International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14218-14231
© Research India Publications. http://www.ripublication.com

GOA-based DBN: Grasshopper Optimization Algorithm-based Deep Belief Neural Networks for Cancer Classification

Praveen Tumuluru
Research Scholar, Department of Computer Science and Engineering
Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, India.

Orcid Id: 0000-0001-9426-205X

Dr. Bhramaramba Ravi
Associate Professor, Department of Information Technology
Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, India.

Abstract

Cancer is a dreadful disease in the world and it drains human life and hence, the accurate identification of the disease is required in the early stages. The early detection causes the need for a better and accurate method that offers the information of the cancer in the patient that enables better decision-making by the clinicians and treating them. Thus, the paper proposes a cancer classification strategy using the gene expression data. The proposed Grasshopper Optimization Algorithm-based Deep belief Neural Networks (GOA-based DBN) aims at performing the cancer classification with improved classification accuracy, for which the Logarithmic transformation and Bhattacharya distance are used. The Logarithmic transformation pre-processes the gene expression data for reducing the complexity associated with the classification and Bhattacharya distance selects the highly informative genes. The weight update in the Deep belief Neural Networks is based on average error estimate using the GOA and Gradient Descent. The experimentation using the colon data and Leukaemia data proves the effectiveness of the proposed cancer classification. The proposed classification method using the gene expression data attains an accuracy rate of 0.9534, FAR of 0.0769, and detection rate of 0.9666.

Keywords: Cancer Classification, Deep Belief Neural Network, Bhattacharya Distance, Grasshopper Optimization Algorithm, Logarithmic transformation.

INTRODUCTION

Cancer is the threatening diseases that cause the death of major humans and the DNA Microarray-based gene expression profiling is the effective technique employed for cancer classification, diagnosis, prognosis, and for treatment. Cancer classification depends on the collection of the significant genes that enables the improved accuracy and leads to the early diagnosis of cancer [13]. The genetic information regarding the function and development of the genes is present in the Deoxyribonucleic acid or DNA and it is referred as the blueprint of the living beings as the information regarding the normal life and maintenance is present in the genetic information. The genetic information is protected and is inherited in all the cells and the new cells acquire the information during cell division that is a process of formation of two or more child cells from a parent cell. The structure of DNA molecules is that they are double twisted helix bond, tied together and arranged in a proper order. Thus, four molecular units form the DNA helix, which is sequenced in a regular manner enabling the bond that holds the individual strands of one component with the components of other strands. DNA duplicates through the breakage of the bonds between the strands with the individual strands forming a matching strand, re-bond and re-twist again and again. The total DNA sequence is essential for synthesizing the RNA molecules, such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) [3] [4].

The main role of DNA is to develop the proteins that perform the cell functions and the proteins are established using two major steps, such as transcription stage and translation stage. In the transcription stage, the DNA molecule is translated as messenger RNA or mRNA and in translation stage, the mRNA is transcribed as proteins' amino acid sequences that perform the cell functions. The gene is expressed soon after the construction of the protein and the effective technique to measure the gene expression is to compute the mRNA rather than the proteins [5]. The importance of using the mRNA sequences is because they possess the capacity to be hybridized along with their complementary RNA or DNA sequences, but there is no need for the proteins [1]. DNA microarray dataset or Gene expression profiling ensures the determination of thousands of genes present in an individual RNA sample through the hybridization of a labeled unknown molecular obtained from a specifically tissue that is significant. It provides an efficient method to gather the data, which is employed to compute the gene expression patterns of the genes present in an organism and this gathering of data is obtained through a single experiment [14], [15]. DNA microarrays identifies the type and category of the genes expressed in a particular type of the cell at a specific time and conditions. Finally, the comparison of the gene expression [6] [7] among various types of the cells or tissue samples, at last the highly informative genes are captured such that the cells or tissues contributing to various types of the disease or cancer is detected [8]. The technique finds valuable application in extracting the essential patterns and in establishing a classification models using the gene expression data and the methods have assisted the cancer prediction [9] and their prognosis. Thus, it stands as an effective platform for analyzing the gene expression for various experiments by the researchers [10].

Gene expression level indicates the measure of RNA developed in a cell under during various biological states such that the diseases present in the parent cell is inherited to the daughter cells and the best example is cancer or malignant tumors and these characters are uncontrollable that affects the other gene expression values [12]. The DNA microarray technology [2] gives the prediction and it enables to understand the life at the molecular level better and requires analytical methods that analyze the huge amount of data precisely and to transfer the analog form of the DNA microarray samples to digital there are several steps. Hybridization detects the presence of certain genes and the dye is scanned to generate the numerical intensity of individual dye, which is then scanned as an image using the image processing techniques. The gene expression level depends on these intensities and the level is generated through the comparison of the gene with variable conditions [1]. The microarray technology analyzes the levels of the gene-expression data [17] [16],[29] of thousands of genes present in the cells used as the test sample. The comparison with other samples facilitates investigations, improvement in disease, accurate diagnosis, medication, and prognosis after treatment [11]. The commonly employed statistical methods in genomic data analysis are machine learning techniques and a large number of the hybrid evolutionary algorithms are employed that enhances the accuracy of the classification methods [24] [4].

The paper proposes the cancer classification strategy that classifies the genes based on the presence or absence of the cancer genes. The genes of the person are organized as gene expression data that carries the information of various genes corresponding to the individuals. Initially, the gene expression data is subjected to pre-processing that is essentially used in cancer classification to refine the gene in the gene expression data such that the complexity in the classification is released and for pre-processing logarithmic transformation is used. Then, the pre-processed data is provided to the gene selection module that selects the informative genes. The informative genes offer all valuable information to perform the classification that ensures accuracy and efficiency of cancer classification. The selection of the genes is done based on the minimum value of the Bhattacharya distance and the selection of the genes is done in comparison with the ground value. Then, the selected or the dimensionally reduced genes are presented to the classification module that possesses the GOA-based DBN to classify the genes as normal or abnormal. Normal class indicates the absence of cancer and abnormal class signifies the presence of cancer. The effective and computationally burden-free computation is ensured using the GOA-based DBN that is duly based on the error estimate.

GOA-based DBN: The contribution of the paper is the proposed GOA-based DBN that determines the optimal weights of DBN based on the average error estimate. The weight update follows the gradient descent if the average error corresponding to the gradient descent is low or otherwise GOA is employed for the weight update.

The paper is organized as: The background of the paper is depicted in section 1, section 2 deliberates the literature works and section 3 depicts the proposed work for cancer classification. The results of the proposed method are deliberated in section 4 and finally, section 5 concludes the paper.

MOTIVATION

The literature review of the works is depicted in this section.

