TurkishJournal of Computer and MathematicsEducation Vol.12 No. 7 (2021), 1372 -1378

Research Article

Classification Of Microarray Gene Expression Using Artificial Neural Network(ANN)

Dr.K.Kalyani
Asst.prof., PG and Research Dept. of Computer Science,
Marudupandiyar College, Thanjavur - 613 403, Tamil Nadu
(Affiliated to Bharathidasan University, Tiruchirappalli)

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: The accurate cancer classification is very important task for cancer treatment. Recently the informative genes are identified from the thousands of genes for correct cancer classification. The collection of microscopic Deoxyribo Nucleic Acid (DNA) microarray is attached in the solid surface. In this study, DNA microarray data is used for cancer classification. The system uses Artificial Neural Network (ANN) for DNA Microarray Data Classification. Initially, the pre-processing step is made by using log transformation method to remove the raw data and feature selection. These selected features are classified by using ANN. Rectified Linear Unit (RELU) activation function isused as the activation function in each ANN layer. Soft Max is used for classification. The performance of the system is made by using leukemia dataset and produces the classification accuracy of 91.65% by using ANN.

Keywords: Microarray Data Classification (MDC), Gene Expression, Flat Pattern Filtering, Feature Selection, Artificial Neural Network (ANN)

I.INTRODUCTION

Microarray data classification (MDC) are widely used in diagnosis handling and disease arrangement through a variability of genes assortment and classification methods. Since cancer diagnosis has seen much claims of microarray predominantly on gene countenance profiles, scholars also have begun sightseeing data analysis by using the technology. This is attributed to its effectiveness in discovering abnormal and normal tissue patterns in speedier time as microarray scales well on large dataset. Microarray is an attractive research avenue as the technology is typically utilized to investigate dataset with high dimensionality, which demands significant memory and processor requirements. There remains room for improvement of microarray in classifying data as the technology struggles with small samples collection yet large features quantity. Selection of suitable features are the keys in this field as numerous research endeavors aim to minimize data dimensionality with improved performance in classification. In the case of classifying cancer cells, numerous machines learning algorithms strive to a number of workable samples is significantly lower than gene count. This situation affects efficiency and effectiveness to a large data dimensionality which impair classification performance. Convolutional neural network is an instance of deep learning strategy is mimicking brain function in processing information. In this paper, multilayered Neural Network, which is a deep learning algorithm, is proposed to classify microarray cancer data in the identification of type of cancer. ANN is proposed due to its ability in dealing with insufficient data and boosting classification performance. In addition, ANN is also powerful in integrating cancer datasets that are strongly linked, which improves performance in classifying data. This is attributed to its capability in detecting latent appearances of cancer from equivalent types.

Dual Tree M-Band Wavelet Features (DTMBWF) based MDC is described in[1]. Initially the input micro array data is given to DTMBWF. The K-Nearest Neighbor (KNN) classifier is used for classification. Gene as sortment using MDC using single and multiple filters is discussed in [2]. At first, the micro array data is given to single filter single wrapper and multiple filter multiple wrappers to improve forcefulness. The classification is made by KNN, Support Vector Machine (SVM) and weighted voting classifier. Genetic data classification based on supervised attribute clustering is discussed in [3]. The input microarray data is preprocessed by gene clustering method. Then the classification is made by naive bayes, KNN and SVM. ANN using data dependent kernel machines is described in [4]. The input microarray data is given to bootstrapping based resampling. The ANN is made by using SVM, KNN and uncorrelated linear discriminant classifiers. Microarray Data Classification (MDC) for cancer classification-based dimension reduction is discussed in [5]. The input microarray data is given to singular value decomposition. The features are selected by recursive feature elimination. SVM classifier is used for classification. MDC using genetic algorithm is described in [6].The input micro array data is given to Pearson, spearman, cosine coefficients, mutual information, signal to noise ratio and information gain to select the efficient features. The genetic algorithm-based classifier is used for classification.

