ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

An Application of Clonal Selection Algorithm in Gene Selection for Cancer Classification using Microarray Data

Gunavathi C [a]*, Premalatha K[b], and Sivasubramanian K[c].

[a]School of Information Technology & Engineering, VIT University, Vellore, India.
[b]Department of CSE, Bannari Amman Institute of Technology, Sathyamangalam, India
[c]Department of ECE, K.S.Rangasamy College of Technology, Tiruchengode, India

ABSTRACT

The main objective of this research paper is to discuss the application of clonal selection algorithm in feature selection for cancer classification using microarray gene expression data. Cancer classification is a vital 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. The proposed method applies the Clonal Selection Algorithm (CSA) of Artificial Immune System for feature selection in cancer classification. The important genes are selected by T-Statistics, Signal-to-Noise Ratio (SNR) and F-Test. The classifier accuracy of k-Nearest Neighbor (kNN) technique is used as the fitness function for CSA. The simulated results are demonstrated and analyzed with 10 different cancer gene expression dataset. The experimental results show that the proposed two-step method performs well in classifying the samples.

Keywords: Cancer Classification, Gene Expression, Feature Selection, Clonal Selection Algorithm, k-Nearest-Neighbor.

Corresponding author

INTRODUCTION

The biological immune system is an organization of cells, tissues, and organs that work together to guard the body against attacks by foreign invaders. It uses learning, memory, and associative retrieval to solve recognition and classification tasks. It is an organ system anticipated to protect the host organism from the threats created from pathogens and toxic substances. Pathogen is an agent of disease which encompasses microorganisms such as bacteria, viruses and fungi. The primary role of the immune system is detection and elimination of pathogen. The immune system is extremely complex. It can distinguish and remember millions of diverse enemies.

Artificial immune systems (AIS) are a class of computationally intelligent systems encouraged by the principles and processes of the vertebrate immune system. The algorithms typically exploit the immune system's characteristics of learning and memory to solve a problem. Clonal Selection Algorithm (CSA) is a class of algorithms inspired by the clonal selection theory of acquired immunity that elucidates how lymphocytes improve their response to antigens over time called affinity maturation. These algorithms focus on the Darwinian attributes of the theory where selection is inspired by the affinity of antigen-antibody interactions, reproduction is inspired by cell division, and variation is inspired by somatic hyper-mutation.

Cancer is featured by an irregular, uncontrollable growth that may destroy and attack neighboring healthy body tissues or somewhere else in the body. Cancer classification refers to the process of building 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 converted into gene expression matrices. The rows in the matrix correspond to genes, and the columns denote samples or experimental conditions. The number in each cell represents 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 and their functions. 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 eradicate 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-m significant or informative genes. In the proposed hybrid approach, CSA is used for feature selection to classify the given samples and its fitness function is measured by the accuracy of kNN technique.

METHODOLOGY

Gene Selection Methods

T-Statistics

Genes who have considerably different expressions between normal and tumor tissues are candidates for selection. A simple T-statistic can be used to measure the degree of gene expression difference between normal and tumor tissues [5]. The top-m genes with the largest T- statistic are selected for the discriminant analysis. The formula for T-statistics is given by (1).

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

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

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[2]. 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 given by (2).

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

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 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 cancer classification. Top-m genes with the largest SNR value are selected and included for the discriminant analysis.

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 of 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-m genes with the smallest F-Test value are selected for inclusion in the discriminant analysis. F-Test formula is given by (3).

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

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

K-Nearest Neighbor Algorithm

The k-Nearest Neighbor algorithm is one of the simplest of all machine learning algorithms. It is one of the Lazy learners in which the learner have to wait until the last instant before building any model for the purpose of classifying a given test tuple. When the training sample is given, a lazy learner simply stores it and waits until it is given a test tuple. It is a way for classifying objects based on closest training examples in the feature space. Here a sample is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k-Nearest Neighbors (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 neighbor.

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.

The Euclidean distance between two samples, S1= (g_{11}, g_{12},….., g_{1n}) and S2= (g_{21}, g_{22},….., g_{2n}), is given by (4).

$$\mathrm{dist}(S1, S2) = \sqrt{\sum_{i=1}^{n} (g_{1i} - g_{2i})^2} \qquad (4)$$

Here g_{1i} and g_{2i} are 'i' th gene in samples S1 and S2 respectively.

