso, I.Q.M. Sancho, A.S. Hurtado, and R.V. Arrabal, "Microarray gene expression classification with few genes: criteria to combine attribute selection and classification methods", *Expert Systems with Applications*, Vol.39, No.8, pp. 7270-7280, 2012.
[13] P. Maji, "Mutual information-based supervised attribute clustering for microarray sample classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol.24, No.1, pp.127-140, 2012.
[14] X. Wang and R. Simon, "Microarray-based cancer prediction using single genes", *BMC Bioinformatics*, Vol.12, Article.391, doi:10.1186/1471-2105-12-391, 2011.
[15] H. Liu, L. Liu and H. Zhang, "Ensemble gene selection for cancer classification", *Pattern Recognition*, Vol.43, No.8, pp.2763-2772, 2010.