Research Journal of Applied Sciences, Engineering and Technology 10(2): 197-205, 2015
DOI:10.19026/rjaset.10.2572
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 22, 2014 Accepted: December 18, 2014 Published: May 20, 2015

**Research Article**

# A Hybrid Classifier for Leukemia Gene Expression Data

[1]S. Jacophine Susmi, [1]H. Khanna Nehemiah, [2]A. Kannan and [1]J. Jabez Christopher
[1]Ramanujan Computing Centre, Anna University, Chennai, 600025, India
[2]Department of Information Science and Technology, Anna University, Chennai, 600025, India

**Abstract:** In this study, a hybrid technique is designed for classification of leukemia gene data by combining two classifiers namely, Input Discretized Neural Network (IDNN) and Genetic Algorithm-based Neural Network (GANN). The leukemia microarray gene expression data is preprocessed using probabilistic principal component analysis for dimension reduction. The dimension reduced data is subjected to two classifiers: first, an input discretized neural network and second, genetic algorithm-based neural network. In input discretized neural network, fuzzy logic is used to discretize the gene data using linguistic labels. The discretized input is used to train the neural network. The genetic algorithm-based neural network involves feature selection. The subset of genes is selected by evaluating fitness for each chromosome (solution). The subset of features with maximum fitness is used to train the neural network. The hybrid classifier designed, is experimented with the test data by subjecting it to both the trained neural networks simultaneously. The hybrid classifier employs a distance based classification that utilizes a mathematical model to predict the class type. The model utilizes the output values of IDNN and GANN with respect to the distances between the output and the median threshold, thereby predicting the class type. The performance of the hybrid classifier is compared with existing classification techniques such as neural network classifier, input discretized neural network and genetic algorithm-based neural network. The comparative result shows that the hybrid classifier technique obtains accuracy rate of 88.23% for leukemia gene data.

**Keywords:** Classification, fuzzy logic, genetic algorithm, microarray gene expression, neural network, PPCA

## INTRODUCTION

With the continuous development of database technology and the extensive applications of database management system, the volume of data stored in database is increasing rapidly and much important information lies hidden in such large amounts of data. Data mining is the process of finding and extracting frequent patterns that can describe the data, or predict unknown or future values (Nittaya and Kittisak, 2007; Brintha and Bhuvaneswari, 2012). Data mining involves several tasks namely classification, estimation, prediction, affinity grouping, clustering, association rule mining and description. Classification is one of the most common data mining tasks and it involves examining the features of a newly presented object in order to assign it to one of the predefined set of classes (Labib and Malek, 2005).

Data mining is utilized in various fields and biological data mining is one of the emerging fields of research and development. One of the important problems in bioinformatics is genomic data mining and knowledge extraction (Shreyas *et al*., 2007). A gene is the molecular unit of heredity of a living organism, which are made up of DNA which contains the formula for the chemical composition of one particular protein. The genetic molecule produces proteins and some additional products and messenger RNA (mRNA) is the first intermediate during the production of any genetically encoded molecule (Anandhavalli, 2008). The genomic information is usually represented by sequences of nucleotide symbols in the strands of DNA molecules, by symbolic codons (triplets of nucleotides), or by symbolic sequences of amino acids in the corresponding polypeptide chains (Anibal *et al*., 2007).

Genes can be extracted from human blood or tissue samples. The extraction process of genes involve the following steps: Initially, microarray slide is obtained containing sequences representing each gene. Microchips of gene expression have made it possible to simultaneously monitor the expression levels of thousands of genes under different experimental conditions (Seo *et al*., 2005; Bose *et al*., 2013; Bidaut *et al*., 2006). Each sample contains thousands of different mRNA sequences representing all of the genes expressed in those cells (Bose *et al*., 2013). In a microarray experiment, messenger RNA (mRNA) present in a cell is extracted. Fluorescent labeled complementary DNA copies of this mRNA are prepared. This cDNA from each sample of mRNA will be labeled with different fluorescent nucleotides. The microarray slide with both of these labeled cDNAs is

**Corresponding Author:** H. Khanna Nehemiah, Ramanujan Computing Centre, Anna University, Chennai, 600025, India

hybridized. Each cDNA will bind to the spots that have complementary sequences. Stringent conditions are used to ensure that the probe sequences are entirely complementary to the microarray spot sequences. The slide is washed to remove excess fluorescent cDNA not bound to spots. The microarray is read using an instrument that measures the fluorescence of each spot at two different wavelengths. Finally, the data is analyzed to determine gene expressions in each cell sample. The microarray chip used collectively reacts to changes in their environments, providing hints about the structures of the involved gene networks (Xu *et al*., 2001). From the expressions obtained the diagnosis of disease can be well established (Qi *et al*., 2009; Huynh *et al*., 2009).