Immune System

The description of immune system is an organ system anticipated to protect the host organism from the threats created from pathogens and toxic substances. Pathogen is an agent of disease which encompasses microorganisms such as bacteria, viruses and fungi. The immune system can be divided into the Adaptive Immune System and Innate Immune System. The difference between them is the Adaptive Immune System has a function of memory, whereas the Innate Immune system does not. The primary role of the immune system is detection and elimination of pathogen. The key to a healthy immune system is its remarkable ability to distinguish between the body's own cells, recognized as self, and foreign cells, or non-self. The body's immune defenses normally coexist peacefully with cells that carry distinctive self-marker molecules. But when immune defenders encounter foreign cells or organisms carrying markers that say non-self, they quickly launch an attack.

Anything that can trigger this immune response is called an antigen. An antigen can be a microbe such as a virus, or a part of a microbe such as a molecule. Antigens are usually proteins or polysaccharides. Antibodies or Immunoglobulins are gamma globulin proteins that are found in blood or other bodily fluids of vertebrates, and are used by the immune system to identify and deactivate foreign objects, such as bacteria and viruses. They simply proteins that are secreted as a result of the antigen provoked immune response. In short, antigens cause the disease and antibodies cure it.

Artificial Immune System

Artificial Immune Systems (AIS) concerned with computational methods inspired by the process and mechanisms of the biological immune system. In AIS, antigen usually refers the problem and its constraints. Antibody refers to the candidate solution. Affinity measure refers to fitness function. Cell cloning refers to solution replication and Somatic hyper-mutation refers to stochastic operator.

Clonal Selection Algorithm

Clonal Selection Algorithm (CSA) is a class of algorithms developed by De Castro and Von zuben [7] inspired by the clonal selection theory of acquired immunity that explains how lymphocytes improve their response to antigens over time called affinity maturation. These algorithms focus on the Darwinian attributes of the theory where selection is inspired by the affinity of antigen-antibody interactions, reproduction is inspired by cell division, and variation is inspired by somatic hyper-mutation.

The Clonal Selection principle is the whole process of antigen recognition, cell proliferation and differentiation into memory cell [8]. Here the selection of antibodies is done based on affinity either by matching against an antigen pattern or via evaluation of a pattern by a cost functions. Selected antibodies are subjected to cloning proportional to affinity. The hyper-mutation of clones is inversely proportional to clone affinity. The resultant clonal set competes with the existent antibody population for membership in the next generation. In addition low-affinity population members are replaced by randomly generated antibody population for the membership in next generation.

The main immune aspects taken into account are maintenance of the memory cells functionally disconnected from the repertoire, selection and cloning of the most stimulated cells, death of non-stimulated cells, affinity maturation and reselection of the clones with higher affinity, generation and maintenance of diversity, hyper-mutation proportional to the cell affinity.

Pseudo code for Clonal Selection Algorithm

1. Generate a set (P) of candidate solutions composed of the subset of memory cells (M) added to the remaining (Pr) population (P = Pr + M).
2. Select the n best individuals from the population based on their Affinity measure.
3. Clone these n best individuals of the population, giving rise to a temporary population of clones (C).
4. Submit the population of clones to a hyper-mutation scheme, where the hyper-mutation is proportional to the Affinity measure of the antibody. A maturated antibody population is generated (C*).
5. Re-select the improved individuals from C* to make up the memory set M. Some members of P can be replaced by other improved members of C*.
6. Replace d antibodies by novel ones. The lower affinity cells have higher probabilities are being replaced.

Cancer Classification using AIS

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

Antibody Representation

The Antibody should contain information about the solution which it represents. The most used way of encoding is a binary string. Figure 1 shows the representation of antibody. Here the binary code ‘1’ or ‘0’ is used to mark whether a gene is selected or not. So each individual 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’.

g1	g2	g3	g4	g5	g6	g7	g8	g9	g10
1	0	1	1	0	0	1	0	1	0

Figure. 1 - Antibody Representation

Affinity measure

The Affinity measure f(x) of an individual is measured by kNN technique. The accuracy of kNN classifier is used as Affinity measure. Affinity measure f(x) is given by (5).

$$fitness(x) = Accuracy(x) \qquad (5)$$

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 (6).

$$Accuracy(x) = (c/t) X 100 \qquad (6)$$

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

Clonal selection

In clonal selection n highest fitness antibodies are selected for cloning with the rate of β. The amount of clones to be generated for all these n selected antibodies is given in equation (7).

$$N_C = \sum_{i=1}^{n} \text{round}\left(\frac{\beta P}{i}\right) \qquad (7)$$

Where N_C is the total amount of clones generated, β is a multiplying factor, P is the total amount of antibodies and round() is the operator that rounds its argument towards the closest integer. Each term of this sum corresponds to the clone size of each selected antibody, e.g., for P = 10 and β = 1, the highest affinity antibody (i = 1) will produce 10 clones, while the second highest affinity antibody produces 5 clones, and so on.