Literature Survey:

Sara Tarek *et al.* [1] proposed Ensemble classifier based on gene expression that performs the cancer classification. The method offered better accuracy and served as a better classifier in classifying the various types of cancers. This method destroyed the effects of overfitting, but classifiers cannot be employed as a base member and this system cannot be applied to other benchmarks especially multiclass datasets. Hanaa Salem *et al.* [2] proposed Information Gain (IG) and Standard Genetic Algorithm (SGA) that was based on the gene profiles to perform the cancer classification. The method employed Information Gain, Genetic Algorithm, and Genetic Programming (GP) for feature selection, reduction, and classification, respectively. The accuracy of classification remained high and it offered robust outcomes, but suffered as a result of time complexity and unable to deal with classification on multi-class labels. Boris Winterhoff *et al.* [3] used Agilent microarrays that determined the gene expression of various types of the ovarian cancers. The merits of the method are that it develops robust biomarker signatures to assist the clinicians in identifying the therapies, but the demerits are that the method offers poor result regarding the survival of the patients that express one of the TCGA signatures and is restricted to stage III/IV patients. M. Dashtban and Mohammadali Balafar [4] proposed Intelligent Dynamic Genetic Algorithm (IDGA) that offered faster convergence for detecting predictive genes and provided required crossover and mutation probability. The disadvantages of the method is the convergence that is not made in proper for various crossover types and the incorporation of median instead of the average point makes the inability of bringing the adaptive crossover relation. Sun-Yuan Hsieh *et al.* [5] proposed Gene Expression Graph (GEG)-based data structure for the classification that offered better performance than the existing methods and correctly detects out-of-class samples in addition to correctly classifying samples in the corresponding classes. The demerits are that the method cannot address the identification of similar diseases and the PSs of large-scale datasets. Jin-Xing Liu *et al.* [6] introduced Robust Principal Component Analysis and Linear Discriminant Analysis (RPCA+LDA), Support Vector Machine (SVM) that were effective and feasible for tumor classification. The method failed on considering the biological meanings of gene selection. Huijuan Lu *et al.* [7] proposed a Mutual Information Maximization and the Adaptive Genetic Algorithm (MIMAGA) Selection Method that significantly reduced the dimension of gene expression data and removed the redundancies for classification and the reduced gene expression dataset provided highest classification accuracy compared to conventional feature selection algorithms. The demerit is that the method took relatively long time for iterative feature selection algorithm and it leads to time complexity.

Challenges:

- The extraction of datasets with right conditions is a tedious process since it involves complex process and degree of noise remains high [7].
- The process of differentiating various genes among the various cancer types and classes with the elimination of the irrelevant genes is a hectic challenge [2].
- The efficiency of the classification relied on several trade-offs that should be balanced or else classifier performance is affected and the tradeoffs include accuracy-generalization, complexity- classifier performance, Performance-memory requirements [12].
- Some of the datasets possess very less number of instances sometimes less than one hundred. These datasets may satisfy the expression level equivalent to several thousands of genes, do not suffer from the high dimensionality issues, and due to the small size of the samples of the experimental data. However, some of the traditional classification methods are not effective for the gene expression classification as they exhibit poor classification accuracy [15].
- The challenge imposed by the Microarray dataset is that they suffer from the dimensionality issues, the less number of the data samples, and the presence of the irrelevant and noisy genes. They offer very poor classification experience and the challenge regarding the irrelevant genes is that they add extra noise during the analysis of the gene expression data that causes the increased computational complexity during classification and clustering [13].
- The ensemble classifier system performs cancer classification, for which the gene expression data is employed. The classifier system enhances the performance of classification using the ensemble of traditional classifiers namely, K-NN, and Naïve Bayes classifiers, but they are unrealistic as they could not classify the huge sized data [1].

PROPOSED METHOD OF CANCER CLASSIFICATION USING THE GOA-BASED DBN

The goal of the research is to develop a classification model using the gene expression data, for which a new classifier is developed. The proposed classifier is based on the error estimate, for which it uses the gradient descent and the GOA for tuning the DBN that performs the optimal classification of the cancer genes. Initially, the gene expression data is pre-processed to ensure the gene ingredients stay in a limited range. The pre-processing stage using the Logarithmic transformation is to enable the classification with no complexity such that the accuracy of the method is preserved to a greater value. The pre-processing is carried out using the logarithmic transformation that is followed by the gene selection, performed using the Bhattacharya distance. The gene selection module selects the highly relevant genes that carry the information about the disease knowledge of the person. The gene selection module removes the redundant genes and minimizes the dimension of the data facilitating the reduction in the complexity of the data. Once the genes are selected, they are presented for classification using the classifier that derives the cancer categorization based on the selected genes. The importance of genes is that the characters of the genes are inherited to the newly formed genes at the time of cell division and this process of transferring the characters is uncontrollable. Figure 1 shows the proposed classification strategy using the gene expression data.

Consider the gene expression data denoted as G and the dimension of the gene data as, $(P \times Q)$ that carries the information of the genes. The gene expression data consists of the gene information of P number of persons. The gene expression data holds a large number of the genes that are present in the person and the process of studying the characteristics of the genes enables us to understand the life of the person better. The gene expression data consists of the gene data for a large number of persons and the classification using this data is advantageous as it facilitates the better understanding of a person very easily and ensures the early prevention.

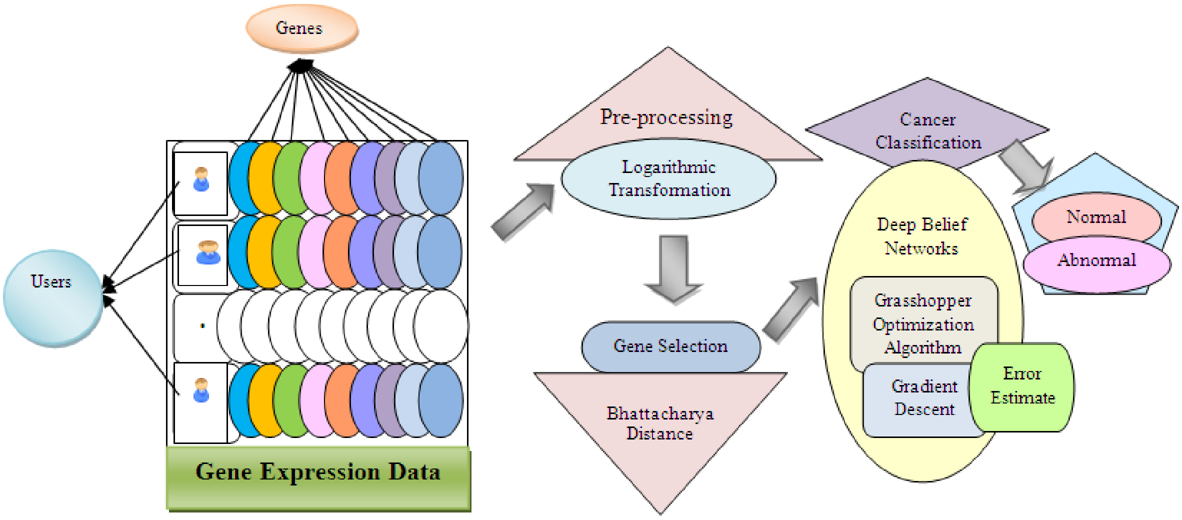


Figure 1. Schematic diagram of the proposed Cancer Classification Strategy

Preprocessing:

The gene expression data is pre-processed [25] using the logarithmic transformation [1] and this step symbolizes the real data processing step that does not alter the dimension of the gene expression data, but pre-processes on the data ratio. Log transformation is employed to reduce the skewness of the data enabling the higher interpretability of the data. The data in the gene expression data possess an unsymmetrical ratio and to make them symmetric, the log transformation is employed, in which the user can select the preferred base for log-transformation. Pre-processing of the gene expression data using the logarithmic transformation is given as,

$$K = \log_{10}(G) \tag{1}$$

where, K is the log-transformed gene expression data. The log-transformation is required to be done before the actual classification and it enhances the gene expression data such that the complexity in the classification is reduced. The pre-processed data is fed to the gene selection module

Gene Selection

The pre-processed data is presented to the gene selection module that is the essential criterion as the pre-processed gene expression data holds a lot of the microarray data. The huge microarray of the gene data yields poor discriminative power and destroys the accuracy of the classifier and thereby, inhibiting the complexity in classification. Therefore, it is essential to minimize the genes such that only the highly significant genes are selected, as they possess the high discriminate power in differentiating the genes for classifying the various diseases. As like the work given in [26], [27], and [28], the genes are extensively analyzed to diagnose the lung cancer. Similarly, in this work, the gene selection is progressed using the Bhattacharya distance [20] to effectively train the classifier. The gene selection aims at removing the irrelevant and redundant genes from the pre-processed gene expression data such that it minimizes the time of computation, complexity in classification, minimizes the storage area. Moreover, gene selection enhances the classifier performance, promotes data visualization, and ensures the better understanding of the data by highlighting the interrelationship existing among the genes. One of the other merits of gene selection is that it controls over-fitting and the absence of generalization of over-fitting has bad impacts on the performance of the classification.