The hypothesis that many or all human diseases may be accompanied by specific changes in gene expression has generated much interest among the bioinformatics community in classification of patient samples based on gene expression for disease diagnosis and treatment (Agrawal and Bala, 2007). An important application of gene expression microarray data is classification of biological samples or prediction of clinical outcomes (Dai *et al*., 2006; Zhou *et al*., 2004). Several combinations of the preprocessing algorithms, feature selection techniques and classifiers can be applied to the data classification tasks (Essam, 2010). The accuracy of the classification model depends strongly on the input data, which is transformed into a feature vector containing a number of features that predicts the output (Jing *et al*., 2010). In this study, we address the classification of microarray gene expression data for cancer diagnosis on Leukemia dataset.

## LITERATURE REVIEW

Peng *et al*. (2007) have proposed a hybrid approach which combines filter and wrapper methods, in which they have used the feature estimation from the filter as the heuristic information for the wrapper. In the first step, a filter gene selection method has been employed to eliminate the irrelevant genes and then a wrapper method has been applied to reduce redundancy. This hybrid approach takes advantages of both, filter methods' efficiency and wrapper methods' high accuracy. The comparative study on gene selection method has shown that Fisher's ratio is a relatively simple and straightforward method than other filter methods. Therefore, the Fisher's ratio has been used as the first step to remove the irrelevant features. In addition, the hybrid approach that combines Fisher's ratio and wrapper method could reduce the effect of the over fitting problem and achieve the goal of maximum relevance with minimum redundancy. With these advantages, the hybrid approach outperforms Fisher's ratio filter method when tested with leukemia dataset by achieving 98.61% accuracy when 3 genes were selected. The hybrid approach also outperforms when tested with breast cancer dataset by achieving 80.41% accuracy when 4 genes were selected and the same achieves 83.51% accuracy when 10 genes were selected.

Zhang *et al*. (2007) have proposed a fast and efficient classification method based on microarray data called the Extreme Learning Machine (ELM) algorithm for a multi-category cancer diagnosis problem. Its performance has been compared with other methods such as the Artificial Neural Network (ANN), Subsequent Artificial Neural Network (SANN) and Support Vector Machine One-Versus-All (SVM-OVA) and Support Vector Machine one-versus-one (SVM-OVO). This has inevitably involved more classifiers, greater system complexities and computational burden and a longer training time. ELM has been capable of performing multi-category classification directly, without any modification. Study results have been consistent with their hypothesis that ELM algorithm achieves higher classification accuracy. The approach also uses a smaller network structure that requires less training time than other algorithms. It has also been confirmed from the results on three microarray datasets GCM data set, Lung data set and Lymphoma dataset that ELM achieves 74, 85 and 97%, respectively.

Alok and Kuldip (2008) have proposed the use of the Linear Discriminant Analysis (LDA) technique to reduce the dimensionality of the feature space for cancer classification using microarray gene expression data. Due to small sample size problem, the conventional LDA technique could not be applied directly on the microarray data. In order to overcome this limitation, the dimensionality of the feature space is reduced through feature selection. LDA is a technique used for feature extraction-based dimensionality reduction. The GLDA technique utilizes gradient descent algorithm to do dimensionality reduction. Once the dimension is reduced through the GLDA algorithm, the k-nearest neighbour classifier is used to classify a tissue sample. The GLDA technique has been applied to three different microarray datasets namely ALL/AML dataset (http://www.broadinstitute.org/cancer/software/genepattern/datasets), SRBCT dataset and Lung adenocarcinoma dataset that shows lower misclassification rate. The GLDA technique compared to other algorithms obtains two misclassification samples whereas other methods Naïve Bayes, Decision Tree and SVM-OVO obtain 6, 7 and 4 misclassifications, respectively.

Huynh *et al*. (2009) have proposed an application of the Single Layer Feed Forward Neural Network (SLFN) trained by the Singular Value Decomposition (SVD) approach for DNA microarray classification. Many non-iterative training algorithms for the single hidden layer feed forward neural networks were compared for DNA microarray classification; they were Extreme Learning Machine (ELM), Regularized Least

Squares Extreme Learning Machine (RLS-ELM) and SVD approach. Also, the Back-propagation algorithm which is the well-known gradient-descent based iterative training method was evaluated and compared in terms of the number of hidden nodes and classification accuracy on test datasets. In SVD neural classifier, the SLFN has activation function and the parameters of the classifier are determined by SVD approach. The architecture consists of P nodes in input layer, N nodes in hidden layer and C nodes in output layer. SVD approach, RLS-ELM and Back-propagation algorithms require the same number of hidden nodes, while ELM needs more hidden nodes. For classification accuracy, SVD-approach and RLS-ELM algorithms are comparable to each other, while better than ELM and Back-propagation algorithm. Data sets used for this study were two binary cancer data sets of DNA microarray: Leukemia and colon cancers. Experimental results have shown that the SVD trained feed forward neural network obtains 95.93% of classification accuracy for leukemia dataset and 83.63% for colon datasets.

