Hybridization of PSO-ABC Based Ensemble Classification Model for High Dimensional Medical Datasets

G. Lalitha Kumari [1]

Acharya Nagarjuna University
Guntur, Andhra Pradesh, India
e-mail : lalithajoy.nuthakki@gmail.com

N. Naga Malleswara Rao [2]

Department of IT
RVR & JC College of Engineering
Guntur, Andhra Pradesh, India
e-mail : nnmrao@rvrjc.ac.in

***Abstract* - As the size and dimensionality of microarray datasets increase, it is vital to select essential features for data classification. Traditionally, ranking and selection measures are used to select the essential features from the high dimensional feature space. However, these measures are used to improve the data classification rate with limited number of instances and features space. Feature selection is one of the challenging issues for microarray datasets due to noise, sparsity and missing values. Traditional feature selection models such as Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) are used to select the highly weighted features for data classification. But these models require high computational memory and time for data classification. In this paper, a hybrid PSO+ABC based feature selection model is designed and implemented on microarray disease datasets. Proposed hybrid feature selection model is applied on multiple classification models to improve the true positive rate and error rate for different dimensional datasets. Experimental results are simulated on different microarray disease dataset and these results proved that the hybrid feature selection model has high true positive rate and minimal mean squared error rate compared to the traditional models.**

Keywords - Artificial Bee Colony (ABC), Classification Algorithm, High Dimensional Data, particle Swarm Optimization(PSO).

I. INTRODUCTION

High dimensionality is a major issue in which a large number of training features are processed to perform the mathematical operations such as data transformation, data classification etc. Traditional dimensionality reduction techniques are used to reduce the number of dimensions using the static and specified number of features.

Particle swarm optimization (PSO) and artificial bee colony (ABC) are the feature selection models which are used to reduce the feature space using the optimization functions. Gene expression datasets are ever increasingly in the field of biomedical databases including gene classification, gene feature selection, and gene based disease prediction etc. However, analysis of gene classification possesses a major challenge due to the large number of gene attributes. Different techniques have been implemented for feature selection on microarray datasets, but these models are independent of data classification. Table 1, [1] describes the accuracy and the computational time on the microarray datasets on limited subset of features. The gene instances in the microarray dataset are represented in vector form with multi-dimensional space. These objects are further interpreted using the machine learning models for decision making process. Traditional gene classification models for gene prediction have limitations such as slow speed due to the complexity in algorithm and number of dimensions and high error rate. Most of the microarray datasets have thousands of gene sets which are associated in large gene expressions or patterns.

Ensemble classification models are used to predict the high dimensional features in the given training datasets with less error rate. Each ensemble learning model combines several base classifiers to improve the accuracy than its individual prediction rate.

Run	No. of Genes	Index no. of selected genes	Cross validation Acc.	*K* for KNN	Training Acc.	Test Acc.	Computing Time (Hrs.)
1	13	187,359,390,742,1051,1074,1374,1738,1841,1842,1947, 1956,1980	98.0159	6	98.4127 (62/63)	90 (18/20)	5.5859
2	13	107,123,246,518,714,742,1181,1206,1286,1343,1954,2097,2106	99.2063	4	100 (63/63)	90 (18/20)	2.8632
3	5	246,257,742,1076,2103,	93.2540	4	93.6508 (59/63)	90 (18/20)	2.8098
4	10	187,231,246,356,428,534,650,742,758,840	95.6349	6	95.2381 (60/63)	90 (18/20)	2.7993
5	6	742,1003,1386,2046,2099,2157	98.0159	4	100 (63/63)	100 (20/20)	2.7956
6	9	125,187,246,362,742,783,1496, 1764,2040	96.8254	3	96.8254 (61/63)	90 (18/20)	2.8111
7	7	742,1601,1645,1932,1955,2046,2144	93.2540	3	100 (63/63)	100 (20/20)	2.8202
8	7	84,251,742,1032,1074,1645,1783	94.4444	4	100 (63/63)	100 (20/20)	2.8098
9	9	255,694,806,1003,1117,1319,1477,2046,2113	96.8254	3	100 (63/63)	95 (19/20)	2.8017
10	6	246,841,1263,1531,1955, 1645	94.8413	3	96.8254 (61/63)	95 (19/20)	2.7981

Artificial neural network is typically having three layered architecture such as input layer, hidden layer and output layer for data classification. In the past few years

tremendous efforts have been made by scientists to improve the network architecture and learning models based on basic ANN. But the major problem in the neural network framework is selection of appropriate activation function using the logistic and hyperbolic functions. Since, the selection of proper activation function improves the classification rate on the high dimensional datasets.

