An Efficient Diseases Classifier based on Microarray Datasets using Clustering ANOVA Extreme Learning Machine (CAELM)

Shamsan Aljamali[1], Zhang Zuping [2] and Long Jun[3]

**[1] School of Information Science and Engineering, Central South University
Changsha, 410083, China**

**[2] School of Information Science and Engineering, Central South University
Changsha, 410083, China**

**[3] School of Information Science and Engineering, Central South University
Changsha, 410083, China**

Abstract

Cancer is a group of diseases distinguished by unregulated growth and spread of cells which has become one of the most dangerous diseases. As a result of the victims of cancer are increasing steadily, the necessity is increasing to find classification techniques for cancer diseases. The present study is aimed to obtain better results of the classification model with high accuracy. Herein, we proposed a method of developing an efficient classifier based on microarray datasets. Moreover, we focused on accuracy, dimensionality reduction and fast classification issues. The proposed method Clustering ANOVA Extreme Learning Machine (CAELM) is a hybrid approach based on Extreme Leaning Machine with RBF kernel function. This hybrid approach consist of two phases: data preprocessing (normalization and genes selection) and data classifying. K-mean clustering was utilized as a method for clustering microarray datasets into three groups, then ANOVA were applied to analysis of variance between this groups to pick out the significant genes which were used in classification process. In case combining clustering with statistical analysis (CAELM) a much better classification accuracy is given of 95,94,100% for leukemia ,prostate and ovarian respectively . In addition, the proposed approach reduced time complexity with good performance.

***Keywords**: CAELM, RBF kernel, K-mean, ANOVA, microarray, genes selection, cancer classification.*

1. Introduction

Cancer is a group of diseases distinguished by unregulated growth and spread of cells [1] which is one of the most challenging studies for researchers in the current century. There has been lot of proposals from various researchers on cancer classification and detailed study is still on in the domain of cancer classification. Many researchers applied various approaches depend on the t test statistic [2,3,4] and ANOVA F test statistic [5,6] to do the process for selecting significant subset of informative genes in order to classify diseases from microarray data with reduce data redundancy, reduce training time and improve classification performance and accuracy. Microarrays gene expression is used with more interesting for cancer classification as effective tool which is a collection of microscopic DNA spots connected to a solid surface which are used to compute the expression levels of vast numbers of genes at the same time or to genotype multiple regions of a genome [7]. High-density DNA microarray [8] gathers the behaviors of various genes concurrently and the gene expression [9] profiles have been utilized for the cancer classification in recent time. But one of the major challenges is that Microarray data contains a large number of genes with a small number of samples. This high dimensionality problem has prevented many existing classification methods from directly dealing with this type of data. In addition, Microarray datasets contains high levels of noise, and these data lead to unreliable and low accuracy analysis as well as the high dimensionality problem. These types of data make the most current classification methods are not robust enough to handle it properly. In this paper, accuracy, dimensionality reduction and fast classification issues are focused on. We discuss the proposed classifier technique, which is based on Extreme Learning Machine with Gaussian (RBF) as kernel function during process of building ELM classification model. In addition, biomedical datasets are used to test the performance of this classifier, having varieties of cancer data which are Leukemia, prostate and ovarian microarray datasets. Moreover, a comparison of the proposed classifier with extreme learning machine and support vector machine classifiers is performed .Extreme learning Machine [10] solves issues such as local minima, improper learning rate and over fitting usually occurs in iterative learning techniques and completes the training very fast. However, the usage of ELM will take more time when large data is used for classification. All the above problems was overcome by using the proposed ELM technique called Clustering ANOVA Extreme Learning Machine CAELM. The proposed technique has the capability to perform the classification in a very short time with high accuracy compared to conventional techniques by clustering microarray datasets into three groups. Using K-mean algorithm, Leukemia microarray dataset was clustered into three groups, prostate into three groups and also ovarian into three groups. After clustering we used one way ANOVA F test for genes selection , which performed an analysis of variance pick out the significant genes from original microarray dataset (genes selection) for classification process.