Kanthida *et al*. (2009) have discussed that the availability of high dimensional biological datasets such as gene expression, proteomic and metabolic experiments could be leveraged for the diagnosis and prognosis of diseases. Many classification methods predict diseased patients and healthy patients. However, existing researchers have focused only on a specific dataset. There has been a lack of generic comparison between classifiers, which might provide a guideline for biologists or bio-informaticians to select the proper algorithm for new datasets. They have compared the performance of popular classifiers, which are Support Vector Machine (SVM), Logistic Regression, k-Nearest Neighbor (k-NN), Naive Bayes, Decision Tree and Random Forest with small and a high dimensional synthetic dataset. The small dimensional dataset comprises 100 features whereas the high dimensional dataset comprises 1000. Both datasets are dichotomous. The experimental result has shown that SVM yields a classification accuracy of 95%.

Chien-Pang *et al*. (2011) have proposed a method for gene selection by dimension reduction on gene expression data. An adaptive genetic algorithm and k-nearest neighbor were used to evolve gene subsets. The proposed system reduces the dimension of the data set and classifies samples accordingly. The experimental results compare the performance of AGA/kNN and GA/kNN with colon data set and mice apo AI data. The results have shown that AGA/kNN reports 90% of accuracy after 40 runs and for 70 runs accuracy increased to 100% with set of 20 genes selected. In contrast, accuracy of GA/kNN obtains 80% for 100 runs.

Kaur and Raghava (2003) have used a Neural Network (NN) to predict gamma turns of proteins in two steps. First, sequence-to-structure network is used to predict the gamma turns from multiple alignment of protein sequence and in the second step, it uses a structure-to-structure network in which input consists of predicted gamma turns obtained from the first step and predicted secondary structure obtained from PSIPRED. From this we infer that the neural network can be trained and used for intermediate steps and not only for the task of classification.

Benítez *et al*. (1997) have proposed a novel method of knowledge acquisition from NN. The cryptic representation of the neural networks is interpreted into human-friendly rules. They have used the Fuzzy Rule Based System (FRBS) to convert the NN interpretation into comprehensible fuzzy rules. Since the rules extracted from the neural network are human-readable, they have concluded that neural networks are not to be considered as black boxes.

Comparing to the works discussed in the literature, the work presented differs in the following ways:

The novelty of the hybrid classifier is the framework that combines the output of two classifiers, Input Discretized Neural Network (IDNN) and Genetic Algo rithm based-Neural Network (GANN). The two classifiers IDNN and GANN are constructed independently. The output of each classifier is taken separately and combined within the framework to obtain final decision in microarray data analysis. The idea of combining multiple classifiers output, results in increasing the accuracy of hybrid classifier (Sung-Bae, 2002). The system framework differs from the existing ones in the way of hybridization. Rather than using three neural networks (Sung-Bae, 2002) this framework makes use of two neural networks. Further, in existing work there are techniques that decides the final classification output using voting technique, fusion technique (Sung-Bae, 2002) whereas, in this designed hybrid classifiers framework the output of two classifiers are combined using a mathematical model to classify gene data. The mathematical model employs a distance based classification that utilizes the output values of IDNN and GANN with respect to the distances between the output and the median threshold, thereby predicting the class type. The mathematical model devised computes confidence and makes a decision based on that. The distance measurement is based on the distance between the output of IDNN and median threshold as well as distance between the output of GANN and median threshold. The distance measurement normally (Hela *et al*., 2004) measures the distance between first classifier and second classifier, whereas our hybrid classifier measures the distance from the median threshold.

## METHODOLOGY

**System framework:** The framework of the system is illustrated in Fig. 1. The major components used in the framework are:

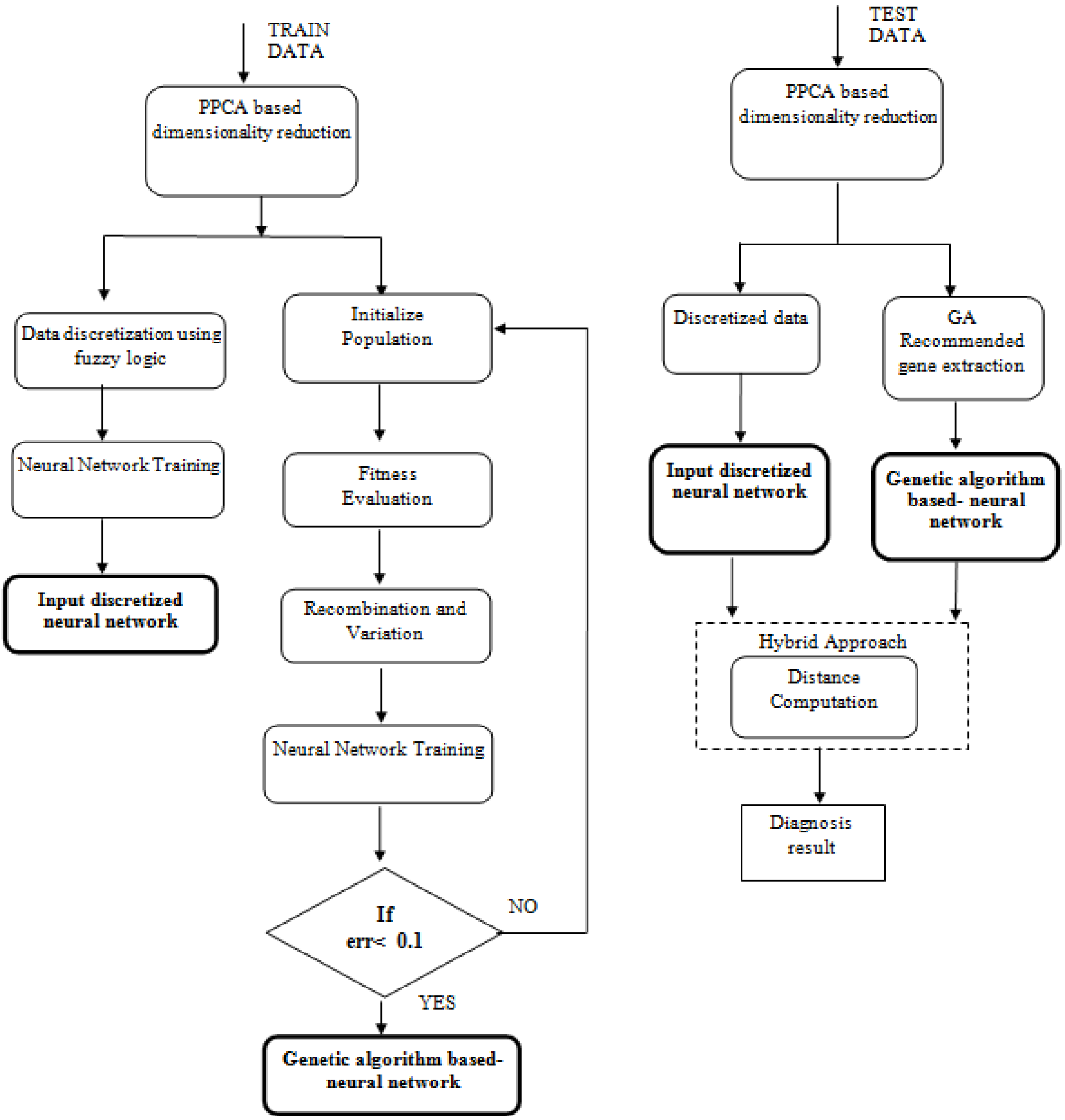
TRAIN DATA
PPCA based dimensionality reduction
Data discretization using fuzzy logic
Neural Network Training
Input discretized neural network
Initialize Population
Fitness Evaluation
Recombination and Variation
Neural Network Training
If err< 0.1
NO
YES
Genetic algorithm based- neural network
TEST DATA
PPCA based dimensionality reduction
Discretized data
GA Recommended gene extraction
Input discretized neural network
Genetic algorithm based- neural network
Hybrid Approach
Distance Computation
Diagnosis result

Fig. 1: System framework

i. Input Discretized Neural Network (IDNN)
ii. Genetic Algorithm-Based Neural Network (GANN)
iii. Hybrid classifier

The microarray gene expression dataset considered is first reduced using Probabilistic Principal Component Analysis (PPCA). The data is discretized using fuzzy logic and fed as input to the IDNN and simultaneously the PPCA reduced data is fed as input to GANN. Both the neural networks are trained using Back-Propagation algorithm (BP). The trained neural networks are subjected to the test data and then hybrid classifier is used to predict the class type.

The notations used in the mathematical model (Jiang *et al*., 2004) are as follows:

$G_{ijk}$ is the microarray gene expression dataset,
$N_c$ is the number of classes present in the microarray data
where, $0 \le i \le N_c - 1$, $0 \le j \le N_s^{(i)} - 1$ and $0 \le k \le N_g^{(j)} - 1$
$N_s^{(i)}$ is the number of data samples present in the $i^{th}$ class
$N_g^{(j)}$ is the number of genes present in every $j^{th}$ sample

**PPCA-based dimensionality reduction:** Probability distribution of a high dimensional data and a low dimensional representation can be determined using the PPCA algorithm formulated by Tipping and Bishop (1999). Reducing or eliminating statistical redundancy between the components of high-dimensional vector data enables a lower-dimensional representation without significant loss of information (Kambhatla and Leen, 1997). In our work, PPCA is used for reducing high dimensional data and eliminates redundancy between components without loss of information (Kambhatla and Leen, 1997). PPCA is a proper choice when the number of samples is large and the data matrix is densely populated (Ilin and Raiko, 2010).

