Advanced Approach for Classifying DNA Microarray Data using Machine Learning

[1]Ankita Choudhary, [2]Prof. Niti Shah
[1]Student of Master of Engineering, [2]Assistant Professor
Computer Engineering Department,
Silver Oak College Of Engineering & Technology, Ahmedabad, India.

Abstract : Now a days Ensemble classification has been a frequent topic of research, especially in bioinformatics and especially true for DNA microarray data experiment and large levels of noise inherent in data. DNA Microarray data is a high-dimensional data that enables the researchers to analyze the expression of many genes in a single reaction quickly and in an efficient manner. Microarray's characteristic such as small sample size, class imbalance, and data complexity causes difficulty to classify the diseases. The information obtained from the analysis of DNA microarray data is relevant to identify and predict illness, improve treatment and determine which genes are responsible to provoke a specific disease. This research aims to improve the accuracy of the classification algorithm by working on implementing an efficient Algorithm using Ensemble Classification for classifying the type of diseases using DNA microarrays data.

Keywords **-DNA microarray, Principal Component Analysis, Feature Selection, Classification Technique, ANN.**

I. INTRODUCTION

Data mining is one of the newest analytical methods that have been used to serve medical science research and has been shown to be a valid, sensitive, and reliable method to discover patterns and relationships. It involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. While data mining represents a significant advance in the type of analytical tools currently available, medical research studies have benefitted from its application in many areas of interest. [10] An important problem in deoxyribonucleic acid (DNA) microarray experiments is the classification of biological samples using gene expression data. To date, this problem has received the most attention in the context of cancer research; we thus begin this work with a review of disease classification using microarray gene expression data. A reliable and precise classification of disease is essential for successful diagnosis and treatment of it. Current methods for classifying human malignancies rely on a variety of clinical, morphological, and molecular variables. [10]

II. MICROARRAY

A microarray is a multiplex lab-on-a-chip. It is a 2D array on a solid substrate (usually a glass slide or silicon thin-film cell) that tests a large amounts of biological material using high-throughput screening multiplexed, parallel processing and detection methods. "Microarray" has become a general term; there are many types of microarray [12] -

- DNA microarrays
- Protein microarrays
- Antibody microarray
- Chemical compound microarray

Microarray Steps-

1. Experiment and Data Acquisition
2. Sample preparation and labeling
3. Hybridization
4. Washing
5. Image acquisition
6. Data normalization
7. Data analysis
8. Biological interpretation

III. DNA MICROARRAY

A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. DNA microarray also known as biochip or DNA chip. DNA microarray classification is a technique widely applied to discover valuable information about diseases. i.e. Cancer. It allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression). Each DNA spot contains picomoles (10^{-12} $moles$) of a specific DNA sequence, known as probes. [12]

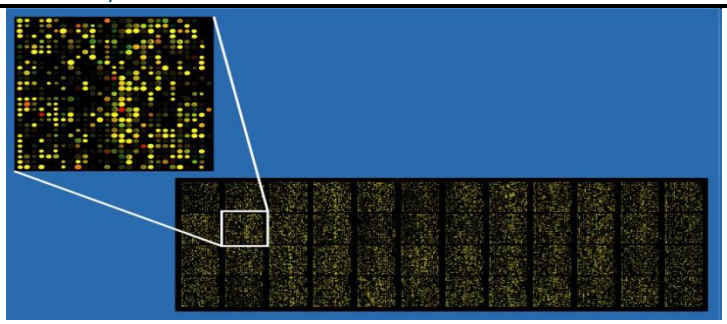

figure 1. Biological Samples in 2D Arrays on Membrane [12]

GREEN represents **Control DNA**, where either DNA or cDNA derived from normal tissue is hybridized to the target DNA.
RED represents **Sample DNA**, where either DNA or cDNA is derived from diseased tissue hybridized to the target DNA.
YELLOW represents **a combination of Control and Sample DNA**, where both hybridized equally to the target DNA.
BLACK represents areas where **neither the Control nor Sample DNA** hybridized to the target DNA.