SVM Map Reduce based MDC is described in [7]. Initially the gene filtering is performed to remove gene ranking, noisy data and filtering informative genes. The SVM Map Reduce is selected by using feature subset

selection method. Classification is made by SVM. Gene expression data using associative classification is discussed in[8].At first, the gene expression data is given to gene filtering. Then discretization is done for data preprocessing. Associative classification is used for classification.MDC for high dimensional data-based feature selection is discussed in [9]. Initially the gene filtering is used to select and reduce the features. The C4.5 classifier is used for classification. DNA based MDC for cancer classification is discussed in[10].The input micro array data is given to novel strategy for extraction. Then gene rank selection is performed. The classification is made by simple rule-based ensemble classifiers.

MDC for gene selection algorithm and combination of ranking and clustering is discussed in [11]. The genes are selected by clustering and non-clustering gene selection. MDC is made by using SVM classifier. MDC using fuzzy inference system is described in [12]. The features are selected by t-statistics method. The selected features are classified by using fuzzy inference system. An efficient approach for MDC system using ANN is discussed in this study. The organization of paper is as follows: Section 2 describes the methods and materials used for MDC system. Results and discussion of MDC system is explained in section 3. The last section concludes the MDC system.

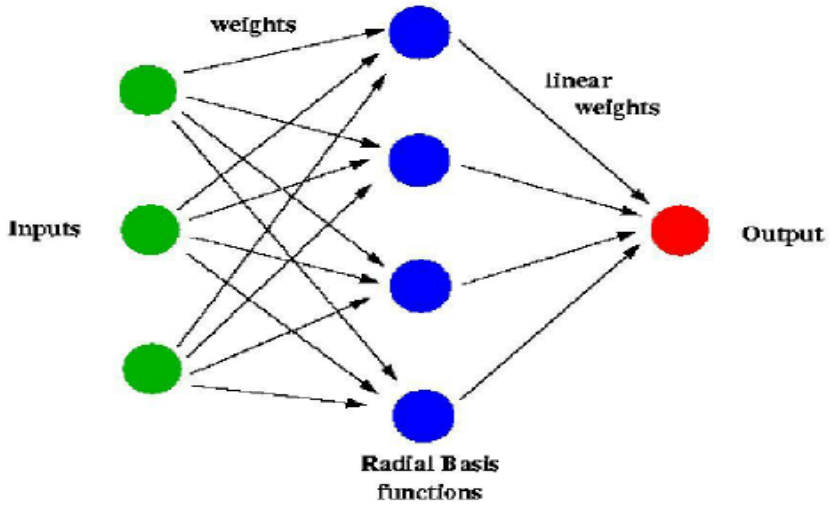

Fig. 1. Artificial Neural Network

II. METHODS ANDMATERIALS

The ANN based MDC system is shown in figure1. At first the input micro array data is given to preprocessing step using log transformation to remove raw data to get clear data and also it selects the efficient features. Then ANN is used for the classification of MDC system.

Microarray gene expression data have been utilized in past researches to perform cancer type classification by using machine learning strategies. Decision tree (DT) was the most primitive machine learning strategy introduced in comparing human proteins to informative gene in proteins containing diseases [6]. Diagnoses of cancer have been largely assisted by exploring gene expression data with the application of technology available in microarray technique. The technology enables genes to be measured simultaneously in a large quantity. In assessing significant genes, parametric statistical analysis has been typically employed to establish statistical significance [7]. In literature, numerous algorithms and mathematical models have been constructed and proposed to interpret and analyze gene expression data. In analyzing gene expression data, two dominant strategies which have been focused are clustering and classification [8]. Additionally, there are also numerous techniques which have been executed previously in classifying gene expression data, including, k-nearest neighbors (k-NN) [9], Support Vector Machines (SVM) [10], Multilayer 2018 International Conference on Advanced Science and Engineering (ICOASE), Kurdistan Region, Iraq 146 perceptron (MLP) [11], and variants of Artificial Neural Networks (ANNs) [12].