**Input discretized neural network:** The input discretized neural network consists of the following two stages namely:

i. Discretization using fuzzy logic
ii. Neural network training using discretized leukemia dataset.

The design of input discretized neural network is detailed below:

**Discretization using fuzzy logic:** The dimensionality reduced gene expression dataset, is used as the input to fuzzy logic for data discretization. The discretizations of gene data are constructed in the form of if-then statements using linguistic label. The microarray gene expression data is discretized into five set called fuzzy sets. The gene values are mapped into fuzzy set based on the following criteria illustrated below:

$$
\begin{aligned}
&G_{ijk}^{'} \in FC_1; \quad if\ G_{ijk}^{'} < FC_{T1} \\
&G_{ijk}^{'} \in FC_2; \quad if\ FC_{T1} \le G_{ijk}^{'} < FC_{T2} \\
&G_{ijk}^{'} \in FC_3; \quad if\ FC_{T2} \le G_{ijk}^{'} < FC_{T3} \\
&\vdots \\
&G_{ijk}^{'} \in FC_{N_{fc}}; \quad if\ FC_{T3} \le G_{ijk}^{'} \ge FC_{TN_{fc-1}}
\end{aligned}
$$

where, $FC_1$, $FC_2$, $FC_3$, $FC_4$ and $FC_5$ are fuzzy sets and $FC_{T1}$, $FC_{T2}$, $FC_{T3}$, $FC_{\mathrm{T4}}$ and $FC_{T5}$ are fuzzy set thresholds. Based on the fuzzy sets, the gene data are discretized. The determined fuzzy sets are represented by their fuzzy weights. For instance, $FC_1$ has a fuzzy weight of $\alpha_1$, $FC_2$ has a fuzzy weight of $\alpha_2$ and so on. Henceforth, the gene expression values are represented by the fuzzy weights.

**Neural network training using discretized leukemia dataset:** Fuzzy logic is used to discretize the training dataset. The discretized data is given as the input to a supervised feed forward neural network. The network consists of an input layer, a hidden layer and an output layer.

The basis function (Rajasekaran and Pai, 2003) defined by Eq. (1) is used for the designed network:

$$y_j = \tau + \sum_{k=0}^{N_g^{'}-1} w_{jk}\, \hat{G}_{jk} \quad 0 \le j \le N_s^{'} - 1 \tag{1}$$

$$y^{(1)} = \frac{1}{1+e^{-y}} \tag{2}$$

$$y^{(2)} = y^{(1)} \tag{3}$$

Eq. (1) is the basis function, Eq. (2) and (3) represents the sigmoid and linear transfer function selected for the hidden layer and output layer respectively (Rajasekaran and Pai, 2003). In Eq. (1), $\hat{G}$ is the dimensionality reduced microarray gene data, $w_{jk}$ is the weight of the neurons. $\hat{G}$ is given as input to the neural network for training. The neural network is trained using Back-Propagation (BP) algorithm. The steps of Back-Propagation (BP) algorithm are as detailed below:

**Step 1:** Generate and assign weights for input layer to hidden layer and hidden layer to output layer neurons.

**Step 2:** Determine the error rate, $e$ by calculating the difference between the obtained gene expression level ($G_{OUT}$) and target gene expression level ($G_T$) for all the training dataset $\hat{G}$:

$$e = G_T - G_{OUT} \tag{4}$$

where,
$G_T$ is the target output
$G_{OUT}$ is the network output

**Step 3:** Compute new weights for each neuron, (i.e.,) new weights from input to hidden layer and hidden layer to output layer:

$$w = w + \Delta w$$

where,

$$\Delta w = \chi . G_{OUT} . e \tag{5}$$

In Eq. (5),
$\chi$ is the learning rate
$\Delta w$ is the change in weights
$e$ is the error rate calculated in step 2

**Step 4:** Repeat the process from step 2, until a minimized least value of BP error $e$<0.1 is obtained.

Once the training process gets completed, the network gets trained and it becomes suitable for classification, thus $O^{IDNN}$ is obtained.

## GENETIC ALGORITHM-BASED NEURAL NETWORK

Genetic algorithm-based neural network obtains a set of genes as input from dimension reduced microarray gene data. The dimension reduced dataset is given to the genetic algorithm for selection of subset of genes. The subset of gene is selected by initializing population, evaluating fitness, performing genetic operations crossover and mutation. The selected features are used to train the neural network.

**Process of genetic algorithm:** The genetic algorithm is an evolutionary approach to effectively explore the search using fitness evaluation. With the chromosomes selected based on maximum fitness, crossover and mutation operation is performed to obtain optimal solution. The input to the genetic algorithm is the dimension reduced dataset. The steps in the genetic algorithm are detailed below:

**Step 1:** A population set $X_a$; $0 \leq a \leq N_p -1$ is initialized, where, $X_a$ is the $a^{th}$ chromosome of length $L$ and $N_p$ is the population size.