2-Related work

Guyon et al. [11] utilized Support Vector Machines as method of Gene Selection for Cancer Classification. In this study, the author proposed a solution for the problem of selection of a small subset of genes from many samples of DNA microarrays that contains many patterns of gene expression data. Using available training examples from cancer and normal patients, the approach develop suitable classifier for Diagnosis of genetic diseases, as well as drug discovery. Previous attempts to solve this problem select genes with correlation techniques. The author proposes another method of gene selection using Support Vector Machine methods depends on Recursive Feature Elimination (RFE). This experiment shown that the genes selected by proposed method produced better classification performance and are biologically relevant to cancer. Lipo wang et al., [12] presented the accurate classification of cancer with the use of expression of very few genes, the author focuses at choice the smallest set of genes that can guarantee classification of cancers with high accuracy from microarray data with the use of supervised machine learning techniques. The importance of determining the smallest gene subsets is in three phases as below: 1-It significantly decreases the computational time and noise is caused by unrelated genes. In the illustrations examined in this paper, determining the minimum gene subsets still allow for extraction of simple diagnostic rules that directs to accurate diagnosis without the requirement for any classifier. 2- It makes gene expression examinations easier to contain subset has a very small number of genes slightly than thousands of genes that can decrease the cost for cancer testing appreciably. 3-It terms for additional examinations into the probable biological relationship among these few numbers of genes and cancer expansion and treatment. Cınar et al. [13] support vector machines and artificial neural networks used in prostate cancer diagnosis. The aim of this study is to develop a classifier based expert system for early diagnosis of the organ in constraint stage to obtain masterful decision making without biopsy by utilizing some selected features. The other objective is to Determine whether any relationship between prostate cancer, Body Mass Index (BMI) and smoking. This study were used the data which were taken from 300 men as samples. (Benign prostatic hyperplasia, 100: prostate adenocarcinoma and 200: chronic prostatitis). Weight, height, Free PSA,BMI, age ,prostate volume, density, Prostate Specific Antigen (PSA), smoking, systolic, diastolic, pulse and Gleason score features were used and independent sample t-test was used for feature selection. For classifying related data, it was chosen the following classifiers; Scaled Conjugate Gradient (SCG), Broyden-Fletcher-Goldfarb-Shanno (BFGS) and Levenberg-Marquardt (LM) training algorithms of Artificial Neural Networks (ANN) and linear, polynomial and radial based kernel functions of Support Vector Machine (SVM). This study was determined that the prostate cancer is not affected by BMI whereas smoking has relationship to increase the prostate cancer risk. Since PSA, smoking, volume and density features were to be statistically significant, they were chosen for classification process as significant features. The proposed method was implemented with polynomial based kernel function, which was its accuracy is 79% as the best performance.

3- Methodology

One of the major problems is that Microarray data contains a large number of genes with a small number of samples. This high dimensionality problem has prevented many existing classification methods from directly dealing with this type of data. Microarray data contains high levels of noise, and these data lead to unreliable and low accuracy analysis as well as the high dimensionality problem. Most current classification methods are not robust enough to handle these types of data properly. To address these problems, in this paper, a new methodology called Clustering ANOVA Extreme Learning Machine

CAELM is proposed. In this approach RBF kernel is used with ELM for better performance. CAELM is a hybrid approach, consisting of two main phases: the data preprocessing (clustering and genes selection) phase and the classification phase, as shown on Fig. (4.1). In the data preprocessing phase, an integrated k-means clustering algorithm and ANOVA statistical test are used to filter the data. In the classification phase, ELM classifier is used to classify the proposed data. Each phase will be explained in details in the next sections.

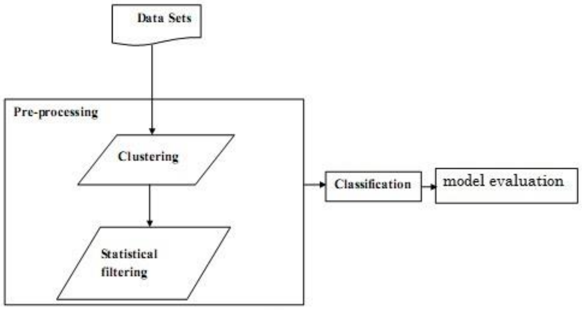

Fig.1 the proposed classifier CAELM

3.1 Preprocessing Phase

For classifying required data, a hybrid approach for data preprocessing having both (k-means) clustering and statistical filtering (ANOVA test) were used.

3.1.1 k-means clustering

K-means is one of the most commonly used clustering algorithms which Requires selecting K initial centroids randomly where K is a user defined number of required clusters. Each point is then assigned to a closest centroid and the set of points close to a centroid form a cluster. The centroid is updated according to the points in the cluster and this process continues until the points stop changing their clusters. In other words, its objective is to find:

$$\sum_{i=1}^{k}\sum_{x\in S_i} ||x-\mu_i||$$

Where μi is the mean of points in Si.

In this paper K-means divided microarray data into three groups: Leukemia to three groups, Prostate to three groups and ovarian cancer to three groups. In each process of clustering we set k = 3 and maximum iteration = 50. Iteration process was uneven in the each process of clustering, but did not reach to 50.

3.1.2 ANOVA Model

ANOVA test is an extension of the t-test to more than two experimental conditions. It selects genes that have significant differences in means across three or more groups of samples. P- Values in ANOVA are computed from the theoretical F-distribution. F-statistics are calculated for each gene, and a gene Take in consideration as significant if p-value associated with its F-statistic is smaller than the user-specified alpha or critical p-value. The most significant varying information has the smallest p-values. Within groups estimate of:

$$\sigma_y^2 = \frac{\sum_{ij}(y_{ij}-\bar{y}_j)^2}{\sum_j(n_j-1)} = \frac{SS_{WG}}{df_{WG}} = MS_{WG}$$

Between groups estimate of:

$$\sigma_y^2 = \frac{\sum_j n_j(y_j-\bar{y}_{ij})^2}{\sum_j(k-1)} = \frac{SS_{BG}}{df_{BG}} = MS_{BG}$$

$$F(df_{BG},df_{WG}) = b\,\frac{between\ group\ estimate\ of\ \sigma_y^2}{within\ group\ estimate\ of\ \sigma_y^2} = \frac{MS_{WG}}{MS_{BG}}$$

In this paper ANOVA test was applied when α (P-value) = 0.01, P values in ANOVA test are computed from the theoretical F-distribution. By using the k best ranking genes according to ANOVA test-statistic we would select high correlated genes.

