ADAPTIVE SPARSE K-MEANS AND OPTIMIZATION ENABLED NEURAL NETWORK FROM GENE-EXPRESSION DATA FOR CANCER CLASSIFICATION

LAKSHMI RAMANI BURRA¹, BHARAMARAMBA RAVI², BANOTHU RAMJI³, APPARNA ALLADA⁴, PRAVEEN TUMULURU⁵

¹Assistant Professor, Dept. of CSE, PVP Siddhartha Institute of Technology, Vijayawada, ramaninyythili@gmail.com
²Professor, Dept. of CSE, GIT, GITAM Deemed to be University, Visakhapatnam, bharamarambaravi@gmail.com
³Assistant Professor, Dept. of CSE, CMR Technical campus, Hyderabad, Telangana, vasmramji@gmail.com
⁴Assistant Professor, Dept. of CSE, Gudlavalleru Engineering College, Gudlavalleru, alladaaparna9@gmail.com
⁵Assistant Professor, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, praveenluru@gmail.com

ABSTRACT

Cancer is one of the malignant diseases existing globally and the people affected with cancer are rescued only when the disease is recognized at the earliest possible stage. Identify in advance of disease is essential as in the final stage; since the chance of living/existence is partial. The indications of cancers are difficult and thus, all the indications should be considered accurately earlier to the diagnosis. Thus, an automatic prediction system is essential for classifying the tumor to malignant or benign. This work introduces a cancer classification approach using Chronological Grasshopper Optimization Algorithm (Chronological-GOA) for classification of cancer. For reducing the dimension of gene-expression data, log transformation is applied to the database. Then, the adaptive sparse K-means clustering selects the necessary data, which is provided to the Deep Belief Network (DBN). Here, the DBN is trained using Chronological-GOA. At last, the DBN classifies the selected gene sequences as normal and abnormal gene, and thereby identify the cancer. The performance of the cancer classification based on MSparse Kmeans + Chronological GOA-DBN is computed based on accuracy, detection rate, and False Alarm Rate (FAR). The developed method attains the accuracy of 0.9876, maximal detection rate of 0.9893, and the minimal FAR of 0.0596.

Keywords: Gene-expression data, Adaptive sparse K-means clustering, Chronological GOA, Deep belief network, log transformation.

1. INTRODUCTION

One of the dangerous diseases caused by most of the living organisms is cancer. Research on Cancers has been paid more attention in the field of medicine. With the rapid growth of microarray approach, the tumor classification research has been paid novel breakthrough in the recent years [4]. In the last few years, the mortality rate of cancer has been grown rapidly, causing a serious problem to human health. Uncontrolled proliferation and metastasis of cancer cells is challenging due to the identification of other types of cancer. Most of the cancers are diagnosed at the very initial stage [1], but becomes deadly at the final stage. Cancer classification is utilized for enhancing the health-care of patients, and life quality of the individuals, and, also it is suitable for drug discovery, and the diagnosis of cancer [2] [8] [29].

Gene is nothing, but the physical and functional unit of heredity and genes are made by Deoxyribonucleic acid (DNA). Few genes are act as instructions for making molecules termed proteins [3]. The gene expression level is utilized to solve the fundamental issues related to biological evolution mechanisms, drug discovery, prevention
and cure of diseases. Thus, the classification is performed with respect to some conditions based on morphological appearance and analysis of gene expression using the micro array data. Thus, the researchers focused their attention in the field of class prediction and class discovery [4][30]. The individual genes may be inactive or active based on the requirements and circumstances of the body cells under certain condition. Hence, any abnormalities of gene expression level may cause uncontrolled growth, death of cells [7][5]. Several recent researchers have been studying the issues of cancer classification based on deep learning methods [6][9]. Deep learning has been broadly utilized for image and audio applications [1]. There is the substantial amount of research with various machine learning approaches, like Radial Basis Function, Bayes Network, Nearest Neighbors algorithm, Single Conjunctive Rule Learner Decision tree and pruning.

This paper introduces automatic classification method using Chronological Grasshopper optimization algorithm (Chronological GOA) in which the chronological concept is integrated in GOA algorithm. Initially, the gene expression data is pre-processed using the Logarithmic transformation and the selection of gene is carried out using the proposed adaptive sparse K-means clustering. Then, the certain selected genes are provided to the Deep Belief Network (DBN), trained using chronological-GOA.