The dimensional reduction ensures the effective classification accuracy and the genes of higher importance are drawn out from the pre-processed gene expression data. Each of the individuals possess a number of the genes and they provide the information regarding the individual. The genes corresponding to the people are compared with the ground truth individually and the comparison is made using the Bhattacharya distance and the genes that hold minimum distance with the ground truth is selected as the highly relevant genes. The relevant genes selection is based on the following formulations. The gene selection intends to determine $(P \times Q)$ dimensional linear transformation matrix ψ that maps the Q-dimensional original features to the P-dimensional reduced feature space through minimizing the upper bound of error probability given as,

$$(X)_p = (\psi^T)_{P \times Q} (K)_Q \tag{2}$$

The above equation in terms of the Bayes classification error probability ρ_v is given as,

$$\psi = \arg\max_{\psi} (\rho_v) \tag{3}$$

Equation (3) appears to be hard for analytical performance analysis and hence, the minimum bound of the classification error is indulged for the calculation of the minimum Bayes error rate. The upper bound of the Bayes minimum error probability $\partial_{c_1 c_2}$ for the classes c_1 and c_2 is given as,

$$\rho_{vc_1c_2} \le \partial_{c_1 c_2} = \sqrt{\rho(\omega_{c_1})\rho(\omega_{c_2})}\int_{-\infty}^{\infty}\sqrt{\rho(Y|\omega_{c_1})\rho(Y|\omega_{c_1})}dY \tag{4}$$

where, $\rho(\omega_{c_1})$, $\rho(\omega_{c_2})$, $\rho(Y|\omega_{c_1})$, and $\rho(Y|\omega_{c_1})$are the prior probabilities and conditional probabilities of the classes $c_1$and c_2. When the distribution of samples is unknown, the upper bound of the class error probability based on normal distribution is,

$$\partial_{c_1c_2} = \sqrt{\rho(\omega_{c_1})\rho(\omega_{c_2})}\ \exp\left(-\lambda(Y_2)\right) \tag{5}$$

where, $\lambda(Y_2)$is the Bhattacharya distance and is denoted as,

$$\lambda(Y_2) = \frac{1}{8}\theta(V_{c_1c_2}\ \beta_{c_1c_2}) + \frac{1}{2}\ln\frac{|V_{c_1c_2}|}{|\beta_{c_2}|^{\frac{1}{2}}\ |\beta_{c_1}|^{\frac{1}{2}}} \tag{6}$$

where, V_{c_1} and V_{c_2} is the expected within-class scatter matrices of classes $c_1$and c_2, respectively. The average of the V_{c_1} and V_{c_2} is denoted as, $V_{c_1c_2}$ and $\beta_{c_1c_2}$ denotes the between-class scatter matrix for c_1 and c_2.

$$V_{c_1} = E\left[(K - \Re_{c_1})(K - \Re_{c_2})^T\right] \tag{7}$$

$$V_{c2} = E\left[(K - \Re_{c_2})(K - \Re_{c_2})^T\right] \tag{8}$$

$$V_{c_1c_2} = \frac{V_{c_1} + V_{c_2}}{2} \tag{9}$$

$$\beta_{c_1c_2} = (\Re_{c_1} - \Re_{c_2})(\Re_{c_1} - \Re_{c_2})^T \tag{10}$$

where, $\Re_{c_1}$ and $\Re_{c_2}$ are the mean vectors of classes $c_1$and c_2, respectively. The upper bound of classification error beneath the dimensional reduction $\partial_{c_1c_2\psi}$ and the Bhattacharya distance is denoted as,

$$\partial_{c_1c_2\psi} = \sqrt{\rho(\omega_{c_1})\rho(\omega_{c_2})}\ \exp\left(-\lambda_{c_1c_2\psi}(Y_2)\right) \tag{11}$$

$$\lambda_{c_1c_2\psi}(Y_2) = \frac{1}{8}\theta\left[(\psi^T * V_{c_1c_2}\ \psi)^{-1}(\psi^T * V_{c_1c_2}\ \psi)\right] + \frac{1}{2}\ln\frac{|\psi^T V_{c_1c_2}\ \psi|}{|\psi^T V_{c_1c_2}\ \psi|^{\frac{1}{2}}\ |\psi^T V_{c_1c_2}\ \psi|^{\frac{1}{2}}} \tag{12}$$

To deal with the multiclass problems, the L normal distributed classes are employed that possess equal prior probabilities and hence, the upper bound of the Bayes error probability in the reduced gene space is given as the individual two-class pair as given as,

$$\rho(\omega_{c_1}) = \rho(\omega_{c_2}) = \frac{1}{N(N-1)} \tag{13}$$

$$\partial_{N\psi} = \sum_{c_1 > c_2}^{N}\sum_{c_2=1}^{N}\partial_{c_1c_2\psi} = \frac{1}{N(N-1)}\sum_{c_1 > c_2}^{N}\sum_{c_2=1}^{N}\exp\left[-\lambda_{c_1c_2\psi}(Y_2)\right] \tag{14}$$

where, ψ is the transform matrix. The equation (14) is differentiated with respect to the transform matrix and equated to zero that yields a highly non-linear equation of transform matrix that does not yield the analytical solution. Hence, the solution is provided using the recursive algorithm that depends on the gradient method that is given as,

$$\psi_{Re+1} = fn(\psi_{Re}) \tag{15}$$

where, $\psi_{Re+1}$is based on the recursive algorithm, Re denotes the number of the recursions, χ specifies the step size, and the gradient-based Bhattacharya coefficient is given as,

$$\psi_{Re+1} = \psi_{Re} - \chi.\nabla_{\psi}.(\partial_{N\psi}) \tag{16}$$

The gradient-based Bhattacharya distance facilities analytical analysis that is followed by the simpler computation and the recursive algorithm is iterated until the condition is attained. The condition for iteration is, $\|\psi_{Re+1} - \psi_{Re}\| \le Th$. Thus, the selected genes is given as, K^*that is of dimension, $(P \times e)$. The genes corresponding to the individuals are minimized using the Bhattacharya-based distance for dimensional reduction.

Cancer Classification using the proposed GOA-based DBN:

The classification of the cancer cells using the proposed method is depicted in this section. The proposed GOA-based DBN aims at computing the optimal weights of the DBN such that the global optimal solution regarding the presence or absence of the cancer is obtained. The classification accuracy is ensured using the proposed algorithm and the classifier, termed as GOA-based DBN, is a binary classifier that provides the class label as normal or abnormal. The main advantage of the proposed strategy is that the genes are employed for classification of the cancer cells. The proposed algorithm based on the error estimate ensures the decision at the global optimal level and with less convergence time. The architecture and the working of the proposed position update are depicted below.

Architecture of the Deep Belief Network:

DBNs are the generative neural networks with two layers, such as the multiple layers of Restricted Boltzmann Machines (RBM) and a Multi-Layer Perception (MLP) layer. Each of the layers possesses the input layers and the hidden layers with a series of the neurons. The hidden layer in the RBM acts as the output layer for the successive layers and the MLP layer possess the input layers, hidden layers, and output layers. The input to the DBN is the output from the gene selection block that offers the dimensionally reduced genes that are highly useful for the identification of the cancer cells. The cancer classification using the GOA-based DBN exhibits high classification accuracy and the selected genes ensure the higher value of the classification accuracy.