3.2 Extreme Learning Machine Classifier

After clustering and filtering process, data are passed to ELM classifier for the classification process .Extreme learning machine (ELM) meant for Single Hidden Layer Feed-forward Neural Networks [5] (SLFNs) will randomly selected the input weights and analytically determines the output weights of SLFNs. ELM contains an input layer, hidden layer and an output layer. In this paper Gaussian (RBF) radial basis function kernel used as kernel function in the process of building SVM, ELM and CAELM classification models. Usually, the recommended kernel function [14] for nonlinear problems is the Gaussian radial basis function because it resembles the sigmoid kernel for certain parameters and it requires less parameter than a polynomial kernel. The parameter of kernel function γ and the parameter C, which controls the complexity of the decision function versus the training error minimization, can be determined by running two dimensional grid search, which means that the values for pairs of parameters (C, γ) are generated in a predefined interval with a fixed step. The Tolerance value with Gaussian function was (0.001).The important properties of this solution are develop an efficient and effective classifier with High accuracy, High learning speed Minimum training error and Best generalization performance.

3.3 Datasets

In this section, we discuss microarray datasets. We collected this datasets from NCBI web site (GEO database). Table (1) shows the summary of the characteristics of the three data sets. Each Microarray dataset is described by the following parameters: Genes: the number of genes or attributes, Class: the number of classes and Sample: the number of samples in the dataset.

Table 1: microarray datasets

Name	Leukemia	Ovarian	Prostate
Sample	72	253	136
Gene	7129	15154	12600
class	2	2	2

4- Results and Discussion

In this section, we discuss the implementation of the proposed classifier microarray datasets. Applying K-means to cluster three datasets, there is three Microarray datasets as bioinformatics data (leukemia, Prostate and ovarian). K-means used to divide microarray data into three groups: Leukemia to three groups, Prostate to three groups and ovarian cancer and to three groups. In each the process of clustering we set k = 3 and maximum iteration = 50. Iteration process was uneven in the each process of clustering, but did not reach to 50. After clustering the datasets, we will move to the second step from preprocessing data, a process of genes selection microarray data by ANOVA test.

4.1 Clustering Microarray Data

Applying clustering to three microarray datasets, Leukemia, prostate cancer and ovarian, Table (2) displays results clustering for each datasets.

Table 2: clustering microarray datasets

Datasets	Group1	Group2	Group3
Leukemia	22	14	36
Prostate	34	62	40
Ovarian	98	80	75

4.1.1 Clustering leukemia microarray

Fig. (2, 3, 4) show the result of clustering Leukemia into three groups by k-means cluster. The red line represents the graph of the mean of all the samples across the genes while the mean for all samples of a gene is represented by a dot, and the standard deviations for the samples in that gene are shown above and below it. Fig. (2) clarifies result of accumulation of genes where the number of genes is 7129, the centers averages do not appear to some extent, but it is not far from the halfway line. However, we can see that the standard deviation of the data shows that there is a gap in the leukemia data.

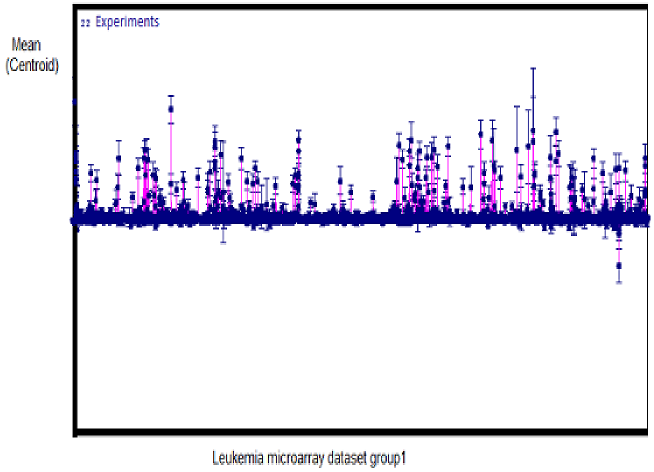

Fig. 2 Group one (22 samples) by k-means cluster (Leukemia datasets)

Fig. (3 and 4) clarify the result of number many genes do not appear centers means are clear, and also shows the standard deviation of the means of the data shows the dispersion of this data.