The main contribution of this work is the development of sparse fuzzy k-means clustering algorithm for selecting the suitable gene data and optimizing the weights for DBN classifier.

The organization of the paper is as follows: Section 2 analyses the existing methods of cancer classification. Section 3 discusses the proposed method and section 4 describes the results and discussion of the proposed method and finally, concludes the paper in section 5.

## II. LITERATURE REVIEW

The challenges and performance of cancer classification is reviewed with four methods.

Yang Guo et al. [14] designed deep learning model, termed as BCD Forest for cancer detection. Initially, multi-class-grained scanning method was introduced for training the multiple binary classifiers to encourage diversity of ensemble. Subsequently, the model fitting quality was considered in representation learning using the sliding window scanning and then, variation-based strategy was established for boosting the significant features in forest learning. At last, by using the deep learning the classification of cancer was done. The method failed to use various types of genomic data for better performance.

Ashok Kumar Dwivedi [15] developed a method using supervised machine learning for acute myeloid leukemia, and discrimination of acute lymphoblastic leukemia based on gene expression data. Here, ANN was employed for cancer classification.

Zhang, L. et al. [9, 34] proposed a model the Gene Interaction Regularized Elastic Net (GIREN) for classifying the gene-expression data and identifying existence rate of people affected by the disease. In GIREN model, classification is done by predicting suitable features sets for the classification. Further, the scheme developed iterative gradient descent algorithm for regularizing optimization process. The model makes the most of incorporated data environment for predicting the tumor cells, but sustained from the advanced computational burden.

Seah, C.S. et al. [13, 27, 28] developed the tuning parameter selection for gene prediction and thus, allowed cancer classification. The authors have introduced the directed random walk for classifying the gene expression data, and the scheme used pathway dataset for building the directed graph. The scheme improved sensitive predictions for cancer classification.

### 2.1 Challenges

The challenges encountered by existing methods are described and as follows,

- Cancer microarray gene-expression data is very puzzling because microarray data is high dimensional-low sample dataset with several irrelevant or noisy genes and missing data [6].
• In [14] [31], Boosting Cascade Deep Forest (BCD Forest) is developed for cancer classification. Here, the prediction accuracy is found better, but BCD Forest face ensemble diversity and over fitting challenges working on the small-scale biology data.

• The case of distinguishing several genes between the different types of cancers and classes with the removal of the insignificant genes is an excited assessment [13].

• Various challenges faced by classification algorithms in cancer classification are high dimensional, imbalanced noisy data with small sample size [17] [33]. Besides these challenges, classifiers need to reduce the classification cost.

III. PROPOSED CANCER CLASSIFICATION USING ADAPTIVE SPARSE K-MEANS AND CHRONOLOGICAL GRASSHOPPER OPTIMIZATION ALGORITHM-BASED DEEP BELIEF NETWORK

This section presents a method for cancer classification using adaptive K-means and chronological-GOA-based algorithm. The Figure (1) shows the schematic diagram of the proposed method for classification of cancer. In the beginning, the gene-expression data is given for pre-processing and it is performed using log transformation. Later, the gene selection is performed using adaptive sparse K-means clustering algorithm for dimensionality reduction and is indirectly associated with handling the complexity of classification. Finally, the classification is achieved based on the selected genes using Deep Belief Network (DBN) that are trained using Chronological-GOA.

3.1 Pre-processing Technique using Logarithmic transformation

Preprocessing is the primary step which is involved in the cancer classification. For complexity reduction, and enhancement in the classification accuracy, the input data is handled by using the log transformation that adapts the data into range of uniform value. Let us assume the gene-expression database $B$ of size $X \times Y$.

Log Transformation: The log transformation [17] [32] is the widely employed method for addressing skewed data, which is utilized in the field of psychosocial and biomedical research. Because of its ease of popularity and use, the log transformation is integrated in the statistical software packages, such as SAS, Splus and SPSS. The log transformation is beneficial for generating the patterns using more interpretable and helps fulfill the supposition, and to reduce skew, and for normalizing the data. The log transformation expression is given as,

$$K = \log_{10}(B)$$ (1)

The pre-processed dataset $B$ dimension is $X \times Y$.