**Step 2:** The neural network is trained and BP error rate $e$, is generated for every chromosome.

**Step 3:** The fitness of the chromosomes that are present in the population pool is determined as follows (Rajasekaran and Pai, 2003):

$$f_a = \frac{1}{e_a} \tag{6}$$

where, $e_a$ is the error rate for each chromosome that can be determined as in Eq. (6).

**Step 4:** The chromosomes with maximum fitness are selected and placed in the selection pool for crossover and mutation.

**Step 5:** A single point crossover operation at a crossover rate $C_R$ is performed over the chromosomes that are in the selection pool.

**Step 6:** Mutation is performed over the child chromosomes. The mutation operation to be performed over a child chromosome is described below.

i. The gene $X^{child}$ is randomly mutated (i.e.,) $x_{r_1}^{child} = r_2$

ii. The new child is checked to find out if it satisfies the criterion $x_0^{child} \neq x_1^{child} \neq \cdots \neq x_{L-1}^{child}$. If the criterion is not satisfied by the new chromosome, then step (i) is repeated until it satisfies the same.

**Step 7:** The new chromosomes obtained are then placed in the selection pool.

**Step 8:** Repeat the process from Step 3 for number of times, until the best chromosome, which has maximum fitness, is extracted from the population pool.

Genetic algorithm renders the neural network trained with best chromosomes resulting in $O^{GANN}$.

## DESIGN OF HYBRID CLASSIFIER

The framework of the designed hybrid classifier considers two separate classifiers. First, the input discretized neural network and second, the genetic algorithm based neural network. In this process two independent classifiers are fused within the framework thus producing the hybrid classification system as shown in Fig. 1. We use the system framework to make a classification based on thresholding the classifier measure that is employed to make a decision. The designed hybrid classifier combines fuzzy logic, genetic algorithm and neural network that overcome the drawbacks of each, while maintaining the advantages of each technique. The combined classifier system includes data discretization, exploration of feature extraction and the fusion method to produce an optimal classifier. The optimal classifier defined in Eq. (8) predicts the class type. The Eq. (8) is based on the sum of outputs of each classifier with respect to distance of each classifier. The distance measured in this work is based on the confidence values associated to the class labels belonging to both classifier outputs. Thus, the distance measure estimates the classifier's confidence.

In the input discretized neural network the fuzzy logic and neural network are used in a combined way while fuzzy logic is used to adjust the inputs using linguistic variables. The neural network is characterized by its effective learning capability. The input discretized neural network learns from the given fuzzified training data.

In Genetic algorithm-based neural network, each individual in the population represents a candidate solution. The GA initializes a population (set) of individuals, computes the associated fitness value, into a new generation of the population using reproduction, crossover and mutation. Genetic algorithm is characterized by the feature extraction from the initial population.

**Algorithm:**

```
Input:
    $O_A$ - Output of classifier A
    $O_B$ - Output of classifier B
    $T_i$ - Threshold of classifiers
    $D_A$, $D_B$-Distance between classifiers output and threshold
Output:
    $O_C$ - Output of classifier C
Begin
    for each output i = 1 to n do
    Determine the output of classifier $O_C$ using Eq. (8)
    for each selected output do
        if $T_i<O_C$ then select a class label $C_i$
    else
        select a class label $C_{i+1}$
End
```

**Determining IDNN distance:** The IDNN distance $D^{IDNN}$ is the distance between the $O^{IDNN}$ and the class threshold, $T^{IDNN}$.

**Determining GANN distance:** The GANN distance $D^{GANN}$ is the distance between the $O^{GANN}$ and the class threshold, $T^{GANN}$:

$$C^{final} = \begin{cases} C_0; & if \ C^H < T^H \\ C_1; & if \ T^H \geq C^H \end{cases} \tag{7}$$

where,

$C^{final}$ is the final class obtained by the designed hybrid classification technique
$C^H$ is the classification result of hybrid approach
$T^H$ is the threshold point

$$C^H = \left( \frac{O^{IDNN}}{D^{IDNN}} + \frac{O^{IDNN}}{D^{GANN}} \right) \tag{8}$$

where,

$$D^{IDNN} = O^{IDNN} - T^{IDNN}$$
$$D^{GANN} = O^{GANN} - T^{GANN}$$
$$T^H = \frac{Upperbound \ + Lowerbound}{2}$$

$O^{IDNN}$ is the output of input discretized neural network
$O^{GANN}$ is the output of genetic algorithm-based neural network
$D^{IDNN}$ is the distance of input discretized neural network
$D^{GANN}$ is the distance of genetic algorithm-based neural network
$T^{IDNN}$ is the threshold of input discretized neural network
$T^{GANN}$ is the threshold of genetic algorithm-based neural network