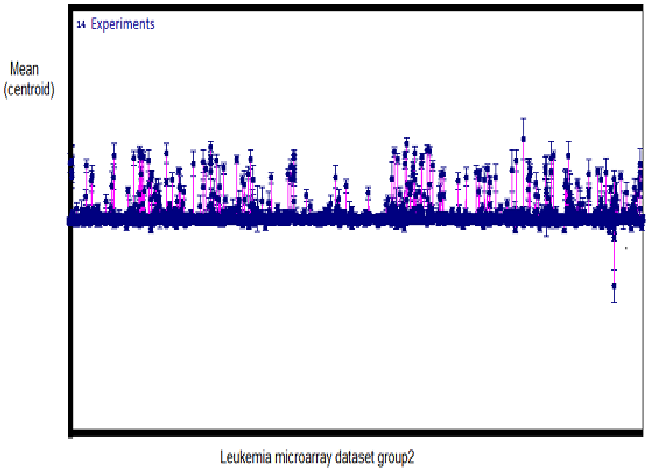

Fig. 3 Group two (14 samples) by k-means cluster (Leukemia datasets)

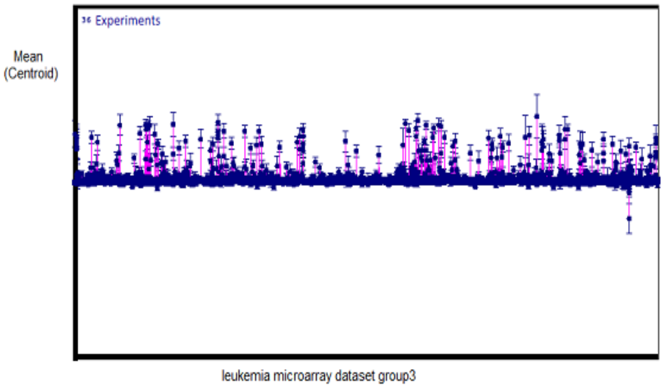

Fig. 4 Group three (36 samples) by k-means cluster (Leukemia datasets)

4.1.2 Clustering prostate microarray

Fig. (5, 6, 7) illustrate clustering Prostate datasets into three groups by k-means cluster. The red line represents the graph of the mean of all the samples across the genes. That is, the mean for all the samples of a gene is represented by a dot, and the standard deviation for the samples in that gene are shown above and below it. Fig. (5) represents the first group of data of the prostate and show overlapping centers means, for the large number of genes, and shows the standard deviation of some of the data significantly, and some are small. Fig. (6 , 7) illustrate the result of number many genes do not appear centers means are clear, and also shows the standard deviation of the means of the data shows the dispersion of this data.

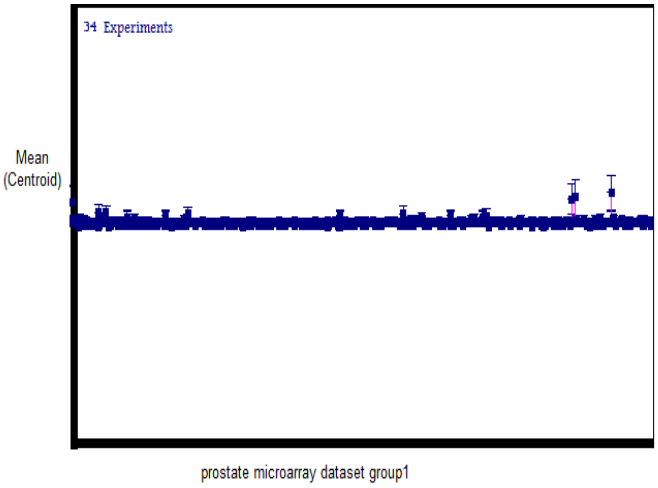

Fig. 5 Group one (34 samples) by k-means cluster (Prostate datasets)

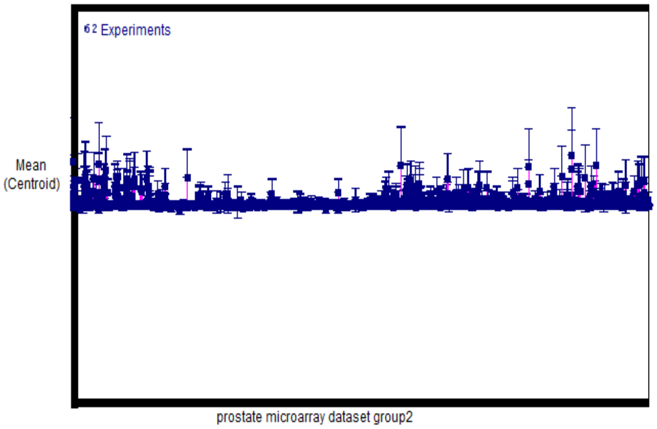

Fig. 6 Group two (62 samples) by k-means cluster (Prostate datasets)