The architecture of the DBN is depicted in figure 2. The input layer and hidden layer of the RBM1 is given as,

$$a^1 = \left\{a_1^1, a_2^1, \ldots, a_l^1, \ldots, a_n^1\right\} \tag{17}$$

$$h^1 = \left\{h_1^1, h_2^1, \ldots, h_\lambda^1, \ldots, h_u^1\right\} \tag{18}$$

where, n and u are the total number of the input neurons and the hidden neurons present in the RBM1 of the DBN. The total number of the input neurons is based on the total number of the features, $n = (P \times e)$. The weights of the RBM1 is given as, $w_{l\lambda}^1$ and the dimension of the weights of the RBM1 is given as, $[n \times u]$. The output from the RBM1 is given as,

$$h_\lambda^1 = \omega\left[\varphi_\lambda^1 + \sum_\ell a_l^1 \times w_{l\lambda}^1\right] \tag{19}$$

where, φ_λ^1 is the bias of the λ^{th} hidden layer in the RBM1 and a_l^1 is the l^{th} input neuron of RBM1. The λ^{th} hidden neuron of RBM1 is denoted as, h_λ^1. The input layer and hidden layer of the RBM2 are given as,

$$a^2 = \left\{a_1^2, a_2^2, \ldots, a_\lambda^2, \ldots, a_u^2\right\} \tag{20}$$

$$h^2 = \left\{h_1^2, h_2^2, \ldots, h_\lambda^2, \ldots, h_u^2\right\} \tag{21}$$

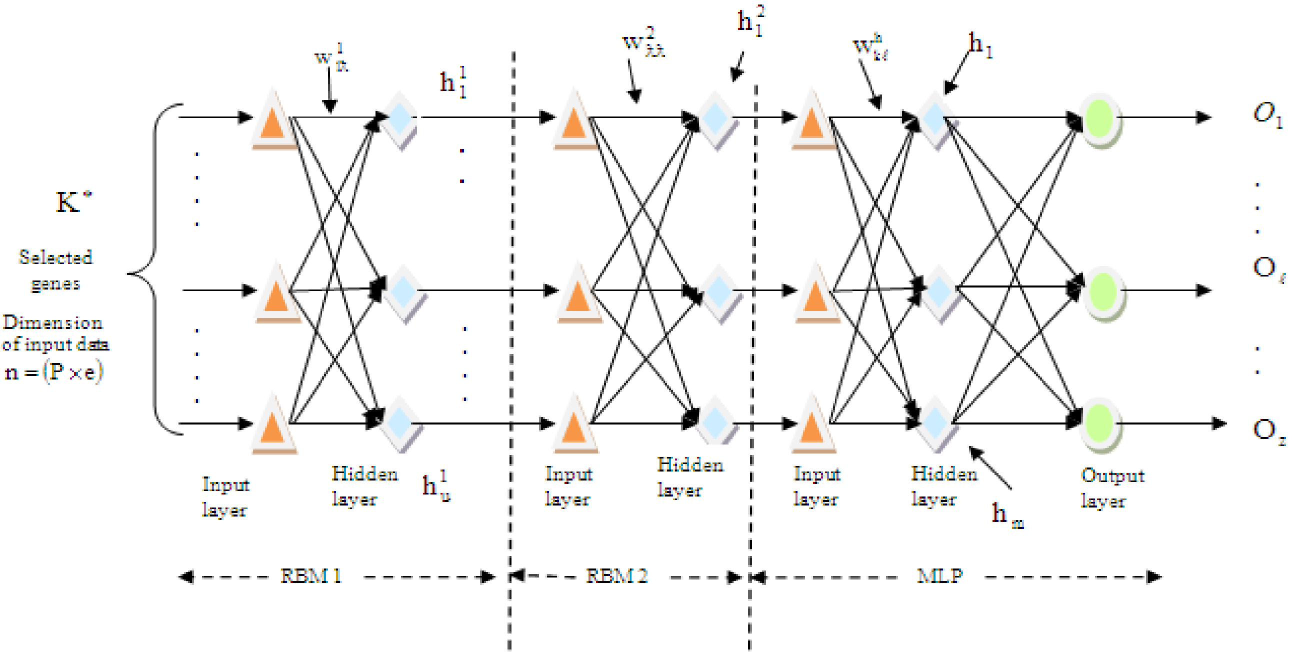


Figure 2: Architecture of DBN

The number of neurons in the hidden layer of the RBM1 is equal to the number of the neurons in the input layer of the RBM2. The weight of RBM2 is denoted as, $w^2 = \left\{ w^2_{\lambda\lambda} \right\}$. The weights between the λ^{th} visible neuron of RBM1 and the λ^{th} hidden neuron of the RBM2 is indicated as, $w_{\lambda\lambda}$. The output of the RBM2 is given as,

$$h^2_{\lambda} = \omega\left[\varphi^2_{\lambda} + \sum_{\lambda} a^2_{\lambda} \times w^2_{\lambda\lambda} \right] \forall a^2_{\lambda} \approx h^1_{\lambda} \tag{22}$$

The output from the hidden layer of the RBM2 is the input to the MLP layer. The input layer in MLP is given as,

$$M = \left\{ M_1, M_2, ..., M_{\lambda}, ..., M_b \right\} = h^2_{\lambda}; \left(1 \le \lambda \le b\right) \tag{23}$$

$$h = \left\{ h_1, h_2, ..., h_k, ..., h_m \right\}; \left(1 \le k \le m\right) \tag{24}$$

The output of the MLP layer is given as,

$$O = \left\{ O_1, O_2, ..., O_{\ell}, ..., O_z \right\} \tag{25}$$

where, z implies the total number of the outputs and O_{ℓ} denotes the output of the ℓ^{th} output neuron. The output from the MLP layer is given as,

$$O_{\ell} = \sum_{k=1}^{m} w^h_{k\ell} * h_k\ ; \left(1 \le k \le m\right); \left(1 \le \ell \le z\right) \tag{26}$$

where, $w^h_{k\ell}$ is the weight between the k^{th} hidden neuron and the ℓ^{th} output neuron and the dimension of the weights, $w_{k\ell}$ is $\left(m \times z\right)$. The output of the k^{th} neuron is given as,

$$h_k = \left[\sum_{\lambda=1}^{u} w_{\lambda k} * M_{\lambda} \right] B_k \forall M_{\lambda} = h^2_{\lambda} \tag{27}$$

where, B_k refers to the bias of the hidden neuron and $w_{\lambda k}$ denotes the weight between the λ^{th} input neuron and the k^{th} hidden neuron and the dimension is $\left(u \times m\right)$. The weights of the DBN are determined using the grasshopper optimization algorithm.

Training Phase of DBN

The training in the MLP is based on the GOA and the gradient descent that depends duly on the average error of the outputs. The RBMs undergoes unsupervised learning using the gradient descent, whereas the MLP employs the supervised learning using the proposed algorithm gradient descent-GOA. GOA-DBN is the introduction of DBN [19] [23] with the GOA [18] such that the weight update is performed optimally using the proposed algorithm.

Step 1: Training of RBM1 and RBM2: The training sample that consists of the selected genes is fed to the first layer of the RBM1 that calculates the probability distribution of the data and encodes the probability with the weights and the output obtained from the RBM1 is given as input to RBM2. The same procedure is performed in RBM 2 and the output is computed that enters as the input to the MLP layer.