To overcome the data dimensionality and error rate in the traditional feature selection and classification models, feature selection based ensemble classification is used to handle large and heterogeneous datasets. This optimal feature selection model is used to solve the complex patterns using the novel optimization functions. The main objective of this model is to improve the classification true positive rate by minimizing the error rate using the hybrid feature selection measure.

II. RELATED WORK

PSO [5] is an optimization algorithm initially designed for numerical optimization problems. An improved version of PSO called geometric PSO to virtually improve the search space of the traditional PSO model. Several binary versions of PSO can be used in the literature [6] for binary classification problems. The key issues of GPSO and BPSO is algorithm complexity on high dimensional datasets. Also the updation of particle position is initially performed for continuous datasets using the Euclidean, hamming and manhattan distance.

Chinnaswamyet.al,[6] proposed a novel intelligent model for disease classification. They have used three different methods fuzzy based KNN, back propagation and gradient descent. The extension of this method is developed in [7] to improve the convergence and divergence of the PSO using fuzzy.

Juang, et.al,[8] implemented a novel approach to detect the gene subset selection and classification of cancer patterns using the hybrid PSO and tabu search approach. They used statistical t rank test to find the top most gene sets from the training dataset. The results of this tabu search is optimized by li et.al,[9] using genetic algorithm and PSO. Li[9] used wilconxon ranking measure to find the relational gene sets on the training dataset.

Güneyet.al,[10] implemented a heuristic optimization model for PSO for micro array datasets. This model is applicable to limited feature space and limited number of instance in each training data.

Yijuanet.al,[11] suggested classification of microarray dataset and gene selection on the small training dataset. They used fisher's linear discriminate model for classification. The main limitation of this model is that they used limited gene sets for accuracy computation and error rate optimization.

Sun, et.al,[12] implemented a novel ensemble classification method on microarray dataset. They used pattern matching model to find and extract the essential patterns form the training micro array dataset.

Ngomliu et.al,[13] proposed a gene selection based ensemble classification model on micro array datasets. Finding essential and relevant gene from large number of micro array datasets is major issue on large datasets due to inter and intra cluster relationships and noise.

Yu et.al,[14] designed and implemented a genetic algorithm and independent component analysis based ensemble learning model on microarray dataset. This model is limited to small instances with high dimensional space. The complexity of the model depends on the training datasize and number of features in the feature selection process.

Jaison et.al,[15] developed a novel classification model on microarray dataset for disease prediction. They used a set of base classifiers such as naïve bayes, SVM, KNN and DWT for data classification. In the experimental results, they used a limited number of features in each disease dataset of microarray data.

III. A NEW PROPOSED MODEL

Proposed PSO based ensemble classifier is a multi-objective technique which finds the local and global optimum solutions by iteratively searching in a high dimensional space. Traditionally, as the size of the training dataset is small, medical disease prediction rate could be dramatically reduced due to class imbalance and high dimension space.

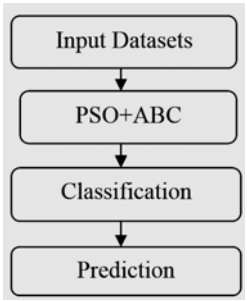

Fig. 1. The Proposed Model

Ant Bee Colony optimization is a population based heuristic technique and tries to optimize the initialization parameters of each particle in the high dimensional space. The execution process of the ABC includes three phases; employed bee phase, onlooker bee phase and scout bee phase. Here, number of onlooker bees equals to number of employed bees. Both the employed bees and onlooker bees are used to find the nectar sources equals to the employed bees. Proposed ABC based ensemble classifier is a multi-objective technique which finds the local and global optimum solutions by iteratively searching in a high dimensional space. Traditionally, as the size of the training dataset is small, medical disease prediction rate could be

dramatically reduced due to class imbalance and high dimension space.

Proposed PSO and ABC are integrated to improve the classification rate of the ensemble classification model with weak classifiers on high dimensional feature selection as shown in figure 1. Most of the ensemble classification technique is designed and implemented using the set of weak classifiers to optimize the overall classification rate and to minimize the error rate. Here, search space of the traditional PSO and ABC methods are optimized using the proposed optimized fitness measure and chaos gauss based randomization measure.

Step 1: Data pre-processing on Training high dimensional data.