3.2 Selection of Gene data using Adaptive Sparse K-means clustering

After pre-processing, the gene selection is performed based on adaptive sparse K-means clustering. The sparse K-means clustering algorithm clusters the nodes so that the communication between the nodes occurs through the cluster head and ensures the effective detection of the attacker node. Generally, in case of sparse K-means clustering [18, 25, 26], dimensional reduction is a major advantage, but the cluster error is increased. In order to mitigate the cluster error attained from K-means clustering, the cluster membership function is formed based on optimization algorithm. The pre-processed database $B$ of size $X \times Y$ is subjected to the gene selection. The adaptive sparse k-means clustering selects the optimal genes from the pre-processed data set $B$. The steps involved for the adaptive sparse K-means clustering are given below,

Step 1: In the first step, the gene Id is initialized as input with micro array expression value points. Let us consider the data points, expressed as $I = \{h_1, h_2, ..., h_q\}$, where $q = 1 to 1431$, and the patient Id is expressed as $T = \{I_1, I_2, I_3, ..., I_s\}$, where 

$s$ varies from 0 to 28. The set of clusters is denoted as $W = w_1, w_2, ..., w_q$. 

www.turkjphysiotherrehabil.org
Step 2: The cluster centers $C$ are selected randomly on each data point, and then the distance between each data point and cluster centers are computed. Here, the data point is denoted as $G(i)$, and the cluster centers are represented as $V(k)$. where, $i$ ranges from 1 to $q$, and $j$ varies between 1 and $s$.

Step 3: The optimization form based on clustering error are formulated, and is expressed as,

$$
X \in \left\{ \min_{0,1} \right\}^{C \times C} \left\{ \sum_{k=1}^{C} \sum_{j=1}^{q} X_{i,k} \right\}^{C} \sum_{k=1}^{C} \sum_{j=1}^{q} \frac{X_{i,k}}{q} \left( c_{k} - k \left( y_{i}^{*} \right) \right)^{2}
$$

(2)

where, $X$ denotes the cluster membership matrix which satisfies $X_{i,j} = 1$, if the data point $h_{i}$ belongs to the cluster $k$ and 0 otherwise, $\left\{ c_{k} \right\}_{k=1}^{C}$ represents the cluster center, and the factor $\frac{1}{q}$ is employed for time span.

Step 4: If $C < V(i)$, then distribute the data point to the cluster center $C(k)$.

Step 5: Then the new cluster center is recomputed by using the equation as,

$$
W_{i} = \left( \frac{1}{c_{i}} \right) \sum_{j=1}^{c_{i}} y_{i}
$$

(3)

where, $c_{i}$ indicates the no. of data points in $i^{th}$ cluster.

Step 6: After that, the $L_{1}$ regularization are applied to acquire the sparse optimal points.

Step 7: After the determination of sparse optimal points, the distance is recomputed between every data point and recently achieved cluster centers, and check the best solution is attained or not.

Step 8: The steps from 2 to 7 are repeated till the iteration reaches the maximum count. The pseudo code adaptive sparse K-means clustering is depicted in Algorithm 1.

Algorithm 1. Adaptive sparse K-Means clustering Algorithm
Input: \( K = \log_{10}(B) \)

Output: Best Features

Procedure:

Begin

1. Initialization: \( I = \{h_1, h_2, ..., h_q\} \)
2. Formulate the cluster centers \( C \).
3. Distance between each data point and the cluster centers are calculated
4. Solve the optimization form in terms of clustering error using equation (2)
5. Recompute the new cluster center using equation (3)
6. Apply \( L_1 \) regularization for sparse optimal weights.
7. Continue the iteration until the end of stopping criterion

Terminates

3.3 Cancer classification using Chronological GOA

Once the features are selected using adaptive sparse- \( K \)-means clustering, the particular genes are subjected to the proposed Chronological-GOA for cancer classification. For efficient gene-expression data classification, the DBN classifier model is used and the weights-biases are analyzed by using optimally the proposed algorithm.

3.3.1 Solution encoding

In optimization problems, the solution encoding determines the optimum solution. It achieves the optimum solution to the algorithms from a collection of solutions. The proposed system solution indicates a number of solutions, from which the best solutions are selected by using the fitness function.

3.3.2 Fitness evaluation

The fitness function for selecting genes based upon the proposed Chronological GOA is expressed as,

\[
Fitness = \left\{ E + \left[ 1 - \frac{R}{U} \right] \right\} \times \frac{1}{2}
\]

(2)

where, \( R \) and \( U \) signifies the chosen genes, and the number of genes present in the pre-processed database. The classification accuracy is denoted as \( E \) that depends on the output obtained from the DBN classifier.