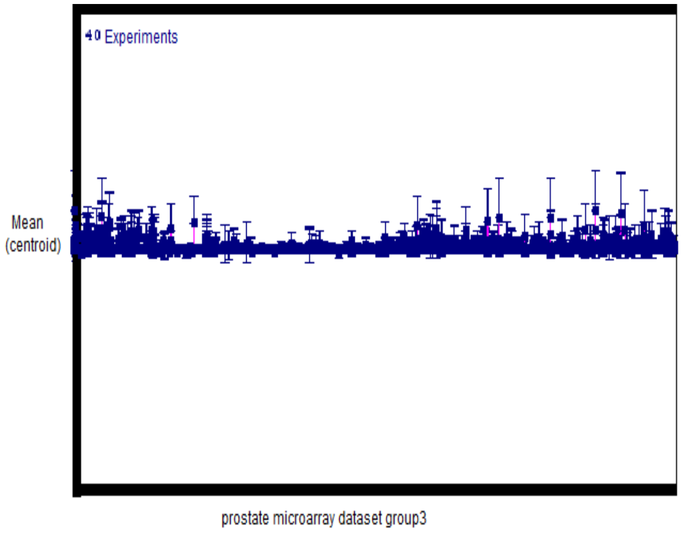

Fig. (7): Group three (40 samples) by k-means cluster (Prostate datasets)

4.1.3 Clustering Ovarian microarray

Fig. (8, 9, 10) show clustering of ovarian datasets into three groups using k-means cluster. As in Leukemia and prostate datasets, the red line represents the graph of the mean of all the samples across the genes. That is, the mean for all the samples of a gene is represented by a dot, and the standard deviation for the samples in that gene are shown above and below it. Ovarian cancer datasets contain 15154 the number of genes, which affect the clarity of centers means with it seem far apart and the data appear scattered, and negative values and positive deviations measure the dispersion.

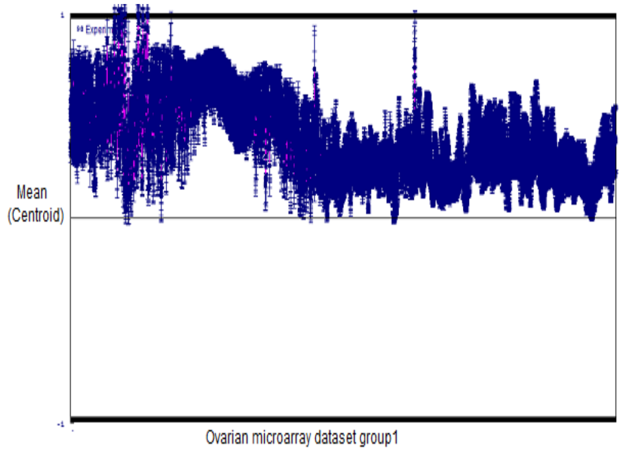

Fig. 8 Group one (98 samples) by k-means cluster (Ovarian cancer datasets)

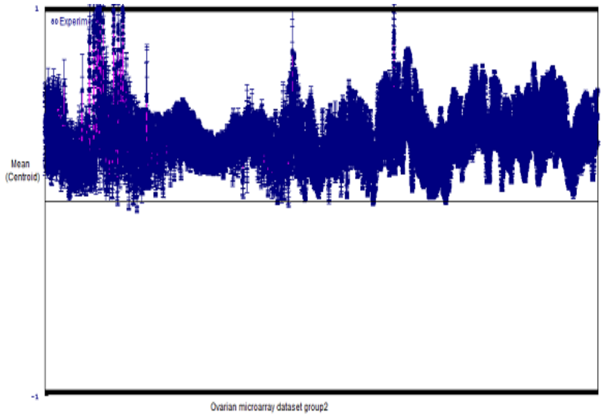

Fig. (9): Group two (80 samples) by k-means cluster (Ovarian cancer datasets)

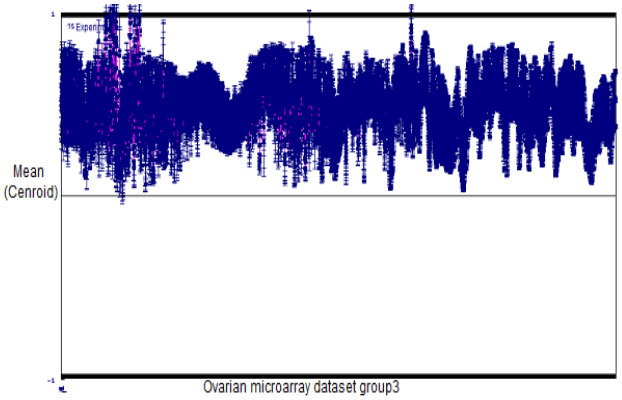

Fig. 10 Group three (75) samples by k-means cluster (Ovarian cancer datasets)

4.2 Filtering Microarray Data (genes selection)

After microarray datasets were clustered, we applied ANOVA test, and we got a number of significant genes. Table (3) shows the number of significant genes that have been selected from Microarray datasets, 713 Leukemia, 832 ovarian Cancer and 725 Prostate genes respectively. ANOVA test was applied when α (P-value) = 0.01, P values in ANOVA test are computed from the theoretical F-distribution. By using the k best ranking genes according to ANOVA F-statistic we would select highly correlated genes.