Load dataset $D^1, D^2 \ldots D^n$

For each attribute Att(i) in the $D^1, D^2 \ldots D^n$ do

For each record value R(j) in the Att(i) do

if(isNum(R(j)) AND R(j)==null)

then

$$I(j) = \sum_{j=1/i \neq j}^{n} \frac{(I(j) - \mu_{A(I(j))}|)}{(Max_{A(I(j))} - Min_{A(I(j))})} \text{ -- (1)}$$

end if

if (isCat(A_i) && A_i(I)==null)

then

$$I(j) = PriorProb(A(i), class(m)); \text{ ---(2)}$$

Here, mth class of the missing value is used to find the prior probability in place of missing value

end if

End for.

A. Proposed PSO + ABC Algorithm

Step 1: Initializing particles with feature space, number of iterations, velocity, number of particles etc.

Step 2: For each Particle[i] do,

Compute hybrid velocity and position for each particle in'd dimensions using the following equations.

In this proposed PSO model, a random value between 0 to 1 is selected using the following equation as

$$R_i = \max\{\mu.\beta_j^k(1-\beta_j^k), \frac{1}{\sqrt{2\pi\sigma_X}} e^{-\frac{(X-\mu_X)^2}{\sigma_X^2}}\}$$

$$K = 1, 2 \ldots iterations$$

$$\beta_j^k \in (0,1)$$

$$v(d+1,i) = \psi.[\omega(d,i).v(d,i).\theta_{chaos1}(pBest_i - X(d,i)) + \theta_{chaos2}(gBest_i - X(d,i))]$$

$$X(d+1,i) = X(d,i) + v(d+1,i)$$

ψ is the convergence factor computed as

$$\psi = \frac{2*\eta}{|2 - (\theta_{chaos1} + \theta_{chaos2}) - \sqrt{(\theta_{chaos1} + \theta_{chaos2})^2 - 4(\theta_{chaos1} + \theta_{chaos2})}|}$$

where $\theta_{chaos1}, \theta_{chaos2} \in$ Ortho chaos gauss randomization, η is the value lies in [0,1].

In this optimized model, inertia weight is computed as:

$$\omega(d,i) = \omega_{max} - (I_{current} / I_{max}).(\omega_{max} - \omega_{min})$$

ω_{max} : max inertia

ω_{min} : min inertia

I_{max} : max iteration

step 3:if(f(pBest[i])>f(X(d+1,i)) then

pBest[i]=X(d+1,i);

end if

if(f(pBest[i])<f(gBest[i]))then

gBest[i]=min{f(pBest[i])}; i=1,2,3…n particles.

where f is the predefined benchmark objective functions.

end if

Step 4: Computing Fitness Value

Proposed feature selection fitness measure is given as

end for.

Step 5: Initialization of ABC Parameters

Initialization of ABC parameters Phi $\in$ [0,1], rho $\in$ [0,1],, alpha, beta, #iterations, initialize all bees solutions to false.

Initial optimal solutions are derived using the following equation.

$$\phi_i = \max\{\mu.\beta_j^k(1-\beta_j^k), \frac{1}{\sqrt{2\pi}\sigma_D} e^{-\frac{(X-\mu_D)^2}{\sigma_X^2}}\}$$

$K = 1,2 \ldots iterations$

$\beta_j^k \in (0,1)$

$$x(p,q) \leftarrow x_{min}^q + \phi(x_q^{max} - x_q^{min}), p = 12 \ldots N, q = 12 \ldots D$$

where x(p,q) is pth employed bee with qth dimension.

x_q^{max} is the upper bounds of qth dimension

x_q^{min} is the lower bounds of qth dimension

ϕ is the optimized random number from 0 to 1.

N total employed bees

D is the dimensionality

Step 6: To each particle[i] do,
Update the particle velocity and position using the following equations:

$$v(p,q) \leftarrow x(p,q) + \eta(x(p,q) - x(r,q)), p = 12 \ldots N, q = 12 \ldots D, p \neq r$$

where x(p,q) is pth employed bee with qth dimension.

$\eta = \phi.[x_q^{max} - x_q^{min}]$

x_q^{max} is the upper bounds of qth dimension

x_q^{min} is the lower bounds of qth dimension

ϕ is the optimized random number from 0 to 1.