3.3.2 Cancer classification using the proposed

**Chronological-GOA algorithm**

By integration of the chronological concept and GOA algorithm, the proposed Chronological-GOA is recommended to select the best solution. GOA [19] [35] algorithm can improve accuracy and effectively search for the best solution. Also, the algorithm exchanges complex operations, so the performance of the GOA algorithm is simple and flexible. The concept chronological is suitable for updating solutions based on past events related to time. Therefore, a method is provided to integrate the concept of time sequence into the GOA algorithm to find the best solution. The steps involved in the proposed Chronologic-GOA are as follows:

**Step 1: Initialization:** In the initialization process the first step is to initialize the solutions, which is represented as,

\[
M = \{M_1, M_2, ..., M_r, ..., M_p\}
\]

(3)
where, \( M_i \) indicates the \( r^{th} \) solution, and it varies between 1 and \( P \).

**Step 2: Fitness function evaluation:** By equation (2) the fitness function is evaluated for each solution. The solution provided that maximum fitness value is examined as the optimal solution.

**Step 3: Solution update:** The solution is updated based upon the Chronological concept. The term chronological signifies the past history. Here, the chronological-GOA algorithm includes the solution obtained at past iterations in current iteration. The past solutions are included in the solution update to enhance the convergence of optimization process.

The location update equation of the proposed chronological-GOA algorithm is expressed as,

\[
M(\tau + 1) = \frac{1}{2} \left[ 2k \sum_{i=1}^{P} \left( \frac{v_i - l_i}{2} \right) k \left| M_i - M_J \right| \frac{M_i - M_J}{d_{ij}} \right] + M(\tau) + M(\tau - 1) \tag{4}
\]

where, \( k \) signifies the decreasing constant to define the search space, \( d_{ij} \) represents the distance between \( r^{th} \) and the \( i^{th} \) grasshopper. The lower, and upper bounds for the search space is denoted as \( l_i \), and \( v_i \).

**Step 4: Find a feasible solution:** Determine an effective solution according to the fitness metric in equation (2) to test the probability of the new solution. The fitness function of the resulting solution is combined with the earlier best solution. Therefore, the location is updated until a better result is obtained.

**Step 5: Termination:** The steps (2) to (4) are repeated till the algorithm reaches the maximum iteration, represented as \( L_{\text{max}} \), and at the end of the iteration, the best sequences, denoted as \( R \), are chosen.

### 3.4 Cancer classification by training DBN using the proposed Chronological-GOA

The selected genes are then subjected to the DBN [16], which is trained by chronological-GOA for cancer classification. The DBN classifier is utilized for allowing deep learning of data patterns, and thus provides enhanced classification accuracy than other learning approaches. DBN is categorized into two RBM layers, and one MLP layer. The gene data selected is fed as the training sample to the DBN. Here, the RBM layer is trained using back propagation algorithm, whereas the MLP layer is trained by proposed Chronological-GOA algorithm. Figure 2 presents the block diagram of the proposed Chronological-GOA based DBN classifier developed for cancer classification.

As shown in figure 2, DBN comprises of three layers, such as RBM layer1, RBM layer 2, and MLP layer. Each layer consists of hidden neurons, output neurons, and input neurons. The selected training samples are provided as the input of the first RBM, and the output obtained from the first RBM is provided to the second RBM layer. Then, the result obtained from the second RBM is provided to the MLP layer.

#### 3.4.1 Deep Belief Network Training:

In this section, it identifies the proposed chronological-GOA-based DBN classifier training process. The RBM has unsupervised learning based on gradient descent, while MLP uses standard backpropagation algorithms to perform supervised learning methods.

**a) Training of RBM layers**

The selected genes are provided as input to the first RBM layer, and its weights are chosen by the back propagation algorithm. Inputting the output of first RBM to the succeeding layer, second RBM layer is processed. The weights are chosen by the back propagation algorithm for calculating the second RBM layer output. Using the computed weights, the output of second RBM is found.
Figure 2. Cancer classification with DBN classifier using the proposed Chronological GOA algorithm

b) Training of MLP layer

MLP layer is trained using the proposed algorithm to select the optimal weights for cancer classification. The training procedure is summarized as below,

1) At first, the training MLP layer is assigning the random weights to the hidden, and input layers of MLP, and the initialized weights are expressed as $K_{rx}^{F}$, and $K_{xv}^{I}$.