A. Gene data preprocessing

The preprocessing of microarray data is an essential stage to remove raw data in the dataset. In this study, the log transformation technique is used for preprocessing. It is a popular transformation technique to transform the data into normal data. The log-normal distribution is followed by original data. The log transformation technique is used in this study to remove raw data and select the efficient features for classification. The highly skewed distributions are reduced by log transformation. It can be used in the data to

help the assumptions of inferential statistics.

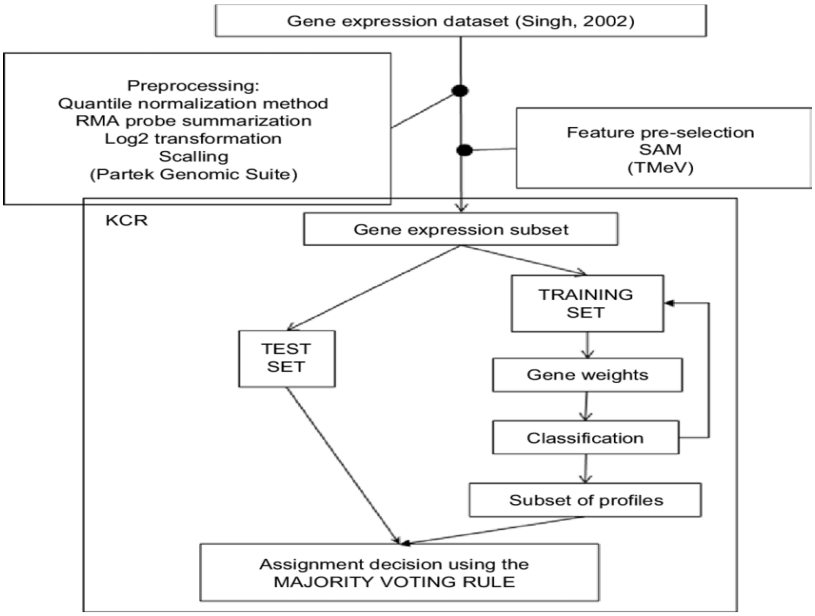

Fig: 2. Gene data preprocessing

B. MDC using ANN

ANN model is based on structure functions. ANN structure may change due to the input and output. ANN has input, output and hidden layer. It is also used in hand moment classification, heartbeat classification and electrocardiogram signal classification. The input and output patterns with complex relationship are found in the nonlinear statistical data modeling tools. It has a group of nodes which is inter connected by the neurons in the brain. The node in one layer is connected with the other layer.

The artificial neurons are represented as circular nodes and the connection of artificial neurons are made by using arrows from input to output layer. The first layer neuron is connected to the next layer. The neuron in each layer is connected to the next layer. ReLU activation function is most commonly used activation function in neural networks. ReLU function is its derivative both are monotonic. The function returns 0 if it receives any negative input, but for any positive value x, it returns that value back. Thus, it gives an output that has a range from 0 to infinity. The activations functions that were used mostly before ReLU such as sigmoid or tanh activation function saturated. This means that large values snap to 1.0 and small values snap to -1 or 0 for tanh and sigmoid respectively. Further, the functions are only really sensitive to changes around their mid-point of their input, such as 0.5 for sigmoid and 0.0 for tanh. This caused them to have a problem called vanishing gradient problem.

Neural Networks are trained using the process gradient descent. The gradient descent consists of the backward propagation step which is basically chain rule to get the change in weights in order to reduce the loss after every epoch. It is important to note that the derivatives play an important role in updating of weights. Now when we use activation functions such as sigmoid or tanh, whose derivatives have only decent values from a range of -2 to 2 and are flat elsewhere, the gradient keeps decreasing with the increasing number of layers. This reduces the value of the gradient for the initial layers and those layers are not able to learn properly. In other words, their gradients tend to vanish because of the depth of the network and the activation shifting the value to zero. This is called the vanishing gradient problem.