N total employed bees

D is the feature space

Fitness value of the candidate solutions can be computed using

$$fit_i = \frac{1}{1 + fitFunc_i}; if(fitFunc_i >= 0)$$
$$= 1 + abs(fitFunc_i); otherwise$$

- **Onlooker Bee Phase**

In this phase, each onlooker-bee selects an employed_bee to optimize its feasible solution. This selection is performed using the fitness values of employed bees by roulette wheel as

$$\kappa_2[P_i] = Prob_i = \frac{fit_i}{\sum_{i=1}^{N} fit_i}$$

Where $Prob_i$ is the ith employee bee probability.

- **Scout Bee Phase**

In this phase, each bee which doesn't optimize their feasible solution until the maximum limit is reached becomes a scout bee.

Done

For each particle P[i] do

$if(\kappa_1[P_i] > \kappa_2[P_i])$

Then

Select features related to κ_1 for classification

Else

Select features related to κ_2 for classification

Done

Step 7: Update particle velocity, position, global best and particle best according to the fitness value conditions .
Step 8: This process is continuous until max iteration is reached.

B. Proposed Classification Model:

Input : Selected Feature weights in the hybrid PSO+ABC algorithm , Input training dataset, weak classifiers.
Output: Test data prediction
Procedure:
Step 1: Partition the dataset into k disjoint sets D_k.
Step 2. For each partition D_k
Do
Step 3: Apply base classifiers
C[]={"IFFNN","SVM","IDT","NAIVEBAYES"};

a) The improved activation function used in the FFNN is given below:

The activation function used in the IFFNN is the maximization of hyperbolic tangent function and tan sigmoid hyperbolic function as

$$Actfunct = f(x) = \max\{\frac{e^x - e^{-x}}{e^x + e^{-x}}, tansig(x)\}$$

$$\frac{df}{dx} = \frac{d}{dx}\frac{e^x - e^{-x}}{e^x + e^{-x}} dx$$
$$= \frac{d}{dx}\frac{sinh(x)}{cosh(x)}$$
$$= \frac{cosh^2(x) - sinh^2(x)}{cosh^2(x)}$$
$$= 1 - tanh^2(x) \in [-1, 1]$$

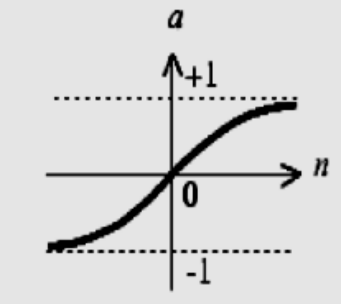

Fig.2 Hyperbolic tan sigmoid function

b) The enhanced decision tree technique implemented in the ensemble classification is the integration of Correlation, Information Gain, Gain Ratio, to evaluate attribute selection measure for the tree building and evaluation.

The computational measure used to evaluate the attributes in the training dataset is given as:

$$CFSM(D_i, D_{m(i!=m)}) = \sqrt[3]{\left(\sum_{i=1}^{|A_i|}\sum_{j=1}^{|A_m|}\left(\sqrt[3]{D_i/|D_i|} - \sqrt[3]{D_m/|D_m|}\right)^2\right)} \quad \text{--- (A)}$$

Where , D_i is the ith class instances. D_m is the remaining mth class instances.

$$HIFSM(A) = \frac{\sqrt[3]{GainRatio(A) * CFSM(D_i, D_{m(i!=m)})}}{\min\{IG(A), Corr(D_i, D_{m(i!=m)})\}} \quad \text{---(B)}$$

Step 4: Find the maximum of the voting in the classifiers for accuracy prediction:

MV=Max{TP{C{"IFFNN","SVM","IDT","NAIVEBAYES"}}}

IV. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed model to the existing models, different microarray datasets were selected from the biomedical repository. Different dataset used for experimental evaluation are summarized in Table 2. In the experimental results, 10% of the training data are used as testing data for performance evaluation. Proposed feature selection based ensemble methods increase the performance of true positive rate and accuracy on entire high dimensional datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models. From the experimental results, it is clear that hybrid PSO+ABE based ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

TABLE I. DATASETS AND ITS CHARACTERISTICS

Micro array Datasets	Gene sets	Data-Type
lung-Michigan	7000	Continuous/Numeric
lungCancer_train	12000	Continuous/Numeric
DLBCL-Stanford	4000	Continuous/Numeric
Breast cancer	24481	Continuous/Numeric
Leukemia	7129	Continuous/Numeric

Proposed model increase the performance of true positive rate and accuracy on entire high dimensional microarray datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models.

True negative measure evaluates the ratio of number of instances that are not affected with disease, which are identified correctly.