Somatic hyper-mutation

A speedy accumulation of mutations is necessary for a fast maturation of the immune response. The selection mechanism provides a means by which the regulation of the hyper-mutation process is made dependent on receptor affinity. Cells with low affinity receptors may be further mutated and die if they do not improve their clone size or antigenic affinity. In cells with high-affinity antibody receptors the hyper-mutation becomes inactive [9] in a gradual manner.

For mutation the Cauchy mutation operator [10] is applied. The one dimensional Cauchy density function centered at the origin is defined as follows:

$$f(x) = \frac{1}{\pi}\frac{t}{t^2 + x^2} \quad -\infty < x < \infty \qquad (8)$$

Where t>0 is a scale parameter.

The adaptive mutation rate P_m depends on the fitness values of the Antibodies. The adaptation allows the individuals having fitness values of over-average to maintain their value and the individuals with below average fitness values to disturb. The mutation rate adaptation rule is given in equation (9).

$$P_m = \begin{cases} k_1 \times \dfrac{F_{max} - F}{F_{max} - F_{avg}} & F \geq F_{avg} \\ k_2 & F < F_{avg} \end{cases} \qquad (9)$$

In this equation, F denotes the fitness value of the individual, F_{max} denotes the best fitness value of the current generation, and F_{avg} denotes the average fitness value of the current generation. The constants k_1 and k_2 are chosen as $1.75/(n \times N^{1/2})$, where n and N represents number of antibodies and length of antibody (Back 1992).

Experimental Result Analysis

The proposed method uses T-statistics, Signal-to-Noise Ratio and F-Test to select top-m genes. The m values are taken as 10, 50 and 100. These genes alone used for further classification. Clonal selection algorithm of AIS is applied on the selected genes. For classification purpose 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 an Affinity measure in CSA. The kNN with 5-fold cross validation method gives the classification accuracy as output. The CSA is configured to have 20 individuals and were run for 500 generations in each trial. In CSA Cloning rate is taken as 0.5.

In order to assess the performance of the proposed method, 10 datasets were analyzed. These datasets were collected from Kent Ridge Biomedical Data Repository. The details about the datasets are given in Table 1.

Table 1 - Cancer Microarray Gene Expression Dataset

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 2 gives the Parameters and their values used in this method.

Table 2 - Parameter and their values

Parameter	Value
Antibody size	10, 50 and 100
Population size	20
Maximum no. of Generations	500
Cloning rate β	0.5
d	20% of population size
Distance Measure in kNN	Euclidean distance
k-value is kNN	3

Table 3 shows the results obtained from CSA based method. It gives the Classification accuracy with minimum number of genes with top-m genes when applied different measures like Signal-to-Noise ratio, T-statistics and F-Test.

Table 3 - Classification Accuracy of CSA

S.No.	Dataset	T-statistics			SNR			F-Test		
		Accuracy (%)			*Accuracy (%)*			*Accuracy (%)*		
		10	50	100	10	50	100	10	50	100
1	CNS	75	75	75	81.25	87.5	87.5	87.5	81.25	87.5
2	DLBCL Harvard	80	88	84	96	**100**	96	80	84	84
3	DLBCL outcome	72.72	77.27	86.36	63.63	72.72	68.18	59.09	81.81	72.72
4	Lung Cancer Michigan	91.3	91.3	95.65	**100**	**100**	**100**	**100**	**100**	**100**
5	Ovarian Cancer	67.24	68.97	79.31	98.27	98.27	98.28	96.55	**100**	**100**
6	Prostate outcome	85.71	57.14	57.14	85.71	71.42	85.71	71.42	71.42	57.14
7	AML-ALL	68	77.27	81.81	95.45	95.45	95.45	95.45	95.45	**100**
8	Colon Tumor	70	75	70	90	90	90	75	85	90
9	Lung Harvard2	80	82.5	80	**100**	**100**	**100**	97.5	**100**	**100**
10	Prostate	60.97	70.73	75.61	73.1	63.41	56.09	90.24	89.92	87.8

Table 4 represents the corresponding measure which gives the maximum accuracy with minimum number of genes among top-m genes in CSA based approach.