Step 2: Training Phase of MLP layer: The steps involved in the training the MLP layer of the DBN is depicted below:

Initialization: The weights of the MLP layer are initialized randomly and let us denote the weights of the visible layer and the hidden layer as, $w_{k\ell}$ and $w_{\hbar k}$.

Read the input sample*:* The input to the MLP layer is the output from the RBM2 and is denoted as, h^2_{λ}.

Compute the output of the MLP layer*:* The output of the MLP layer is given as, h_k and O_{ℓ} and the outputs are based on the weight update equation using the gradient descent and the GOA.

Compute the error of the network*:* The error in the network is computed based on the average MSE and is given as,

$$e_{avg} = \frac{1}{n} \times \sum_{l=i}^{n} \left(O^l_{\ell} - O^l_{ground} \right); \ \left(1 \le \ell \le z\right) \tag{28}$$

where, n denotes the number of genes, O^l_{ℓ} is the output of the network, and O^l_{ground} denotes the desired output.

Weight Update in the MLP layer: The weights in the MLP layer are updated based on the error of the outputs from the ground truth and the algorithm corresponding to the less error is employed for the weight update. The weight update equation is based on,

$$w_{\text{ƛ}k}(t+1)=\begin{cases} w_{\text{ƛ}k}^{GOA}(t+1) & ; \quad \text{if } e_{avg}^{GOA}<e_{avg}^{GD} \\ w_{\text{ƛ}k}^{GD}(t+1) & ; \quad \text{Otherwise} \end{cases} \tag{29}$$

$$w_{k\ell}(t+1)=\begin{cases} w_{k\ell}^{GOA}(t+1) & ; \quad \text{if } e_{avg}^{GOA}<e_{avg}^{GD} \\ w_{k\ell}^{GD}(t+1) & ; \quad \text{Otherwise} \end{cases} \tag{30}$$

Weights update using the gradient descent: The weight update in visible and hidden layer and the weight update is performed by taking the partial derivative of the average error, e_{avg}. The derivatives of the equation (28) are employed to determine the incremental weights and are given as,

$$\Delta w_{\text{ƛ}k}^{GD}=-\eta\times\frac{\partial e_{avg}}{\partial w_{\text{ƛ}k}} \tag{31}$$

$$\Delta w_{k\ell}^{GD}=-\eta\times\frac{\partial e_{avg}}{\partial w_{k\ell}} \tag{32}$$

where, η indicates the learning rate. The weight update using the gradient descent is given as,

$$w_{\text{ƛ}k}^{GD}(t+1)=w_{\text{ƛ}k}^{GD}(t)+\Delta w_{\text{ƛ}k}^{GD} \tag{33}$$

$$w_{k\ell}^{GD}(t+1)=w_{k\ell}^{GD}(t)+\Delta w_{k\ell}^{GD} \tag{34}$$

Weights update of MLP using GOA: The weights of the MLP using the grasshopper algorithm (GOA) [18] are based on the characteristics of the grasshopper swarms that exhibit their social attraction. The advantage of the nature-inspired mechanisms in grasshopper includes seeking for food and the optimal solutions using the random operators enable the optimization algorithm to yield the global optimal solution. GOA is characterized using the exploration and exploitation phase that permits abrupt movements and avoiding the convergence to the local optimum. The weight update using GOA is given as,

$$\underset{j\in\{\text{ƛ}k,k\ell\}}{w_j^{GOA}}(t+1)=J\left[\sum_{\substack{i=j\\ i\neq j}}^{G}\left[J\times\frac{U_D-L_D}{2}\right]\times S\left(\left|y_i^D-y_j^D\right|\right)\times\left(\frac{y_i-y_j}{D_{ji}}\right)\right]+\hat{N}_D \tag{35}$$

where, G is the total number of the grasshoppers, J denotes the decreasing constant. The upper bound and the lower bound in the D^{th} dimensional space with $S(R)=g\times e^{-\frac{R}{A}}-e^{-R}$. The term $S(R)$ corresponds to the social forces. The position update of the j^{th} grasshopper is based on the position of the i^{th} grasshopper, distance between the two grasshoppers, and the decreasing coefficient. The decreasing coefficient J that reduces the attraction, comfort, and repulsion zone or in other words, it balances between the exploration and the exploitation, and the direction of the wind A is towards the best position of the target. In short, it is clear from equation (35) that the position update of a grasshopper depends on the current position of the grasshopper, position of the target, and the position of the other grasshoppers in the dimensional space. The term $\left[J\times\frac{U_D-L_D}{2}\right]$ minimizes the search space linearly so as to reduce the space of exploration and exploitation of the grasshoppers. Thus, GOA uses the position of all the search agents to define the best position of a grasshopper. The position of the i^{th} grasshopper is denoted as, y_i in the D^{th} dimensional space and $\hat{N}_D$ refers to the best solution in the D^{th} dimension, g denotes the intensity of the attraction, A refers to the attractive length scale, and R indicates the random number in the interval [0, 1]. The position update of the j^{th} grasshopper at the $(t+1)^{th}$ iteration is denoted as, $\underset{j\in\{\text{ƛ}k,k\ell\}}{w_j^{GOA}}(t+1)$ and the weights of the input-hidden layer, $w_{\text{ƛ}\ell}^{GOA}(t+1)$ and hidden-output layer, $w_{k\ell}^{GOA}(t+1)$ of the MLP are updated based on the above equation.

Re-compute the average error of the outputs of the MLP: The output of the network using the GOA weight update is employed to determine the average error of GOA and the average error is based on equation (28). If the error of GOA is less then, the weights of the MLP is updated using GOA and if the error of the network using the weights of the GOA is greater than gradient descent then, the weight update follows the gradient descent.

Termination: The iteration for the weight update is repeated for the maximum number of the iteration t.

The output from the classifier is either the presence or absence of the cancer and the genes that vary from the normal genes are responsible for the decision-making. The classification accuracy of the proposed method is improved and enhanced when compared with the existing methods. The proposed method conveys the disease-causing genes by comparison with the ground truth genes.

RESULT AND DISCUSSION

The section depicts the results and discussion of the proposed cancer classification strategy and the comparative discussion with the existing methods reveal the effectiveness of the proposed method.

Experimental setup:

The proposed cancer classification is implemented in MATLAB in a computer system that operates in Windows 10 Operating system with 4GB memory.

Performance Metrics:

The performance of the proposed classification method is evaluated using three metrics, such as False Alarm Rate (FAR), accuracy, and detection rate.

1 Accuracy: It is the measure of correctness of the detection as given as,

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{36}$$

where, TP is true positive, TN is true negative, FN is false negative, and FP is false positive.

2 Detection Rate: The sensitivity or the True Positive Rate (TPR) is defined as the number of positives identified correctly.

$$TPR = \frac{TP}{TP + FN} \tag{37}$$

3 False Alarm Rate (FAR): It depends on the specificity as given as,

$$\mathrm{FAR} = 1 - \mathrm{TNR} \tag{38}$$

The specificity or the True Negative Rate (TNR) is defined as the number of negatives identified correctly.