Table 3: features selection from microarray datasets

Datasets	Genes	Significant genes	P value
Leukemia	7129	713	0.01
Prostate	12600	725	0.01
Ovarian	15154	832	0.01

4.3 Classifying microarray datasets

After the process of clustering and filtering, we got a number of significance genes with microarray data. Based on the results of clustering and filtering, we applied the proposed CAELM classifier. A comparison with SVM and ELM with results (CAELM) for this research has been accomplished.

4.3.1 Accuracy evaluation

Tables (4, 5, 6) show the results of SVM, ELM and CAELM classifiers for Leukemia, prostate cancer and Ovarian datasets. The comparison between classifiers based on TPR, FPR, Accuracy, and Precision as evaluation measurements. Tables (4, 5, 6) show the amount of the increase and decrease in accuracy, precision in each microarray datasets. With leukemia datasets in Table (4), the accuracy with SVM 0.75 and ELM is 0.85 while with CAELM is 0.95, and precision 0.56 with SVM , and 0.81 with ELM while with CAELM was 0.86. From these results we conclude that CAELM was the best compared with SVM and ELM as shown in the Table (4).

Table 4 : comparison SVM and ELM with CAELM(Leukemia datasets)

Method	TPR	FPR	Precision	ACC
SVM	0.75	0.75	0.56	0.75
ELM	1	0.44	0.81	0.85
CAELM	**1**	**0.085**	**0.86**	**0.95**

With Prostate datasets, Accuracy with CAELM is 94 while with SVM and ELM was 0.63, 0.93 respectively. In addition, precision, CAELM was better, as shown in Table (5).

Table 5: comparison SVM and ELM with CAELM(prostate datasets)

Method	TPR	FPR	Precision	ACC
SVM	0.63	0.60	0.58	0.63
ELM	1	0.14	0.89	0.93
CAELM	**1**	**0.15**	**0.90**	**0.94**

with Ovarian datasets, it appears from the Table 6 that CAELM was the perfect classifier with the best accuracy and precision , hence Accuracy and precision with CAELM are 100% while SVM and ELM accuracy was 0.85 , 0.98 and precision 0.87,.98 respectively . as will be seen later, even with time CAELM was the best.

Table 6 : comparison SVM and ELM with CAELM(Ovarian datasets)

Method	TPR	FPR	Precision	ACC
SVM	0.81	0.045	0.87	0.85
ELM	1	0.028	0.98	0.98
CAELM	**1**	**0**	**1**	**100%**

4.3.2 Time evaluation

Table 7 illustrates CPU time consuming for building and testing the classification models. The process of build and test was applied in (CAELM) with the genes are 713, 725 and 832 for the Leukemia, prostate and ovarian respectively while SVM and ELM the genes was 7129, 12600 and 15154 for leukemia, prostate and ovarian respectively .The time was taken to build and test, was much less with (CAELM), the different in time between CAELM , ELM and SVM was large because the number of genes for CAELM is less than SVM, ELM. Test and build were run on a machine with the following hardware and software specifications: Intel(R) Core™ i5 – 240M CPU 2.60GHz, 4 GB RAM memory, and Microsoft Windows 10 Professional insider 64 bit operating system.

The following figures(11,12,13,14,15,16) show How the complexity of the time was reduced by the proposed classifier CAELM, figures (11,12) for leukemia show The difference in time between ELM and CAELM . And because the difference in time is extra-large between the SVM on one hand and ELM and CAELM on other hand. Therefore we could not draw a time curve for three classifiers in the same figure, so we did draw time curve for ELM and CAELM in the same figure and SVM In separated figure.

Table 7: CPU time for build and test classification models SVM, ELM and CAELM

Datasets	CPU time(S)					
	SVM		ELM		CAELM	
	Build	Test	Build	Test	Build	Test
Leukemia	0.08	0.04	0.012	0.018	0.0009	0.0019
Prostate	0.44	0.38	0.02	0.03	0.0042	0.0012
Ovarian	1.48	1.2	0.06	0.05	0.006	0.005