Table 4 - Maximum Accuracy of CSA

S.No.	Dataset Name	Maximum Accuracy (with Minimum Genes)		Gene selection method
		Accuracy (%)	*m - value*	
1	CNS	87.5% 87.5% 87.5%	10 50 100	F-Test SNR F-Test
2	DLBCL Harvard	96% 100 % 96%	10 50 100	SNR SNR SNR
3	DLBCL outcome	77.27 % 77.27 % 77.27 %	10 50 100	SNR T-statistics, F-Test T-statistics, F-Test
4	Lung Cancer Michigan	100 % 100 % 100 %	10 50 100	SNR, F-Test SNR, F-Test SNR, F-Test
5	Ovarian Cancer	98.27% 100 % 100 %	10 50 100	SNR F-Test F-Test
6	Prostate outcome	85.71 % 71.42 % 71.42 %	10 50 100	T-statistics, SNR SNR, F-Test SNR, F-Test
7	AML-ALL	95.45 % 100% 100%	10 50 100	F-Test F-Test F-Test
8	Colon Tumor	95 % 90% 90%	10 50 100	SNR SNR SNR
9	Lung Harvard2	100 % 100 % 100 %	10 50 100	SNR SNR, F-Test SNR, F-Test
10	Prostate	92.68 % 90.24% 82.92%	10 50 100	F-Test F-Test F-Test

Figure 2, Figure 3 and Figure 4 depict the Classification Accuracy obtained from different measures when applying CSA of AIS and kNN on top-m genes.