2) The output obtained from the second RBM layer is subjected to the MLP layer, and the training sample to MLP is derived in equation (19).

3) Then, based on the randomly initialized weights, the output of the MLP layer is calculated using equation (20).

4) After that, the average error is achieved based upon the variance among the attained output and the desired output as,

$$ E_{\text{average}} = \frac{1}{n} \sum_{i=1}^{n} (O_{i}^{c} - D_{i}^{c})^{2} ; 1 \leq c \leq y $$  \hspace{1cm} (22)

where, $O_{i}^{c}$ be the output obtained and $D_{i}^{c}$ denotes the desired output.

5) The weights of MLP layer are updated based on the proposed chronological-GOA algorithm, and the following equations show the update done in both layers, respectively.

$$ K_{rx}^{F}(\tau+1) = \frac{1}{2} \left\{ 2k \sum_{r=r'}^{k} \left( \frac{v_{r} - I_{r}}{2} \right) k \left[ K_{rx}^{F}(i) - K_{rx}^{F}(r) \right] \frac{K_{rx}^{F}(i) - K_{rx}^{F}(r)}{d_{n}} \right\} + K_{rx}^{F}(\tau) + K_{rx}^{F}(\tau - 1) $$  \hspace{1cm} (23)

$$ K_{xv}^{I}(\tau+1) = \frac{1}{2} \left\{ 2k \sum_{i=i'}^{k} \left( \frac{v_{i} - I_{i}}{2} \right) k \left[ K_{xv}^{I}(i) - K_{xv}^{I}(r) \right] \frac{K_{xv}^{I}(i) - K_{xv}^{I}(r)}{d_{n}} \right\} + K_{xv}^{I}(\tau) + K_{xv}^{I}(\tau - 1) $$  \hspace{1cm} (24)

where, $K_{vx}^{I}(\tau + 1)$ indicates the updated weight between the hidden layer, and input layers, and $K_{rx}^{F}(\tau + 1)$ refers to the weight updated between the output, and hidden layers of MLP layer using proposed chronological-GOA
algorithm. $K_{r_1}^F(i)$ and $K_{v_1}^I(i)$ signifies the weights at current iteration. The terms $K_{r_1}^F(r)$ and $K_{v_1}^I(r)$ specifies the weights corresponding to various solution at past iterations.

6) Finally, repeat steps (2) to (6) until the iteration reaches the maximum count. After reaching the maximum number of iterations, the best weight is obtained, which is considered the best weight of DBN.

3.4.2 Testing of Deep Belief Network

In the testing phase, while DBN receives the test data, it classifies the data into two classes, either normal or cancerous.

IV. RESULT OF DISCUSSION

The results of the improved method are given, and through comparative analysis, based on accuracy, detection rate and FAR, the effectiveness of the method is proved.

4.1 Experimental setup

The experimentation is performed in MATLAB with Windows 10 OS, with 2GB RAM and Intel i3 processor. The simulation of the proposed method requires some of its parameters to be tuned finely, and the values chosen are given as follows:

Parameters: step ratio = 0.01, dropout ratio = 0.05, initial momentum = 0.5, final momentum = 0.9. Maximum iteration = 50.

4.1.1 Database description


4.1.2 Performance metrics

The evaluation of the proposed technique is performed using three metrics, i.e. accuracy, detection rate, and FAR.

4.1.3 Comparative analysis

The methods used for analysis include: Bhattacharya + GOA-DBN [23], and chronological GOA-DBN [22]. The existing methods are compared with the proposed MSparse Kmeans + Chronologic GOA-DBN algorithm, and comparisons are made according to performance indicators, which involve accuracy, detection rate and FAR.

4.2 Comparative analysis

The Leukemia and Colon databases chosen

4.2.1 Comparative analysis using Leukemia database

Based upon the accuracy, detection rate, and FAR using Leukemia data set the comparative-analysis of the proposed method is examined and is shown in figure 3. Figure 3a) depicts the analysis based on accuracy by varying the training data percentages. When the training data percentage is 40, then the corresponding accuracy values estimated by existing Bhattacharya + GOA-DBN, Chronological GOA-DBN, and proposed MSparse Kmeans + Chronological GOA-DBN are 0.857, 0.9285, and 0.9297, respectively. The comparative-analysis based on decision rate is depicted in figure 3b). When the training data percentage is 50, the existing methods, like Bhattacharya + GOA-DBN, and Chronological GOA-DBN, acquires the detection rate of 0.928, and 0.9642, respectively. Meanwhile, the proposed MSparse Kmeans + Chronological GOA-DBN obtained the detection rate value of 0.9693. The figure 3c) shows the analysis based on FAR. When the percentage of training data is 60, the corresponding FAR values calculated by the existing Bhattacharya + GOA-DBN, Chronological GOA-DBN and the suggested MSparse Kmeans + Chronological GOA-DBN are 0.2, 0.1 and 0.073, respectively. According to the results, it should be noted that based on accuracy, detection rate and FAR, the performance of the proposed MSparse Kmeans + Chronologic GOA-DBN is better than other existing methods.