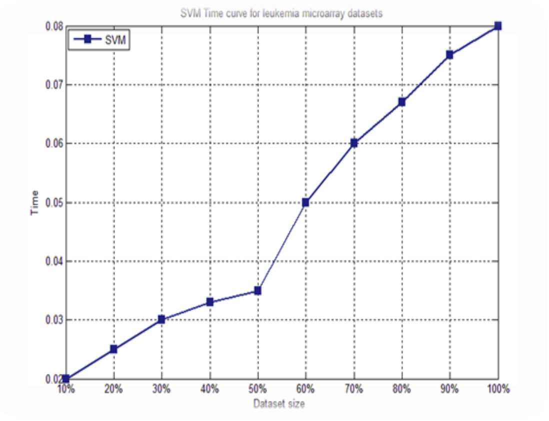

Fig.11 SVM time curve for Leukemia microarray dataset

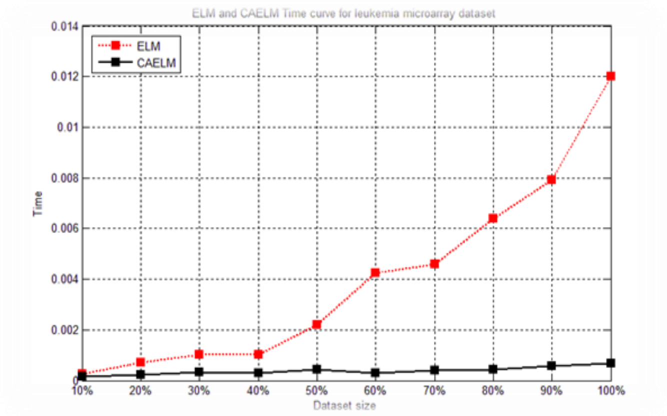

Fig.12 ELM and CAELM time curves for leukemia microarray dataset

As shown in Figures(11,12) The proposed method dramatically reduced the time complexity, also note that the training time of SVM classifier was much more than ELM and CAELM , while the training time of CAELM was much less than SVM and ELM. We also note that the time complexity of the SVM and ELM is $O(n^2)$ approximately , this means poor performance while the time complexity of CAELM is $O(n)$ This means that the proposed method CAELM gave a good performance. As recorded in the table7 note that the Consumer time to build the classification model using the proposed method CAELM was 0.0009s much less than ELM and SVM were 0.012 s ,0.08s respectively.

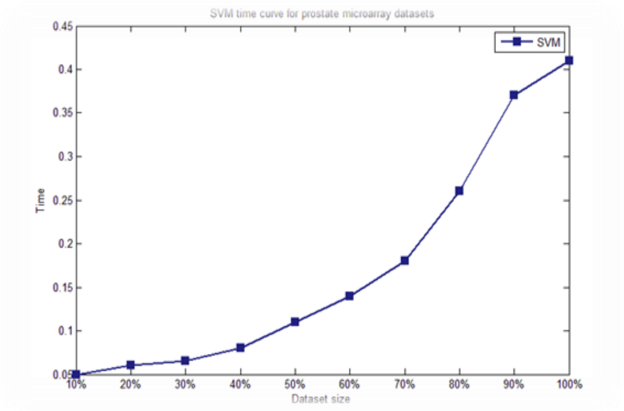

Fig.13 SVM time curve for Prostate microarray dataset

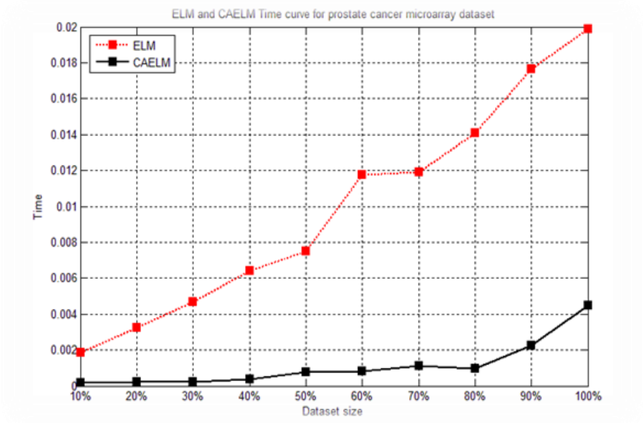

Fig.14 ELM and CAELM time curves for prostate microarray dataset

As shown in figures (13, 14) the proposed method CAELM is still the best in time reduction with a good performance. As recorded in table 7 note that the consumed time to build the classification model using the proposed method CAELM was 0.0042s while the consumed time was 0.02s and 0.44s by using ELM and SVM respectively. Accordingly, we conclude that the proposed method accomplished the classification process of microarray datasets in more less training time with high accuracy and good performance.