IV. PROPOSED SYSTEM

A. Overview

Our Proposed system is based on the Feature extraction and Feature selection method. It consist of mainly following steps such as fetching the microarray dataset, preprocessing the dataset to get the relevant data, feature extracting, feature selecting, classifying the data. The proposed flow diagram is shown in the figure 2. Firstly the Microarray data is load into the dataset. That dataset is divided into the chunks for processing the further process on them. Then the feature extraction method uses the SVD, eigen vector and PCA for extracting the features from the chunks. Then ABC algorithm is used for selecting the most relevant features and then the hybrid classification algorithm SVM with ANN using feed forwarding neural network is used for classifying the DNA microarray data in Leukemia diseases as AML or ALL diseases.

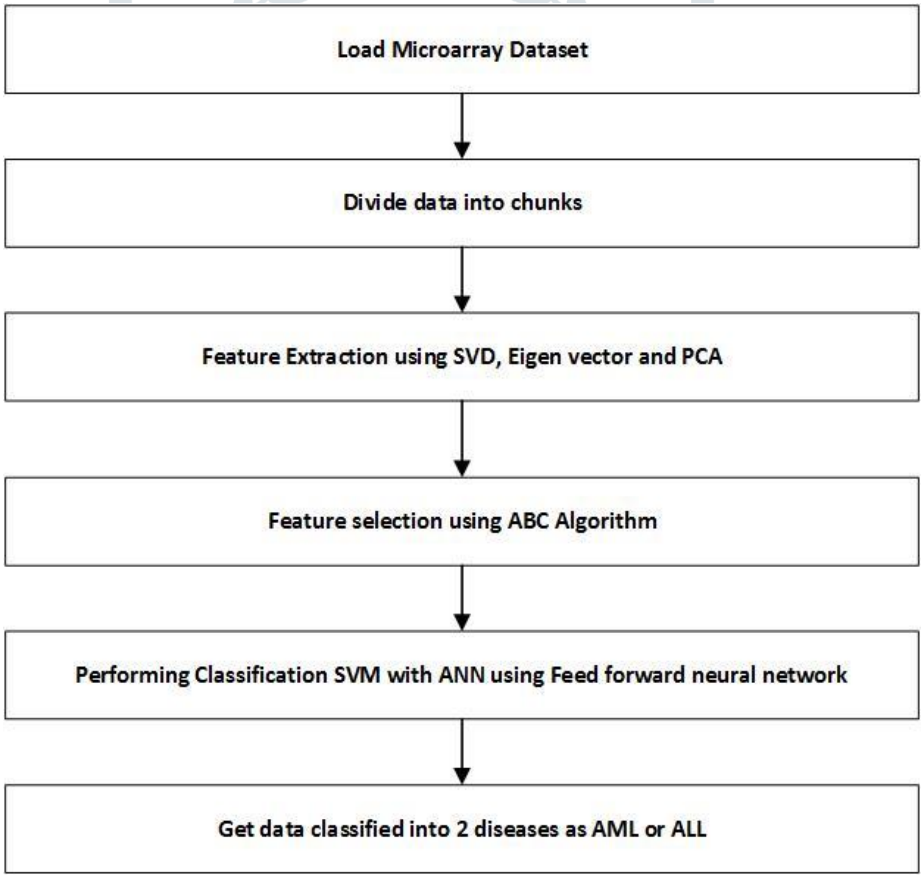

figure 2. Schematic description of the proposed methodology

B. Loading the microarray dataset

Leukemia dataset is taken from a collection of leukemia patient samples in the form of text file. It contains gene expressions of AML samples and ALL samples.The dataset consisted of 93 samples. And each sample is measured over thousand number of

genes. Last column reports the sample's class label (AML or ALL). Gene expression sample is in the form of integer/ pixel value (Range from -32,768 to +32,767).

```
-95,-118,59,270,-229,-383,172,-187,185,157,ALL
-12,-172,12,172,-137,-205,358,-104,-25,147,ALL
-21,-13,8,38,-128,-245,409,-102,85,281,AML
-202,-274,59,309,-456,-581,-159,-343,236,-7,AML
-112,-185,24,170,-197,-400,-215,-227,100,307,AML
-118,-142,212,314,-401,-452,-336,-310,177,-131,AML
-90,-87,102,319,-283,-385,-726,-271,-12,-104,AML
-137,-51,-82,178,-135,-320,-13,-11,112,-176,AML
-157,-370,-77,340,-438,-364,-216,-210,-86,253,AML
-172,-122,38,31,-201,-226,242,-117,-6,179,AML
-47,-442,-21,396,-351,-394,236,-39,95,203,AML
-62,-198,-5,141,-256,-206,-298,-218,-14,100,AML
-58,-217,63,95,-191,-230,-86,-152,-6,-249,AML
-161,-215,-46,146,-172,-596,-122,-341,171,-147,AML
```