But there are some problems with ReLU activation function such as exploding gradient. The exploding gradient is opposite of vanishing gradient and occurs where large error gradients accumulate and result in very large updates to neural network model weights during training. Due to this, the model is unstable and unable to learn from your training data.

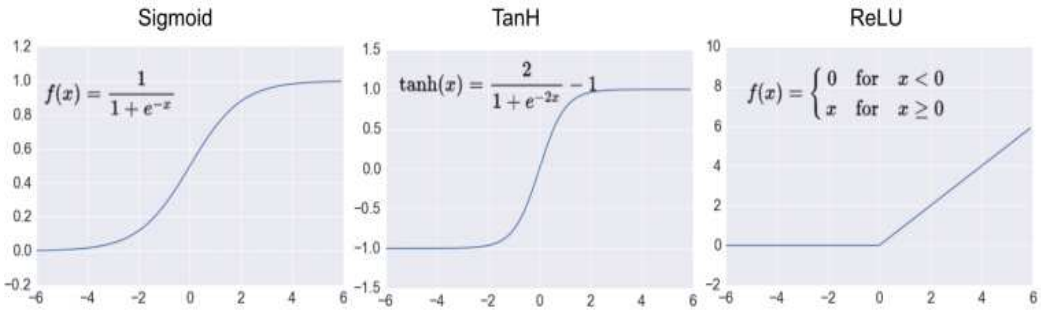

Fig.3. ReLU activation function

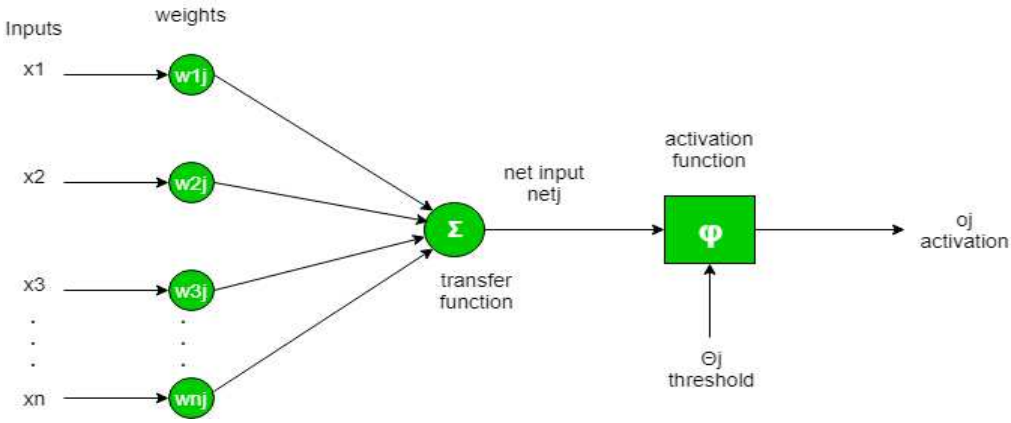

Fig.4. ReLU Activation Function with threshold

The Soft Max layer is used for classification. The n-dimensional vector is taken from the real values with the range of 0 and 1. It is used for binary classification with two or more classes. In this work, the ANN is used for the classification of micro array data in MDC system. The ReLU activation isused in each layer of ANN for the classification. Finally, the Soft Max layer in the fully connected layer is used for the classification of MDC system.