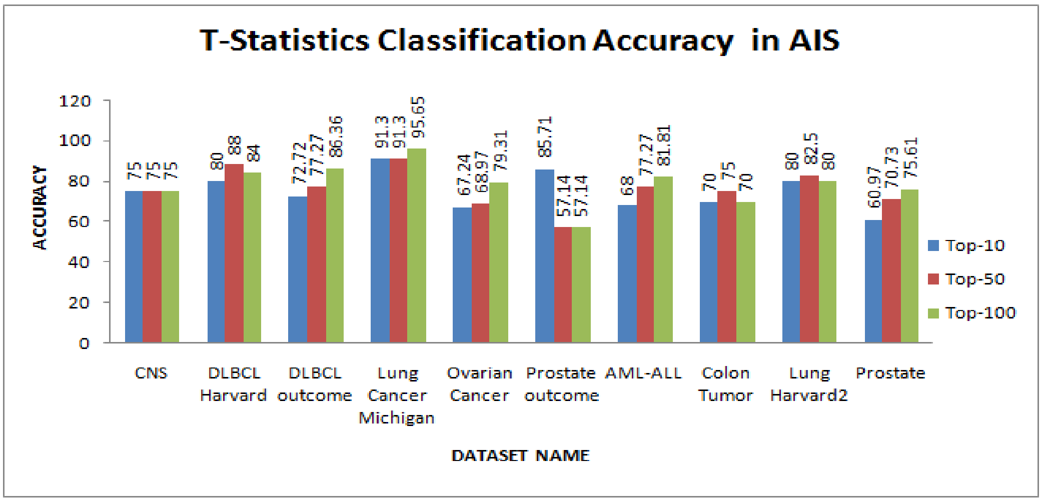

Figure 2 - T-Statistics Accuracy in AIS

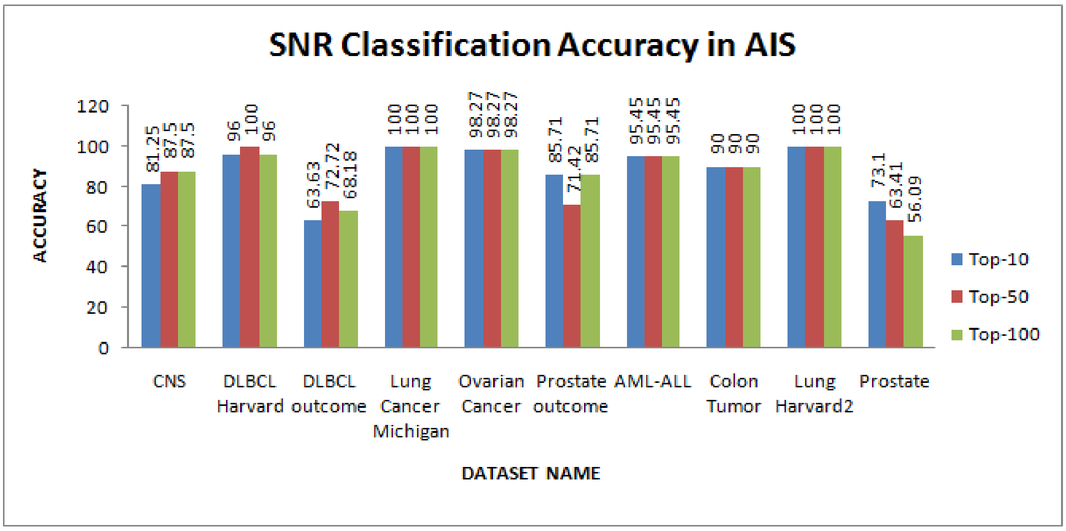

Figure 3 - SNR Accuracy in AIS

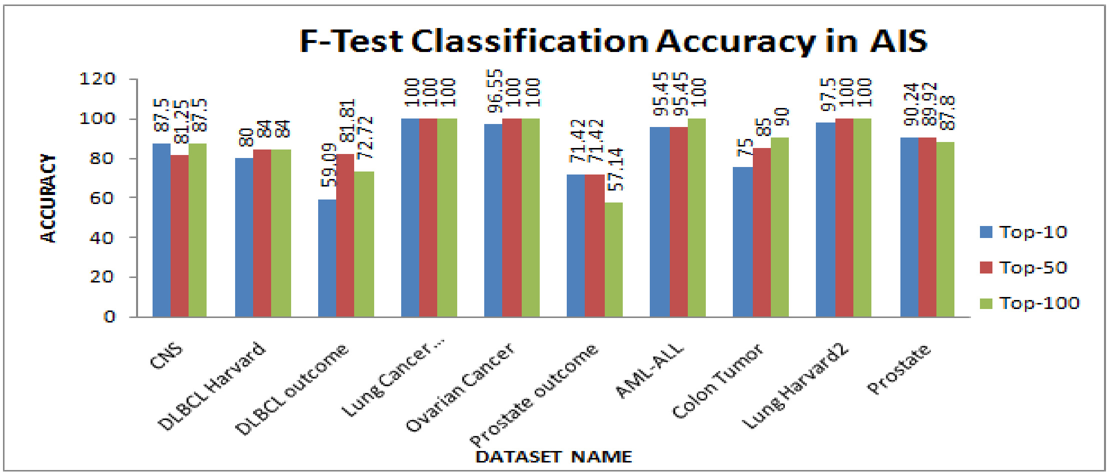

Figure 4 - F-Test Accuracy in AIS

Figure 5 and Figure 6 depict the Maximum Accuracy obtained for different Cancer types and their corresponding measures when applying CSA and kNN on top-m genes.