figure 3. Dataset in text file

C. Features Extracting

For extracting the features from the training dataset SVD & PCA is applied. SVD is the singular value decomposition used as means of decomposing as a matrix into a product of 3 simpler matrices. PCA is a classical statistical method for transforming attributes of a dataset into a new set of uncorrelated attributes called principal components (PCs). PCA can be used to reduce the dimensionality of a dataset, while still retaining as much of the *variability* of the dataset as possible. There are many examples of the use of machine learning to classify high dimensional data, such as gene-expression microarray data[16].

D. Feature Selecting

The DNA microarray has millions of representatives genes, but the question that arises is: which of them contribute to a certain disease or are targets of mutations? To obtain this representative set, it is necessary to eliminate the irrelevant genes and obtain the set of genes whose expression level confirm a connection with a specific disease.
The Artificial Bee Colony (ABC) algorithm is applied to find the best set of genes. The ABC algorithm is a popular optimization technique based on the metaphor of the bees foraging behavior [1]. The population of bees represent solutions in a search space. This algorithm defines three types of bees: employed, onlookers and scouts. The ABC generates a randomly distributed initial population of SN solutions (food sources),

V. Experiments and Results

The proposed algorithm is implemented on the MATLAB (R2015a) platform, and the CPU is the machine of Intel core i5-5200U CPU @ 2.20GHz having the RAM configuration of 8 GB memory. The algorithm is evaluated using the publically available dataset. Firstly we have evaluated using the Leukemia dataset. This dataset contains the several no. of genes expressions and the disease name. The dataset is tested using the training dataset and the testing dataset.

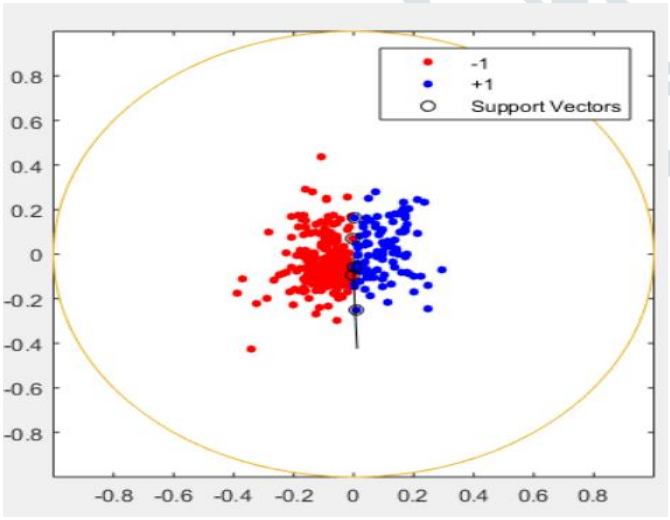

Figure 4. Training dataset
(Result after data partitioning)