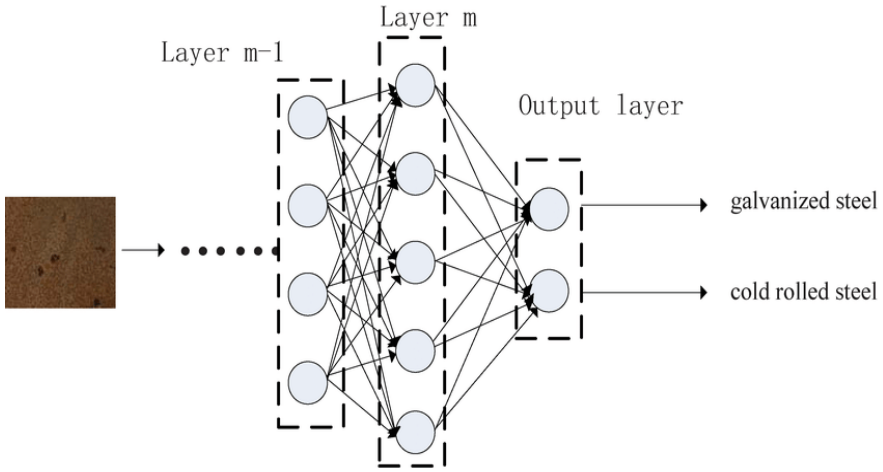

Fig.5 Fully Connected layer of MDC system

III. RESULTS ANDDISCUSSION

Performance of MDC system is analyzed by using gene expression dataset using ANN classifier and publicly available cancer microarray dataset. It consists of leukemia dataset. The table 1 shows the microarray dataset description.

Table- I: Gene expression dataset-Description

Name of gene-expression dataset	Number of features in the dataset	Number of Samples	Number of classes	Reduced number of features with SVD	Number of selected features by InfoGainAttributeEval
(ALL-AML)	7,129	73	2	38; 35	3
Leukemia-1	5,327	72	3	72	6
Colon	2,000	62	2	62	3
SRBCT	2,308	83	4	83	8
Lung Cancer	12,600	203	5	203	9
DLBCL	6,817	77	2	77	7

Table- II: Leukemia gene dataset-Description

Dataset	No of cancer cases	No. of normal cases	Total no. of cases	No of Attributes
Leukemia	47	25	72	7129

The input microarray is given to log transformation for preprocessing to remove raw data and also it selects the efficient features. Then the selected features are used for classification using ANN. Table 2 shows the classification accuracy of MDC system using ANN for ten attempts.

Table-III: Classification accuracies obtained at ten attempts using ANN for MDC system

No. of attempts	Classification accuracy (%)	No. of attempts	Classification accuracy (%)
1	93.50	6	96.00
2	93.00	7	91.50
3	91.50	8	94.50
4	89.00	9	92.00
5	88.50	10	90.00
		Average accuracy(%)	**91.65**

From the above table, it is observed that the overall classification accuracy of 91.65% obtained by using ANN for MDC system. The higher classification accuracy of 96% and the lowest classification accuracy is 88.5% obtained by using ANN for MDC system. Figure 3 shows the performance often attempts for MDC system using ANN.