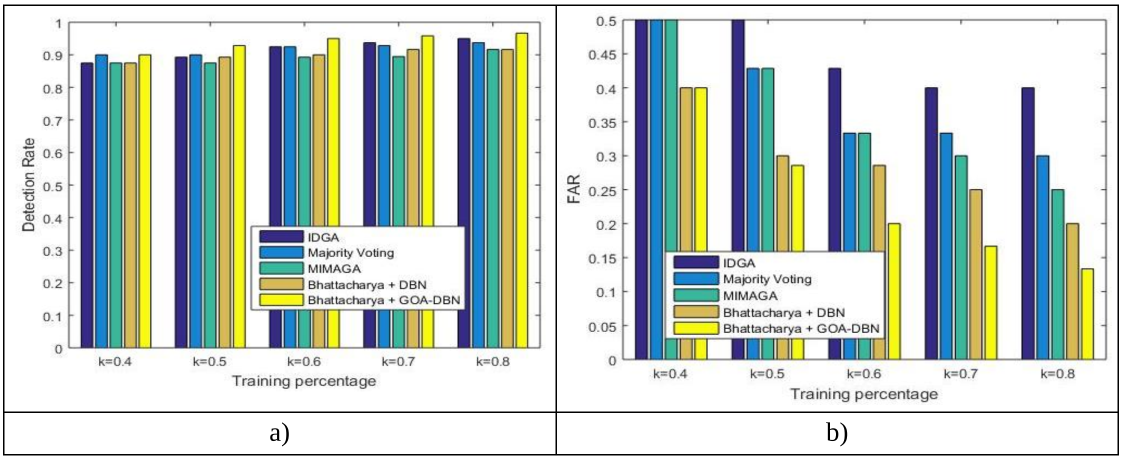
$$TNR = \frac{TN}{TN + FP} \tag{39}$$

Competing Methods:

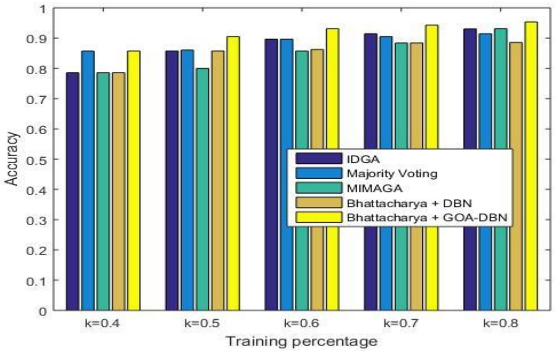
The performance of the proposed classification method Bhattacharya + GOA-DBN is compared with three existing works, such as IDGA [4], Majority Voting [1], MIMAGA [7], and Bhattacharya + DBN.

Datasets used

The datasets employed for analysis include the Colon dataset [21] and Leukemia dataset [22]. The Colon dataset comprises of the I2000 matrix with 2000 genes of the 62 tissues that are descending minimal intensity. This matrix consists of 20 pairs of features corresponding to the gene on the chip. The dataset 2 is a Leukaemia data that comprises of the gene expression data of 72 leukemia patients. The number of genes is 7128 stored in 7128x72 matrix.

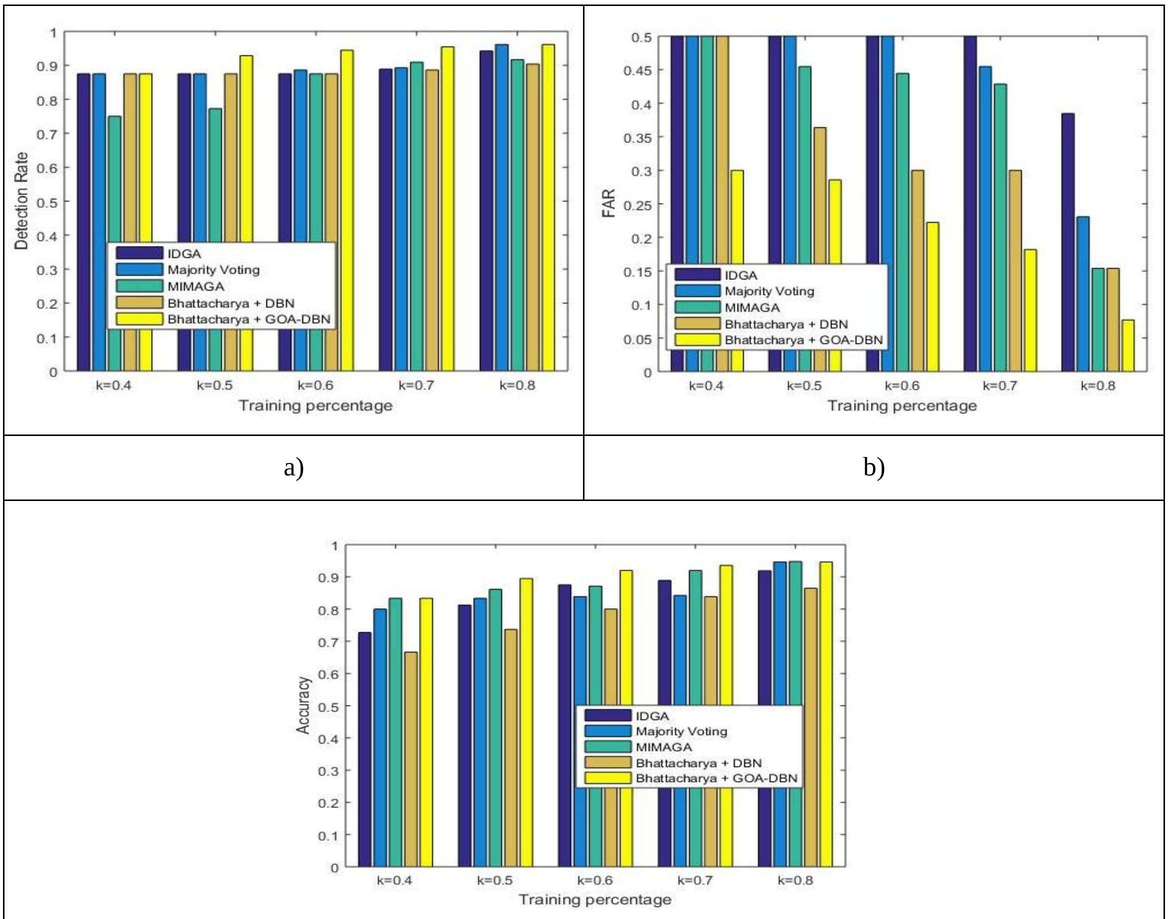


a) b)



c)

Figure 3: Analysis using the dataset 1 a) Detection rate b) FAR c) Accuracy

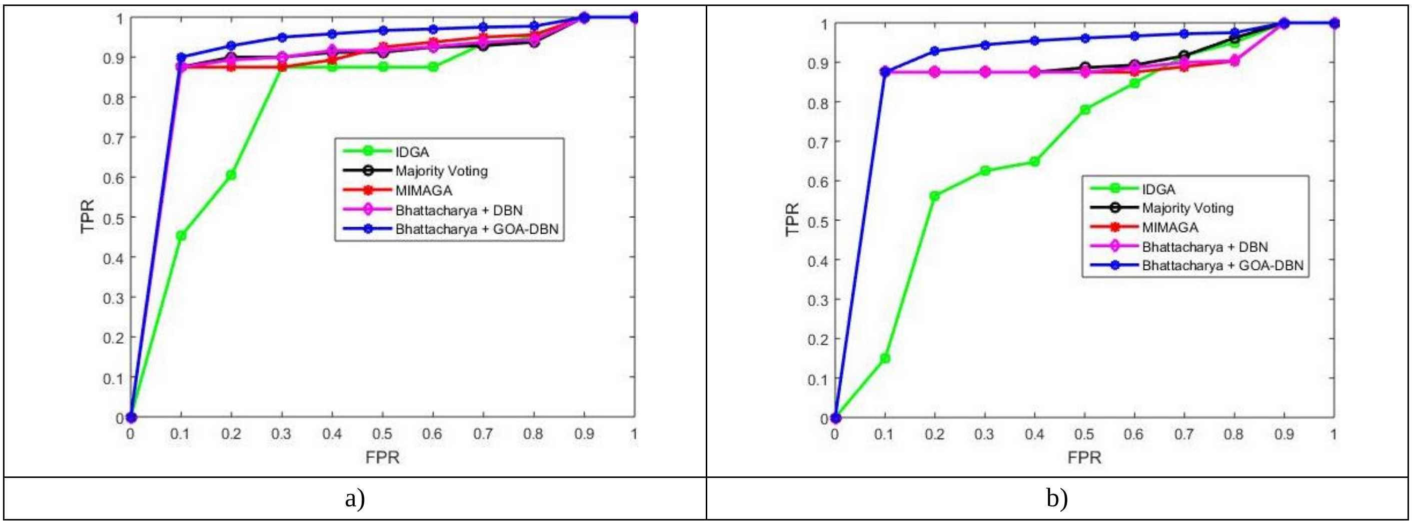


a)

b)

c)