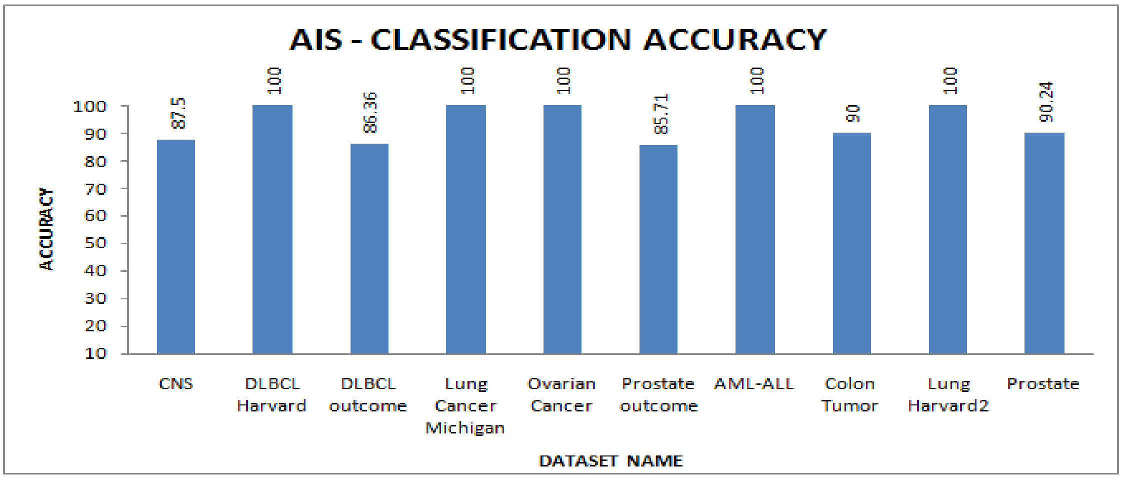

Figure 5 - AIS Classification Accuracy

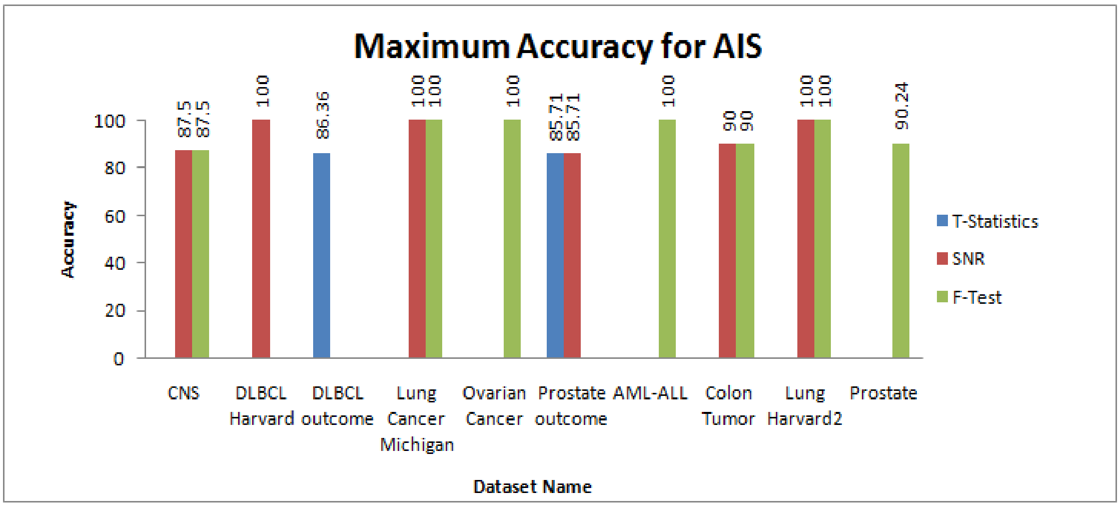

Figure 6 - Maximum Accuracy for AIS

CONCLUSION

T-statistics, Signal-to-Noise Ratio and F-Test are the feature selection methods used to select the important genes. CSA with kNN Classifier method is applied on those top-m genes in this research work. Here the classification accuracy of kNN is considered as the fitness function for the CSA. The kNN classifier is one of the most famous neighborhood classifier in pattern recognition. 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. The CSA based approach provides 100% classification accuracy for 5 datasets. The above 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] Domany E. Cluster Analysis of Gene Expression Data. Journal of Statistical Physics. 2003; 3-6(110), 1117-1139.

[2] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 1999; 5439(286), 531-537.

[3] Chandra B, Gupta M. An efficient statistical feature selection approach for classification of gene expression data. Journal of Biomedical Informatics. 2011; 44(4), 529–535.

[4] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19), 2507-2517.

[5] Yendrapalli K, Basnet R, Mukkamala S, Sung AH. Gene Selection for Tumor Classification Using Microarray Gene Expression Data. *Proceedings of the World Congress on Engineering, London,* 2007, 290-295.

[6] Momiao X, Wuju L, Jinying Z, Li J, Eric B. Feature (Gene) Selection in Gene Expression-Based Tumor Classification. Journal of Molecular Genetics and Metabolism. 2001; 73(3), 239–247.

[7] De Castro LN, Von Zuben FJ. The clonal selection algorithm with engineering applications. *Proceedings of the Genetic and Evolutionary Computation Conference*, Las Vegas, USA, 2000, 36–37.

[8] Burnet F. The clonal selection theory of acquired immunity. Cambridge University Press. 1959.

[9] Berek C, Ziegner M. The Maturation of the Immune Response. Imm. Today. 1993; 8(14), 400-402.

[10] Feller W. An Introduction to Probability Theory and Its Applications. 2nd edition. volume 2. John Wiley & Sons Inc: 1971.