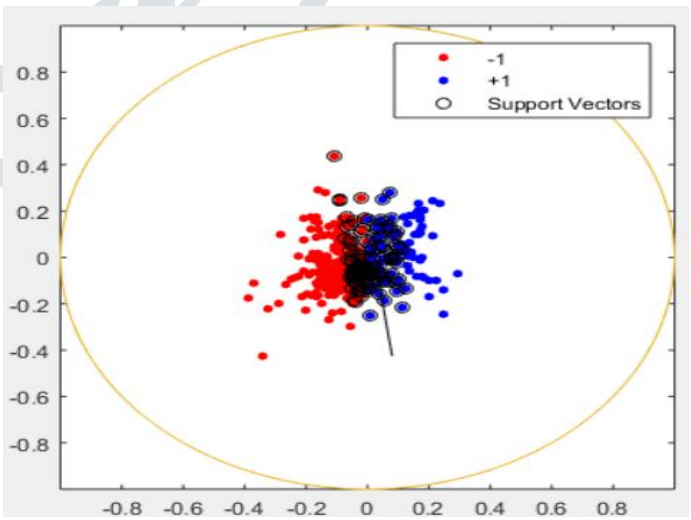

figure 5.Features extracted using svd & PCA

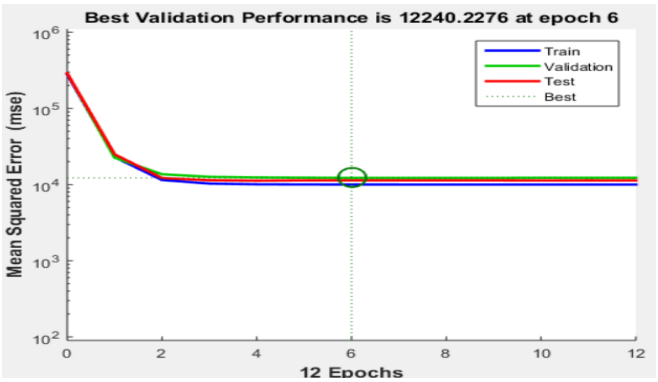

figure 6. Validation Performance

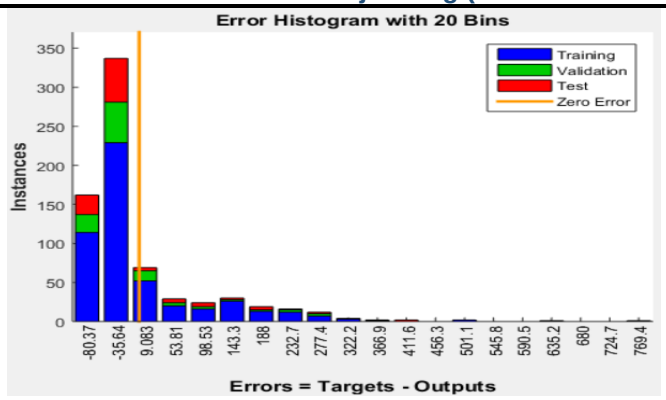

figure 7. Error Histogram Graph

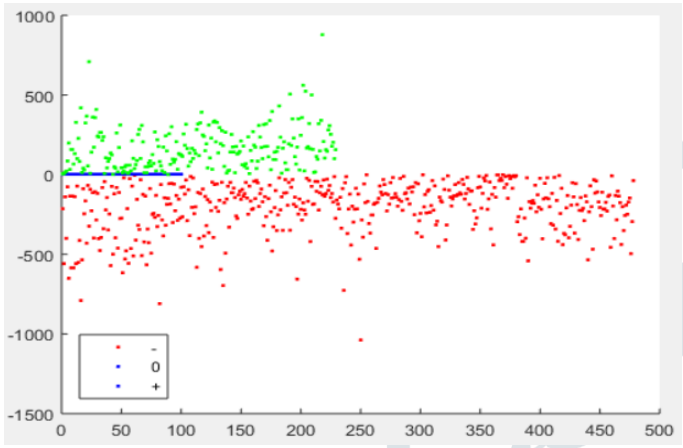

figure 8. Value Matrix

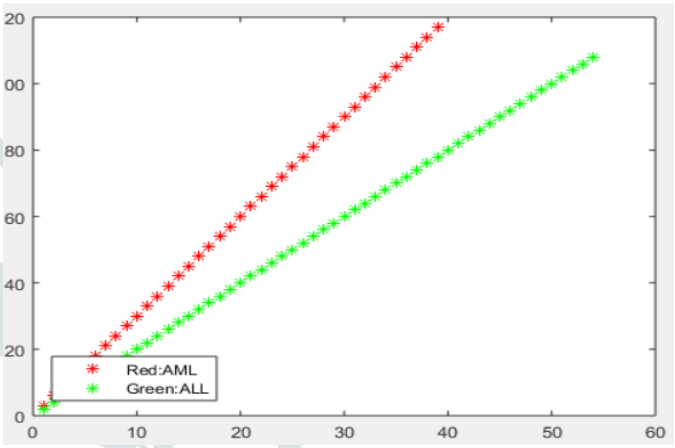

figure 9. AML & ALL count

The performance of the proposed algorithm is evaluated in terms of accuracy. The proposed method and the existing methods namely SVM and ANN are experimented with the same data set and their performance were compared in terms of accuracy. The following formula is used to measure accuracy,

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where, True positive (TP) – correctly predicts the positive class
True negative (TN) – correctly predicts the negative class
False positive (FP) – incorrectly predicts the positive class
False negative (FN) – incorrectly predicts the negative class

Table 1 Comparison with other Classification Techniques

Method	Accuracy (%)	References
SVM	93.10	[1]
ANN	91.20	[1]
Proposed Method	98.73	Proposed methodology

VI. CONCLUSION

In this paper, the described approach is for classifying DNA microarrays data based on the hybrid classification method. The feature selection task was carried out by means of ABC algorithm and for extracting the features SVD , eigen vector and PCA was applied for designing the SVM with ANN including feed forward neural network for classifying the DNA microarrays.
As a result, the proposed methodology reached an accuracy of 98.73%.