True positive measure defines the ratio of number of disease instances that have been predicted as positive rate.

Mean Absolute error: Average misclassification rate of each test data in the cross validation.

Precision measure computes the ratio of correctly predicted cancer instances among the entire disease affected instances.

TABLE II. PERFORMANCE ANALYSIS OF PROPOSED MODEL TO THE EXISTING MODELS IN TERMS OF TRUE POSITIVE RATE AND PRECISION ON CANCER DATASETS.

Avg performance of all Cancer datasets		
Model	**TruePositive**	**Precision**
PSO+Ensemble	**0.8746**	**0.8636**
GPSO+Ensemble	**0.9253**	**0.9163**
Fuzzy PSO+Ensemble	**0.9185**	**0.9153**
PSO+Ensemble	**0.9674**	**0.9609**
ABC+Ensemble	**0.9725**	**0.9783**
ABC+PSO+Ensemble	**0.9857**	**0.9818**

Table II describes the performance of the proposed model on all cancer datasets. Here, all the cancer datasets are evaluated using the proposed model to find the average true positive rate and precision rate on the high dimensional datasets. From the table, it is visualized that the proposed model has high true positive rate and precision over the existing models.

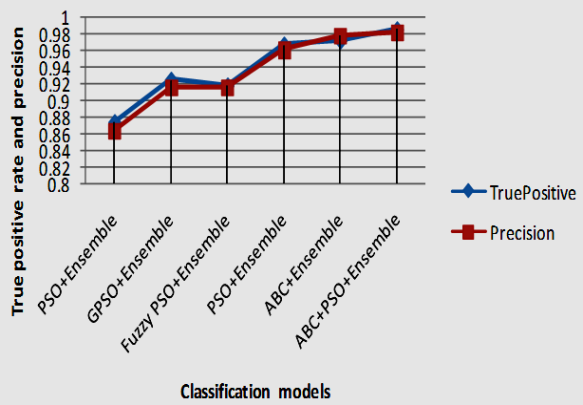

Fig. 3 TP and error rate comparison of proposed model to the existing models

TABLE III. MEAN ERROR RATE AND RUNTIME OF ALL CANCER DATASETS

Model	Error	Runtime (ms)
PSO+Ensemble	0.1956	6963
GPSO+Ensemble	0.0895	6835
Fuzzy PSO+Ensemble	0.0815	6946
PSO+Ensemble	0.0392	5793
ABC+Ensemble	0.0253	5195
ABC+PSO+Ensemble	0.0153	4976

Table III describes the comparison of average mean error rate and runtime of all the cancer datasets. From the table, it is noted that the proposed model has low error rate and runtime compared to the existing models.

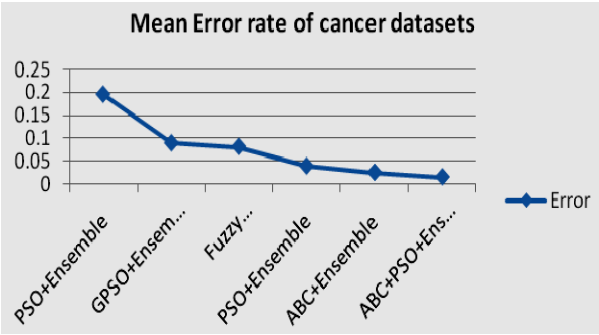

Fig. 4 Average mean error rate of all Cancer datasets

Figure 4, illustrates the average mean error rate of all the microarray cancer datasets with all feature space. From the figure, it is clear that the proposed model has the low error rate compared to the existing models.

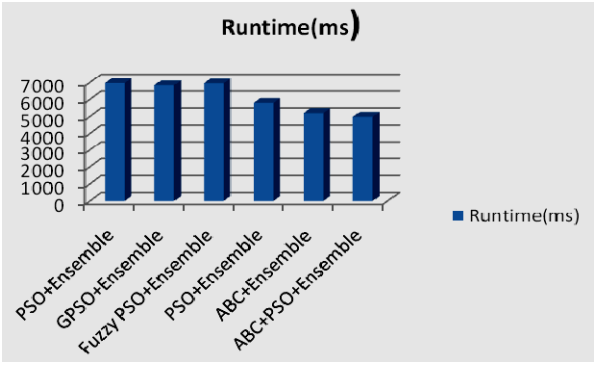

Fig. 5: Avg Runtime (ms) of all Cancer datasets

V. CONCLUSION

In this paper, an integrated PSO and ABC based feature selection method is developed for ensemble classification model on high dimensional datasets. In the traditional models, features are selected in static way or fixed number due to memory and time computation. Traditional PSO based ensemble learning and ABC based ensemble learning are improved using the heuristic activation function and ensemble classification measures. Proposed hybrid feature selection model is applied on multiple classification models to improve the true positive rate and error rate for different dimensional datasets. Experimental results are simulated on different microarray disease dataset and these results proved that the hybrid feature selection model has high true positive rate and minimal mean squared error rate compared to the traditional models.