Figure 4: Analysis using the dataset 2- Leukaemia data a) Detection rate b) FAR c) Accuracy



a) b)

Figure 5: ROC Curves a) dataset 1 b) dataset 2

Table 1: Comparative Discussion of the Classification Methods

Methods of Cancer Classification		Detection Rate	FAR	Accuracy
Dataset 1- Colon Tissue data	IDGA [4]	0.95	0.4	0.9302
	Majority Voting [1]	0.9375	0.3	0.9142
	MIMAGA [7]	0.9166	0.25	0.9310
	Bhattacharya + DBN	0.9166	0.2	0.8857
	Proposed Bhattacharya + GOA-DBN	**0.9666**	**0.1333**	**0.9534**
Methods		Detection Rate	FAR	Accuracy
Dataset 2- Leukaemia data	IDGA [4]	0.9423	0.3846	0.9189
	Majority Voting [1]	0.9615	0.2307	0.9459
	MIMAGA [7]	0.9166	0.1538	0.9473
	Bhattacharya + DBN	0.9038	0.1538	0.8648
	Proposed Bhattacharya + GOA-DBN	**0.9615**	**0.0769**	**0.9459**

Performance analysis using the comparative methods:

Figure 3 shows the analysis of the proposed method and the existing methods using the dataset 1 in terms of the performance metrics, such as detection rate, FAR, and accuracy. The analysis is carried out with respect to the training percentage. When the training percentage increases, the value of the detection rate also increases and at 0.8 as training percentage, the methods IDGA, Majority Voting, MIMAGA, Bhattacharya + DBN, and the proposed Bhattacharya + GOA-DBN obtained a detection rate at a rate of 0.95, 0.9375, 0.9166, 0.9166, and 0.9666, respectively, as depicted in figure 5 a). It is clear from figure 5 a) that the proposed method acquired greater rate of detection rate when compared with the existing methods. The FAR is pictured in figure 3 b) that shows that the FAR decreases with the increase in the training percentage. When the training percentage is 0.8, the FAR for IDGA, Majority Voting, MIMAGA, Bhattacharya + DBN, and the proposed Bhattacharya + GOA-DBN is 0.4, 0.3, 0.25, 0.2, and 0.133 that reveals that the proposed method acquired very less FAR when compared with the existing methods. Similarly, the accuracy of the methods is depicted in figure 3 c) that reveals that the accuracy increases with the increase in the training percentage. When the training percentage is 0.8, the accuracy of the methods IDGA, Majority Voting, MIMAGA, Bhattacharya + DBN, and the proposed Bhattacharya + GOA-DBN is at a rate of 0.9302, 0.9142, 0.9310, 0.8857, and 0.9534, respectively. The proposed method acquired greater accuracy than the existing methods proving that the proposed method is effective for the cancer classification compared with the existing methods.

Figure 4 shows the analysis of the proposed method and the existing methods using the dataset 2 in terms of the performance metrics, such as detection rate, FAR, and accuracy. The analysis is carried out with respect to the training percentage. When the training percentage increases, the value of the detection rate also increases and at 0.8 as training percentage, the methods IDGA, Majority Voting, MIMAGA, Bhattacharya + DBN, and the proposed Bhattacharya + GOA-DBN obtained a detection rate at a rate of 0.9423, 0.9615, 0.9166, 0.9038, and 0.9615 respectively that is depicted in figure 4 a). It is clear from figure 4 a) that the proposed method acquired greater rate of detection rate when compared with the existing methods. The FAR is pictured in figure 4 b) that shows that the FAR decreases with the increase in the training percentage. When the training percentage is 0.8, the FAR for IDGA, Majority Voting, MIMAGA, Bhattacharya + DBN, and the proposed Bhattacharya + GOA-DBN is 0.3846, 0.2307, 0.1538, 0.1538, and 0.0769 that reveals that the proposed method acquired very less FAR when compared with the existing methods. Similarly, the accuracy of the methods is depicted in figure 4 c) that reveals that the accuracy increases with the increase in the training percentage. When the training percentage is 0.8, the accuracy of the methods IDGA, Majority Voting, MIMAGA, Bhattacharya + DBN, and the proposed Bhattacharya + GOA-DBN is at a rate of 0.9189, 0.9459, 0.9473, 0.8648, and 0.9459 respectively. The proposed method acquired greater accuracy than the existing methods proving that the proposed method is effective for the cancer classification compared with the existing methods.

Figure 5 shows the discussion using the ROC curve using the dataset 1 and dataset 2 respectively in figure 5 a) and 5 b). In figure 5 a), for 10% of False positive Rate (FPR), the True Positive Rate (TPR) is 0.875 for IDGA, 0.875 for Majority Voting, 0.4523 for MIMAGA, 0.875 for Bhattacharya + DBN, and 0.9 for the proposed Bhattacharya + GOA-DBN. The proposed method attains a greater value of TPR for 10% of FPR. In figure 5 b), for 10% of False positive Rate (FPR), the True Positive Rate (TPR) is 0.875 for IDGA, 0.875 for Majority Voting, 0.15 for MIMAGA, 0.875 for Bhattacharya + DBN, and 0.875 for the proposed Bhattacharya + GOA-DBN. The proposed method attains a greater value of TPR for 10% of FPR.

Comparative Analysis:

The comparative discussion is performed for the maximum training percentage of 0.8 and the discussion is deliberated in table 1. It is clear from the table that the proposed method acquired the better value when compared with the existing methods. The discussion for the dataset 1 proves that the proposed method attained a maximum accuracy and Detection rate of 0.9534 and 0.9666, whereas the low FAR of 0.1333. The discussion using the dataset2 proves that the proposed method is better with a minimum FAR of 0.0769 and maximum accuracy of 0.9459. The accuracy of the existing methods like IDGA, Majority Voting, MIMAGA, and Bhattacharya + DBN is at a rate of 0.9189, 0.9459, 0.9473, and 0.8648 respectively. The discussion reveals that the proposed method possesses a better detection rate and maximum classification accuracy.

CONCLUSION

The cancer classification using the proposed Grasshopper Optimization Algorithm-based Deep Belief Networks is discussed in this paper. The cancer classification uses the gene expression data for undergoing the classification that enables the effective decision-making at the premises of the physician, prognosis, and diagnosis such that the death toll can be decreased. The gene expression data with the cancer-causing genes and other diseases are easily filtered out using the classification strategy, for which the useful genes that carry the useful information are required. The complexity is reduced and the accuracy of classification is improved using the Logarithmic transformation-based pre-processing scheme that enhances the gene expression data and the pre-processed gene expression data is dimensionally reduced to hold the highly informative genes using the Bhattacharya distance. Thus, the proposed classification strategy inherits a complexity free classification and there is no effect of redundancy and thereby, enhancing the performance of the classification. The analysis using the Colon tissue dataset and Leukaemia dataset proves that the proposed method overcomes the existing methods in terms of classification accuracy that is 0.9534, maximum detection rate at 0.9666 and minimum FAR at 0.0769, respectively. The proposed method of classification ensures burden-free classification using the DBN and achieves higher classification accuracy.