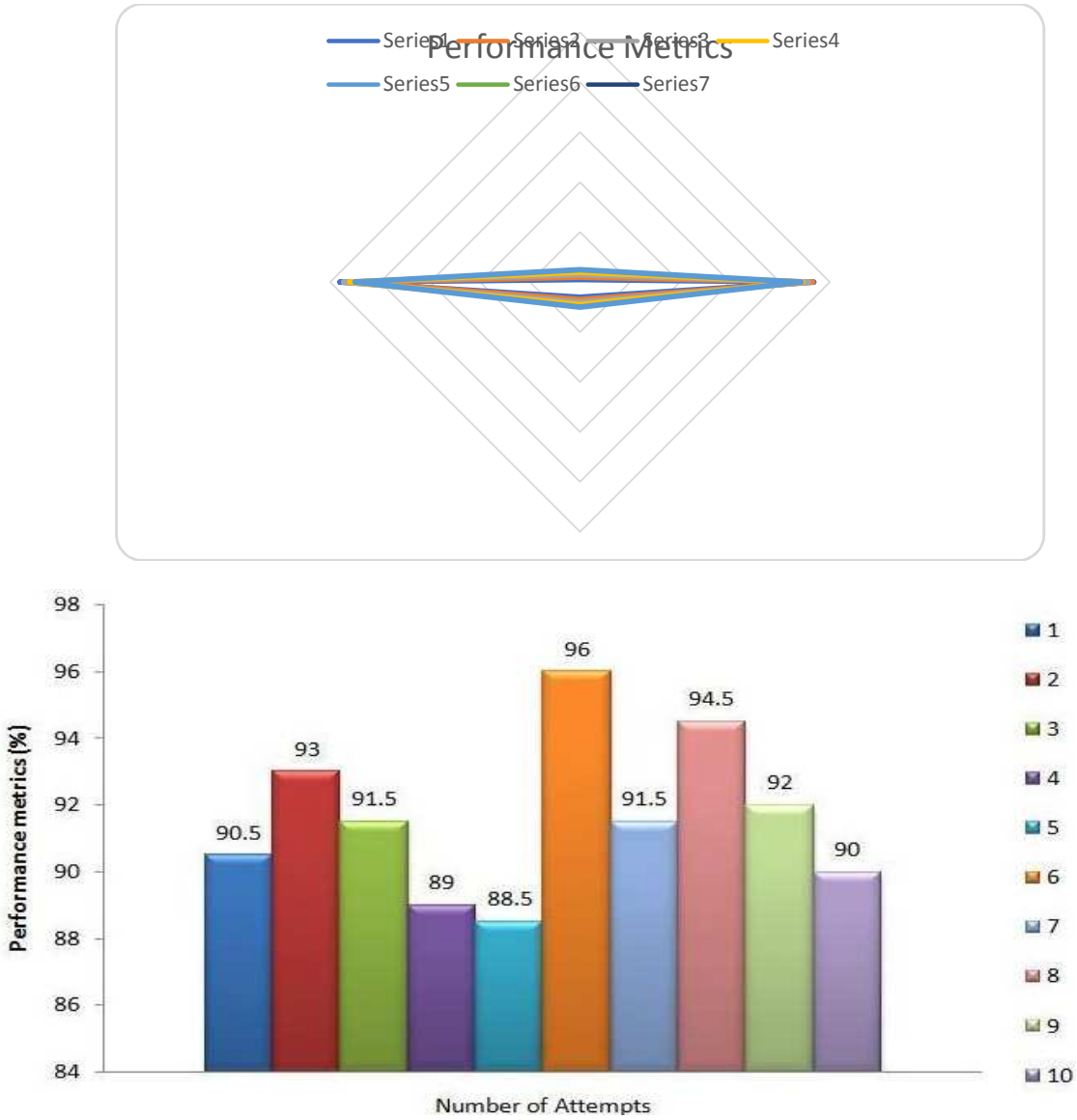

Fig. 6. Classification accuracies of ten attempts using ANN for MDC system

In the above figure, it is clearly observed that the higher classification accuracy is 96% and lower classification accuracy is 88.5% using ten attempts for MDC classification system using ANN classifier.

IV. EVALUATION TECHNIQUE:

In assessing the proposed deep learning CNN, ten cancer data were tested. These data were used in training classification. Mean accuracy was obtained through averaging accuracy scores from the data. This eliminates concerns on redundant tests and optimizes utilization on data that have been obtained. In this paper, accuracy as a measure of performance for the proposed convolutional CNN is defined as following: To evaluate the performance, the accuracy of the result is calculated.

$$\text{Accuracy} = \frac{\textit{Correctly Predicted Data}}{\textit{Total Testing Data}} \times 100\%$$

V. CONCLUSION

An approach for MDC system using log transformation and ANN is described in this study. The performance of the system is made by publicly available gene expression database. The MDC system uses leukemia datasets for performance evaluation. The input leukemia microarray data is given to log transformation to reduce raw data and also to select the efficient features for preprocessing. The preprocesse

dimages are used for further step. ANN classifier is used for the classification of MDC system into normal and abnormal. The average classification accuracy is calculated for ten attempts of classification due to variations in accuracy. The MDC system produces the average classification accuracy of 91.65% obtained by using ANN classifier.