REFERENCES

[1] Kar, S., et al. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications (2014).

[2] P. Mohapatra and S. Chakravarty, "Modified PSO based feature selection for Microarray data classification," 2015 IEEE Power, Communication and Information Technology Conference (PCITC), Bhubaneswar, 2015, pp. 703-709.

[3] P.Maji and C. Das " Relavant and Significant Supervised Gene Clusters for Microarray Cancer Classification ," in IEEE Transaction on NanoBioSciences , Vol.11, No.2,pp.161-168, June

[4] R. Nicole, I. Saha, "Use of multi objective differential fuzzy Clustering with ANN classifier for unsupervised pattern Classification: Application to microarray analysis," 2010 Fourth International Conference on Research Challenges in Information Science (RCIS), Nice, France, 2010, pp. 647-656.

[5] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino and M. Buckland, "A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis," in IEEE Access, vol. 4, pp. 9145-9154, 2016.

[6] E. Alba, J. Garcia-Nieto, L. Jourdan and E. G. Talbi, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," 2007 IEEE Congress on Evolutionary Computation, Singapore, 2007, pp. 284-290.

[7] A. Chinnaswamy and R. Srinivasan, "Hybrid information gain based fuzzy roughset feature selection in cancer microarray data," 2017, Innovations in Power and Advanced Computing Technologies (i- PACT), Vellore, 2017, pp. 1- f
[8] Juang, C. A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), 34(2), pp.997- 1006.
[9] X. Li and M. Yin, "Multiobjective Binary Biogeography Based Optimization for Feature Selection Using Gene Expression Data," IEEE Transactions on NanoBioscience, vol. 12, no. 4, pp. 343-353, Dec. 2013.
[10] H. Güney and H. Öztoprak, "Microarray-based cancer diagnosis: repeated cross- validation-based ensemble feature selection," in Electronics Letters, vol. 54, no. 5, pp. 272-274, 3 8 2018.
[11] Yijuan Lu, Qi Tian, M. Sanchez and Yufeng Wang, "Hybrid PCA and LDA Analysis of Microarray Gene Expression Data," 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2005, pp. 1-6.
[12] Edraki, M., Haghshenas, M., Yektatalab, S., Dehghani, S.M., Haghpanah, S. "The effect of educational-psychological interventions on the resilience of the mothers of 1-to-3-year-old children with chronic liver disease leading to cirrhosis (2018)",International Journal of Pharmaceutical Research, 10 (2), pp. 222-226.
[13] Y. Li and A. Ngom, "Nonnegative Least-Squares Methods for the Classification of High- Dimensional Biological Data," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 2, pp. 447-456, March-April 2013.
[14] Z. Yu, H. S. Wongb, J. You, Q. Yang and H. Liao, "Knowledge BasedCluster Ensemble for Cancer Discovery From Bio-molecular Data," in IEEE Transactions on Nano Bioscience, vol. 10, no. 2, pp. 76-85, June 2011.
[15] Jaison, B., Chilambuchelvan, A. & Mohamed Junaid, K.A. Natl. Acad. Sci. Lett. (2015) 38:415.

BIBLOGRAPHY

1. Lalitha Kumari Gaddala B.Tech, M.Tech, has a work experience of over 13 years. She is pursuing a Ph.D in Computer Science and Engineering at Acharya Nagarjuna University, Guntur and Andhra Pradesh. She is Senior Assistant Professor in Prasad V Potluri Siddhartha Institute of Technology, Vijayawada. Her research interests are Soft Computing-Optimization algorithms and Machine Learning. She has published several papers in international conferences and international Journals.

2. Dr. N. Naga Malleswara Rao, B.Tech, M.Tech and Ph.D, has a work experience of 27 years. He is Professor in the Department of IT, RVR & JC College of Engineering, Guntur (Dt). His research interests are Computer Algorithms, Compilers and Image Processing. He has published few papers in international conferences, National Conferences and international Journals.