REFERENCES

[1] Sara Tarek, Reda Abd Elwahab and Mahmoud Shoman, “Gene expression based cancer classification,” Egyptian Informatics Journal, December 2016.

[2] Hanaa Salem, Gamal Attiya and Nawal El-Fishawy, “Classification of Human Cancer Diseases by Gene Expression Profiles,” Applied Soft Computing, Vol. 50, pp.124-134, January 2017.

[3] Boris Winterhoff , Habib Hamidi, Chen Wang, Kimberly R. Kalli, Brooke L. Fridley, Judy Dering, Hsiao-Wang Chen,William A. Cliby, He-JingWang, Sean Dowdy, Bobbie S. Gostout, Gary L. Keeney, Ellen L. Goode , Gottfried E. Konecny, “Molecular classification of high grade endometrioid and clear cell ovarian cancer using TCGA gene expression signatures,” Gynecologic Oncology, Vol.141, pp. 95–100,2016.

[4] M. Dashtban, Mohammadali Balafar, “Gene selection for microarray cancer classification using a new evolutionary method employing artificial

intelligence concepts," Genomics, Vol.109, pp.91–107, 2017.

[5] Sun-Yuan Hsieh and Yu-Chun Chou, "A Faster cDNA Microarray Gene Expression Data Classifier for Diagnosing Diseases," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.13, No.1, pp.43 – 54, 2016.

[6] Jin-Xing Liu, Yong Xu, Chun-Hou Zheng, Heng Kong and Zhi-Hui Lai,"RPCA-based Tumor Classification Using Gene Expression Data,"IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.12, No.4, pp. 964-970, 2015.

[7] Huijuan Lua, Junying Chena, Ke Yana, Qun Jina, Yu Xuec, Zhigang Gao, "A hybrid feature selection algorithm for gene expression data classification," Neurocomputing, pp.1–7, 2017.

[8] H. M. Alshamlan, G. H. Badr and Y. A. Alohali, "Genetic Bee Colony (GBC) Algorithm: A New Gene Selection Method for Microarray Cancer Classification", Computational Biology and Chemistry, Vol. 56, pp.49–60, 2015.

[9] E. Bard and W. Hu, "Identification of a 12-Gene Signature for Lung Cancer Prognosis through Machine Learning", Journal of Cancer Therapy, Vol. 2, pp.148-156, 2011.

[10] R. M. Luque-Baena, D. Urda, J.L. Subirats, L. Franco and J.M. Jerez, "Analysis of Cancer Microarray Data using Constructive Neural Networks and Genetic Algorithms", Ignacio Rojas & Francisco M. Ortuño Guzman, ed., 'IWBBIO' , Copicentro Editorial, pp.55-63, 2013.

[11] G. Chakraborty and B. Chakraborty, "Multi-objective Optimization Using Pareto GA for Gene-Selection from Microarray Data for Disease Classification", In Proceedings of IEEE International Conference Systems, Man, and Cybernetics (SMC), pp.2629 – 2634, 2013.

[12] RichardSimon, "Analysis of DNA microarray expression data," Best Practice & Research Clinical Haematology, Vol.22, No. 2, pp. 271-282, June 2009.

[13] H. M. Alshamlan, G. H. Badr, Y.A. Alohali, " The performance of bio-inspired evolutionary gene selection methods for cancer classification using microarray dataset," Int. J. Biosci. Biochem. Bioinform, Vol.4, No.3, pp.166–170, 2014

[14] A. Abderrahim, E. Talbi, and M. Khaled, "Hybridization of genetic and quantum algorithm for gene selection and classification of microarray data," In Proceedings of IEEE International Symposium In Parallel Distributed Processing, pp. 1–8, 2009.

[15] E. Alba and J. Garcia-Nieto et al, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," Evolutionary Computation, pp. 284–290, 2007.

[16] Jun S. Wei, Braden T. Greer, Frank Westermann, Seth M. Steinberg, Chang-Gue Son, Qing-Rong Chen,Craig C. Whiteford, Sven Bilke, Alexei L. Krasnoselsky, Nicola Cenacchi, Daniel Catchpoole, Frank Berthold,Manfred Schwab and Javed Khan, "Prediction of Clinical Outcome Using Gene Expression Profiling and Artificial Neural Networks for Patients with Neuroblastoma," Cancer Research, Vol.64, pp. 6883–6891, October 2004.

[17] Andrea Pellagatti, David Vetrie, Cordelia F. Langford, Susana Gama, Helen Eagleton, James S. Wainscoat, and Jacqueline Boultwood, "Gene Expression Profiling in Polycythemia Vera Using cDNA Microarray Technology," Cancer Research, Vol.63, pp. 3940–3944, July 2003.

[18] ShahrzadSaremi, Seyedali Mirjalili, and Andrew Lewis, " Grasshopper Optimisation Algorithm: Theory and application", Advances in Engineering Software, vol.105, pp.30-47, March 2017.

[19] Mostafa A. Salama; Aboul Ella Hassanien; Aly A. Fahmy, " Deep Belief Network for clustering and classification of a continuous data ", In Proceedings of the The 10th IEEE International Symposium on Signal Processing and Information Technology, pp.473 - 477, 2010.

[20] Guorong Xuan; Xiuming Zhu; Peiqi Chai; Zhenping Zhang; Yun Q. Shi; Dongdong Fu, " Feature Selection based on the Bhattacharyya Distance", In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), vol.3, pp.1232 - 1235, 2006.

[21] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, " Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays", vol. 96, no.12, pp.6745-6750, June 8, 1999.

[22] Leukemia data, " https://web.stanford.edu /~hastie /CASI_files/DATA/leukemia.html ", accessed on 07-11-2017.

[23] Yuming Hua; Junhai Guo; Hua Zhao, " Deep Belief Networks and deep learning", In Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things, pp.1-4, 2015.

[24] Thanh Nguyen, Saeid Nahavandi, Douglas Creighton, Abbas Khosravi, " Mass spectrometry

cancer data classification using wavelets and genetic algorithm", FEBS Letters, vol.589, no.24, pp.3879-3886, 21 December 2015.

[25] J. Herrero, R. D´ıaz-Uriarte and J. Dopazo, " Gene expression data preprocessing", Bioinformatics Applications Note, vol.19, no.5, pp.655-656, 2003.

[26] Praveen Tumuluru, Bhramaramba Ravi, "A Framework for Identifying of Gene to Gene Mutation causing Lung Cancer using SPI - Network", International Journal of Computer Applications, vol. 152, no. 10, pp. 21-26, October 2016.

[27] Praveen Tumuluru, Bhramaramba Ravi, Sujatha.Ch., Dr. N.Sudhakar "Credentials of Lung-Cancer Associated Genes Using Protein-Protein Interaction Network", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, No. 3, pp. 82-89, March 2016.

[28] Praveen Tumuluru, Bhramaramba Ravi "Dijkstra's based Identification of Lung Cancer Related Genes using PPI Networks", International Journal of Computer Applications (0975 – 8887), Vol. 163, No. 10, pp. 1-10, April 2017.

[29] Praveen Tumuluru, Bhramaramba Ravi "A Survey on Gene Expression Classification Systems", International Journal of Scientific Research and Review ISSN NO: 2279-543X, Volume 6, Issue 12, 2017.