## RESULTS AND DISCUSSION

The hybrid classifier along with IDNN and GANN are implemented in MATLAB (version 2013a) and the results are evaluated using the leukemia microarray gene expression dataset obtained from broad institute cancer program website. The leukemia data set contains expression levels of 7129 genes taken over 72 samples. Labels indicate that two variants of leukemia are present in the sample (ALL 47 samples and AML 25 samples). The leukemia dataset is subjected to PPCA for dimensionality reduction, so the dataset dimension is reduced to 30 genes. The feed forward back-propagation neural network is used for both training and testing. For training 38 samples (27 ALL 11 AML) were used and for testing 34 (20 ALL 14 AML) samples were used. The network structure uses tansig activation function in hidden layer and purelin activation function in output layer. The number of hidden neurons is calculated using 2n+1, to produce better performance in prediction (Morshed and Kaluarachchi, 1998). The performance of the designed hybrid classifier and existing methods on leukemia datasets are analyzed and the corresponding statistical measures namely specificity and sensitivity are reported.

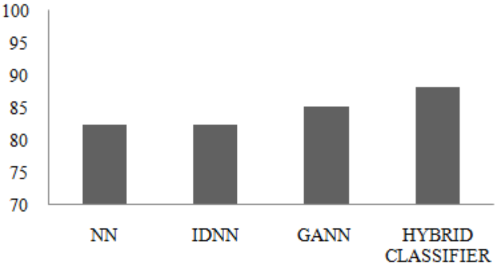

Fig. 2: Performance of designed hybrid classifier, NN, IDNN and GANN in terms of accuracy

Table 1: Accuracy measure of designed hybrid classifier, IDNN, GANN and neural network classifier for Leukemia dataset

| SI. No | Methods | Accuracy |
|---|---|---|
| 1 | NN | 82.35% |
| 2 | IDNN | 82.35% |
| 3 | GANN | 85.29% |
| 4 | Hybrid classifier | 88.23% |

Here, the True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) values are determined as follows:

TP-ALL type correctly identified as ALL type
FP-ALL type incorrectly identified as AML type
TN-AML type correctly identified as AML type
FN-AML type incorrectly identified as ALL type

The statistical performance measures such as accuracy, sensitivity and specificity are calculated using the equations given below:

$$ACCURACY \ = (TP + TN)/(TP + TN + FP + FN) \tag{9}$$

$$SENSITIVITY = TP/(TP + FN) \tag{10}$$

$$SPECIFCITY = TN/(FP + TN) \tag{11}$$

As shown in Fig. 2, our designed hybrid classifier provides 88.23% of accuracy. In case of existing techniques such as, NN, IDNN and GANN, 82.35, 82.35 and 85.29%, of accuracy are obtained (Table 1). The sensitivity and specificity of hybrid classifier obtained are 90 and 85.71%, respectively.

## CONCLUSION

This study uses a hybrid classification approach for leukemia gene expression data. The technique integrates input discretized neural network and genetic algorithm-based neural network into a hybrid classifier approach. The technique has been tested for leukemia microarray gene expression dataset. The technique has classified the leukemia data effectively with remarkable classification efficiency. Performance of the hybrid classifier approach has been compared with the existing approaches such as neural network classifier, input discretized neural network, genetic algorithm based-neural network. The comparative results have shown that the hybrid classifier approach can effectively diagnose leukemia from microarray gene expression data. Further, this study can be extended to more gene expression datasets, multi-class prediction problems and temporal data.

## REFERENCES

Agrawal, R.K. and R. Bala, 2007. A hybrid approach for selection of relevant features for microarray datasets. World Acad. Sci. Eng. Technol., 29(52): 281-287.

Alok, S. and K.P. Kuldip, 2008. Cancer classification by gradient LDA technique using microarray gene expression data. Data Knowl. Eng., 66: 338-347.

Anandhavalli, G., 2008. Analysis of DNA microarray data using association rules: A selective study. World Acad. Sci. Eng. Technol., 42: 12-16.

Anibal, R.F., V.L.G. Juan and R.G. Abalo, 2007. A new predictor of coding regions in genomic sequences using a combination of different approaches. Int. J. Biol. Life Sci., 3(2): 106-110.

Benítez, J.M., J.L. Castro and I. Requena, 1997. Are artificial neural networks black boxes? IEEE T. Neural Networ., 8(5): 1156-1164.

Bidaut, G., F.J. Manion, C. Garcia and M.F. Ochs, 2006. WaveRead: Automatic measurement of relative gene expression levels from microarrays using wavelet analysis. J. Biomed. Inform., 39(4): 379-388.