REFERENCES

1. J.M. Sonawane, S.D. Gaikwad, and G. Prakash, "Microarray Data Classification Using Dual Tree M-Band Wavelet Features", International journal of advances in signal and image sciences, Vol. 3, No. 1, 2017, pp.19-24.
2. Y. Leung, and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 7, No. 1, 2010, pp.108-117.
3. P. Maji, "Mutual information-based supervised attribute clustering for microarray sample classification", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, 2010,pp.127-140.
4. Chang, Wen-Yeau, and Po-Chuan Chang. "Application of Radial Basis Function Neural Network, to Estimate the State of Health for LFP Battery." International Journal of Electrical and Electronics Engineering (IJEEE) 7.1 (2018): 1-6.
5. L. Shen, and E.C. Tan, "Dimension reduction-based penalized logistic regression for cancer classification using microarray data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 2, pp.166-175.
6. X.R. Jenifer, and R. Lawrance, "Classification of microarray data using SVM mapreduce", IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, 2017, pp. 1-6.
7. Achakanalli, Santosh, and G. Sadashivappa. "Statistical Analysis Of Skin Cancer Image–A Case Study." International Journal of Electronics and Communication Engineering (IJECE) 3.3 (2014): 1-10.
8. E. Ahmed, N. El-Gayar, and I.A. El-Azab, "Support Vector Machine ensembles using features distribution among subsets for enhancing microarray data classification", International Conference on Intelligent Systems Design and Applications, 2010, pp.1242-1246.
9. S. Alagukumar, and R. Lawrance, "Classification of microarray gene expression data using associative classification", International Conference on Computing Technologies and Intelligent Data Engineering, 2016, pp.1-8.
10. AZRIYENNI, MW MUSTAFA, DY SUKMA, and ME DAME. "Application of Backpropagation Neural Network for Fault Location in Transmission Line 150 kV." International Journal of Electrical and Electronic Engineering (IJEEE) 2.4 (2013): 21-30.
11. H. Yu, and S. Xu, "Simple rule-based ensemble classifiers for cancer DNA microarray data classification", International Conference on Computer Science and Service System, 2011, pp.2555-2558.
12. M.J.Rani,andD.Devaraj,"ACombinedClusteringandRankingBased Gene Selection Algorithm for Microarray Data Classification", IEEE International Conference on Computational Intelligence and Computing Research, 2017, pp.1-5.
13. Hamdan, S. A. L. A. M., and A. D. N. A. N. Shaout. "Face Recognition Using Neuro-Fuzzy and Eigenface." International Journal of Computer Science and Engineering (IJCSE) 5.4 (2016): 1-10.
14. M. Kumar, and S.K. Rath, "Classification of microarray data using Fuzzy inference system" International Conference on Recent Trends in Information Technology, 2014, pp.1-8.
15. K.P. Shashikala, and K.B. Raja, "Palmprint identification using Log Transformation of Transform Domain Features" International Conference on Electronics and Communication Systems, 2014, pp.1-5.
16. Mengistu, ABRHAM DEBASU, and DAGNACHEW MELESEW Alemayehu. "Robot for visual object tracking based on artificial neural network." International Journal of Robotics Research and Development (IJRRD) 6.1 (2016): 1-6.
17. P. Meaney, T. Grzegorczyk, S.I. Jeon, and K. Paulsen, "Log transformation with Gauss-Newton microwave image reconstruction reduces incidence of local minima convergence", IEEE Antennas and Propagation Society International Symposium, 2009, pp.1-4.
18. Ramachandran, Vedantham, E. Srinivasa Reddy, and B. Sivaiah. "An enhanced facial expression classification system using emotional back propagation artificial neural network with DCT approach." Int. J. Comput. Sci. Eng. Inf. Technol. Res.(IJCSEITR) 5 (2015): 83-94.