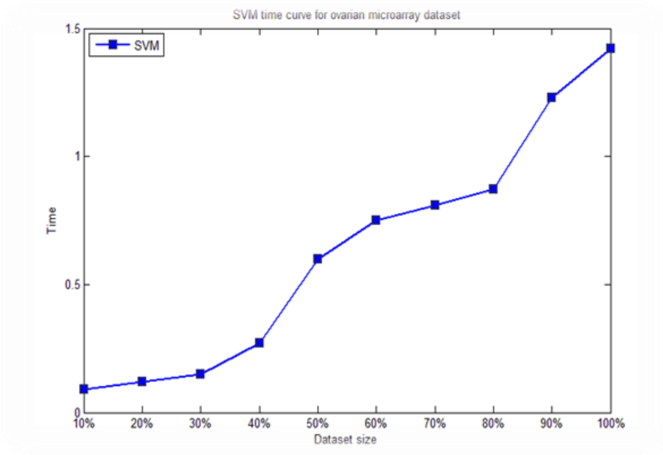

Fig.15 SVM time curve for Ovarian cancer microarray dataset

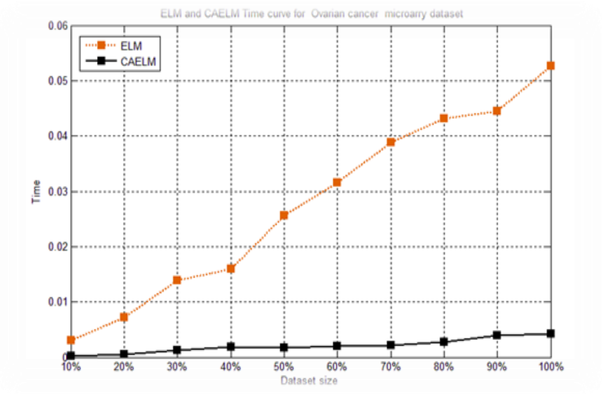

Fig.16 ELM and CAELM time curves for Ovarian cancer microarray dataset

as in leukemia and prostate the figures (15,16) show how the CPU time was reduced by the proposed method CAELM. As recorded in table 7 note that the consumed time to build the classification model using the proposed method CAELM was 0.006s While the time were 0.06s and 1.48s using ELM and SVM respectively.

5- Conclusion

In this paper, a new methodology was proposed which consist of a hybrid approach of K-means and ANOVA test for microarray datasets filtering (genes selection). In addition, Extreme Learning Machine with RBF kernel function was applied as one of the best classification algorithm for classifying cancer diseases.

Three datasets were used: leukemia, prostate and ovarian cancers. By using the proposed method Clustering ANOVA Extreme Learning Machine CAELM we got the highest accuracy compared to the two other classifiers ELM and SVM which was 0.95, 0.94 and 100% for leukemia, prostate and ovarian respectively. Moreover the performance of the proposed method CAELM achieved higher classification accuracy and good performance with less training time. The proposed method dramatically reduced time complexity .We hope using this classifier to help the doctors to diagnose the diseases as a part of a biomedical informatics system.

6- References

[1] W.Kinzler , W.Kenneth , Vogelstein, Bert," Introduction to the genetic basis of human cancer ",New York: McGraw-Hill, Medical Pub. Division, 2002.

[2] M. Xiong , L. Jin , W Li , and E. Boerwinkle , "Computational methods for gene expression-based tumor classification", Biotechniques ,Vol.29,No.6, 2000 , pp. 1264–1270.

[3] D. V Nguyen and D. M. Rocke , "Tumor classification by partial least squares using microarray gene expression data", Bioinformatics. Vol.8,No.1, 2000, pp. 39–50.

[4] H. Liu , J. Li , and L. Wong," A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns", Genome Inform Ser Workshop Genome Inform,Vol.13, 2002 , pp.51–60.

[5] D. Ghosh, "Singular value decomposition regression models for classification of tumors from microarray experiments", Proceedings of the 2002 Pacific Symposium on Biocomputing; Lihue, Hawaii,2002 , pp. 18–29.

[6] D. V Nguyen ,and D. M. Rocke , "Multi-class cancer classification via partial least squares with gene expression profiles" , Bioinformatics, Vol.18, No.9 , 2002 , pp.1216–1226.

[7] M. Ringner, C. Peterson and J. Khan ," Analyzing Array Data Using Supervised Methods, Pharmacogenomics", Vol.3, No.3, 2002 , pp. 403-415.

[8]O. Troyanskaya et al, "Missing Value Estimation Methods for DNA Microarrays", Bioinformatics , Vol.17 , No.6, 2001, pp. 520-525.

[9] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson Jr., J.R. Marks, and J.R. Nevins, "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles", Proc. Nat'l Academy of Sciences USA, Vol.98, No.20,2011, pp. 11 462-11 467.

[10] M.-B , Li, G.-B , Huang , P. Saratchandran , and N. Sundararajan, "Fully Complex Extreme Learning Machine, Neurocomputing" , Vol.68, No1-4,2005, pp. 306-314.

[11]Guyon, I, J. Weston, S. Barnhill , and V. Vapnik, "Gene selection for cancer classification using support vector machines". Mach. Learn., 46: 389422. DOI: 10.1023/A:1012487302797,2002.

[12] L. Wan, F. Chu , and W. Xie," Accurate Cancer Classification Using Expressions of Very Few Genes", IEEE/ACM Transactions on Computational Biology and Bioinformatics, VOL.1, 2007, Pp. 40-53.

[13]Cınar, M., M. Engin, E.Z. Engin , and Y.Z. Atesci , "Early prostate cancer diagnosis by using artificial neural networks and support vector machines". Expert Syst. Appli., Vol.36, 2009, pp. 6357-6361.

[14] N.Cristianini , and J. Shawe-Taylor , "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, Cambridge, England, 2000.