References

[1] Beatriz A. Garro and Katya Rodriguez Ciudad Universitaria , Mexico, D.F. "Designing artificial neural networks using differential evolution for classifying DNA microarrays" 978-1-5090-4601-0/17/$31.00 © 2017 IEEE

[2] Amina Houari, Wassim Ayadi, Sadok Ben Yahia University of Tunis, Tunisia "NBF: An FCA- based Algorithm to Identify Negative Correlation Biclusters of DNA Microarray Data" 1550-445X/18/$31.00 ©2018 IEEE

[3] Hongfei Wang, Ziadong Lv Wuhan, China "Mining Raw Gene Expression Microarray Data for Analyzing Synchronous and Metachronous Liver Metastatic Lesions from Colorectal Cancer" 978-1-5090-3710-0/16/$31.00 ©2016 IEEE

[4] Hongfei Wang, Wenjie Cai Wuhan, China "DNA Probe Signal Processing for Identification of Abnormal Gene Regulation and pathogenetic " 978-1-5090-1345-7/16 $31.00 © 2016 IEEE

[5] Taghi M. Khoshgoftaar, David J. Dittman, Randall Wald, Wael Awada Florida Atlantic university "A Review of Ensemble Classification for DNA Microarrays Data" 1082-3409/13 $31.00 © 2013IEEE

[6] Rupsa Bhattacharjee , Dr. Monisha Chakraborty West Bengal , India "LPG-PCA Algorithm and selective Thresholding based Automated Method: ALL & AML Blast Cells Detection and Counting" 978-1-4673-4700-6/12/$31.00 © January 2012 IEEE

[7] Huang-Cheng Kuo and Pei- Cheng Tsai "Mining Time-delayed Gene Regulation Patterns from Gene Expression Data" ©2012 GSTF

[8] Erin J.Moore, Thirmachos Bouriai Member , IEEE "Expectation Maximization of Frequent Patterns, a Specific, Local, Pattern-Based Biclustering Algorithm for Biological Datasets" 1545-5963 ©2015 IEEE

[9] Wildan Andaru, Iwan Syarif, Ali Ridho Barakbah Subrabaya, Indonesia "Feature Selection Software Development Using Artificial Bee Colony On DNA Microarray Data" 978-1-5386-0716-9/17/$31.00 ©2017 IEEE

[10] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," San Fr. CA, itd Morgan Kaufmann, p. 745, 2012.

[11] L. J. Lancashire, C. Lemetre, and G. R. Ball, "An introduction to artificial neural networks in bioinformatics application to complex microarray and mass spectrometry datasets in cancer studies," Briefings in Bioinformatics, vol. 10, no. 3, pp. 315–329, 2009.

[12] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nat Med, vol. 7, no. 6, pp. 673–679, 2001.

[13] W. Chen, H. Lu, M. Wang, and C. Fang, "Gene expression data classification using artificial neural network ensembles based on samples filtering," in Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on, vol. 1, Nov 2009, pp. 626–628.

[14] "Classification of DNA microarrays using artificial bee colony (ABC) algorithm," in Advances in Swarm Intelligence - ICSI 2014, Proceedings, Part I, ser. LNCS, Y. Tan, Y. Shi, and C. A. C. Coello, Eds., vol. 8794. Springer, 2014, pp. 207–214. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11857-4

[15] B. A. Garro, K. Rodriguez, and R. A. Vazquez, "Classification of {DNA} microarrays using artificial neural networks and (ABC) algorithm," Applied Soft Computing, vol. 38, pp. 548 – 560, 2016.

[16] Tom Howley, Michael G. Madden, Marie-Louise O'Connell and Alan G. Ryder, "The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data", National University of Ireland, Galway