Bose, S., C. Das, T. Gangopadhyay and S. Chattopadhyay, 2013. A modified local least squares-based missing value estimation method in microarray gene expression data. Proceeding of the IEEE 2nd International Conference on Advanced Computing, Networking and Security (ADCONS), pp: 18-23.

Brintha, S.J. and V. Bhuvaneswari, 2012. Clustering microarray gene expression data using type 2 fuzzy logic. Proceeding of the IEEE 3rd National Conference on Emerging Trends and Applications in Computer Science (NCETACS), pp: 147-151.

Chien-Pang, L., L. Wen-Shin, C. Yuh-Min and K. Bo-Jein, 2011. Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Syst. Appl., 38(1): 4661-4667.

Dai, J.J., L. Lieu and D. Rocke, 2006. Dimension reduction for classification with gene expression microarray data. Stat. Appl. Genet. Mo. B., 5(1): 1-19.

Essam, A.D., 2010. Integration of support vector machine and bayesian neural network for data mining and classification. World Acad. Sci. Eng. Technol., 64(35): 202-207.

Hela, Z., H. Laurent, L. Yves and A. Adel, 2004. Building diverse classifier outputs to evaluate the behavior of combination methods: The case of two classifiers. multiple classifier systems. In: Roli, F., J. Kittler and T. Windeatt (Eds.), MCS, 2004. LNCS 3077, Springer-Verlag, Berlin, Heidelberg, pp: 273-282.

Huynh, H.T., J.J. Kim and Y. Won, 2009. Classification study on DNA microarray with feedforward neural network trained by singular value decomposition. Int. J. BioSci. BioTechnol., 1(1).

Ilin, A. and T. Raiko, 2010. Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res., 11: 1957-2000.

Jiang, D., C. Tang and A. Zhang, 2004. Cluster analysis for gene expression data: A survey. IEEE T. Knowl. Data En., 16(11): 1370-1386.

Jing, L., M.K. Ng and T. Zeng, 2010. Novel hybrid method for gene selection and cancer prediction. World Acad. Sci. Eng. Technol., 62(89): 482-489.

Kambhatla, N. and T.K. Leen, 1997. Dimension reduction by local principal component analysis. Neural Comput., 9(7): 1493-1516.

Kanthida, K., N. Michael, P. Bernhard, B. Christian, R.L. Klaus and G. Armin, 2009. Evaluation of the impact of dataset characteristics for classification problems in biological applications. World Acad. Sci. Eng. Technol., 58: 966-970.

Kaur, H. and G.P.S. Raghava, 2003. A neural-network based method for prediction of γ-turns in proteins from multiple sequence alignment. Protein Sci., 12(5): 923-929.

Morshed, J. and J.J. Kaluarachchi, 1998. Parameter estimation using artificial neural network and genetic algorithm for free-product migration and recovery. Water Resour. Res., 34(5): 1101-1113.

Labib, N.M. and M.N. Malek, 2005. Data mining for cancer management in Egypt case study: Childhood acute lymphoblastic leukemia. World Acad. Sci. Eng. Technol., 8(61): 309-314.

Nittaya, K. and K. Kittisak, 2007. Moving data mining tools toward a business intelligence system. World Acad. Sci. Eng. Technol., 25(22): 117-122.

Peng, Y., W. Li and Y. Liu, 2007. A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification. Cancer Inform., 2: 301-311.

Qi, S., S. Wei-Min and K. Wei, 2009. New gene selection method for multiclass tumor classification by class centroid. J. Biomed. Inform., 42(1): 59-65.

Rajasekaran, S. and G.A.V. Pai, 2003. Neural Networks, Fuzzy Logic and Genetic Algorithms: Synthesis and Applications. Prentice-Hall of India, New Delhi.

Seo, Y.K., W.L. Jae and S.B. Jong, 2005. Iterative clustering algorithm for analyzing temporal patterns of gene expression. World Acad. Sci. Eng. Technol., 4(3): 8-11.

Shreyas, S., N. Seetharam and K. Amit, 2007. Biological data mining for genomic clustering using unsupervised neural learning. Eng. Lett., 14(2).

Sung-Bae, C., 2002. Fusion of neural networks with fuzzy logic and genetic algorithm. Integr. Comput-Aid. E., 9: 363-372.

Tipping, M.E. and C.M. Bishop, 1999. Probabilistic principal component analysis. J. Roy. Stat. Soc. B., 21(3): 611-622.

Xu, Y., V. Olman and D. Xu 2001. Minimum spanning trees for gene expression data clustering. Genome Inform., 12: 24-33.

Zhang, G.B. Huang, N. Sundararajan and P. Saratchandran, 2007. Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis. IEEE ACM T. Comput. Bi., 4(3): 485-495.

Zhou, X., K.Y. Liu and S.T.C. Wong, 2004. Cancer classification and prediction using logistic regression with Bayesian gene Selection. J. Biomed. Inform., 37(4): 249-259.