4.2.2 Comparative-analysis using Colon database

www.turkjphysiotherrehabil.org
The analysis of True Positive Rate (TPR) vs. False Positive Rate (FPR) is depicted in figure 3d). The TPR is 0.975 for Bhattacharya + GOA-DBN, 0.9875 for Chronological GOA-DBN, and 0.990 for the proposed MSparse Kmeans + Chronological GOA-DBN, when the FPR is 0.7.

### Comparative-analysis using Colon database

The proposed method is compared and analyzed which is based on accuracy, detection ratio and FAR using the colon dataset, and is shown in Figure 4. Figure 4a) shows the accuracy-based analysis by change in the percentage of training data. When the percentage of training data is 80, the corresponding accuracy values calculated by the existing Bhattacharya + GOA-DBN, Chronological GOA-DBN and the proposed MSparse Kmeans + Chronological GOA-DBN are 0.92, 0.96 and 0.972, respectively.

The decision rate comparative-analysis is shown in Figure 4b). When the percentage of training data is 70, the detection rates of existing methods such as Bhattacharya + GOA-DBN and Chronological GOA-DBN are 0.9545 and 0.9772, respectively. At the same time, the detection rate of the proposed MSparse Kmeans + Chronological GOA-DBN is 0.9840. Figure 4c) shows the analysis based on FAR.

When the training data percentage is 50, the corresponding FAR values calculated by the existing Bhattacharya + GOA-DBN, Chronological GOA-DBN and the proposed MSparse Kmeans + Chronological GOA-DBN are 0.2857, 0.1428, and 0.1349, respectively. According to the results, it can be found that the performance of the proposed MSparse Kmeans + Chronologic GOA-DBN is better than other existing methods based on accuracy, detection rate, and FAR.
Figure 4 Comparative-analysis of Colon database for different training data percentages (a) Accuracy, (b) Decision Rate, (c) FAR and (d) TPR vs. FPR

Table 1. Comparative discussion

<table>
<thead>
<tr>
<th>Databases</th>
<th>Comparative models</th>
<th>Evaluation metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Leukemia database</td>
<td>Bhattacharya + GOA-DBN</td>
<td>0.9534</td>
</tr>
<tr>
<td></td>
<td>Chronological GOA-DBN</td>
<td>0.9767</td>
</tr>
<tr>
<td></td>
<td>Proposed MSparse Kmeans + Chronological GOA-DBN</td>
<td>0.9844</td>
</tr>
<tr>
<td>Colon database</td>
<td>Bhattacharya + GOA-DBN</td>
<td>0.9459</td>
</tr>
<tr>
<td></td>
<td>Chronological GOA-DBN</td>
<td>0.9729</td>
</tr>
<tr>
<td></td>
<td>Proposed MSparse Kmeans + Chronological GOA-DBN</td>
<td>0.9876</td>
</tr>
</tbody>
</table>

The analysis of TPR vs. FPR is shown in figure 4d). The TPR is 0.9666, 0.9833, and 0.9860 for Bhattacharya + GOA-DBN, Chronological GOA-DBN, and the proposed MSparse Kmeans + Chronological GOA-DBN, when the FPR is 0.6.

4.3 Comparative Discussion

Table 1 shows the comparative discussion to reveal the best performance achieved by the cancer classification methods, based upon accuracy, decision rate, and FAR using Leukemia, and colon dataset. The existing Chronological GOA-DBN technique achieved values of 0.9767, 0.9666, and 0.4 for accuracy, FAR, and detection rate using Leukemia database. For Colon database, the Chronological GOA-DBN model attained values of 0.9729, 0.9807, and 0.0769 for accuracy, detection rate, and FAR. Meanwhile, for the colon database, the proposed MSparse Kmeans + Chronological GOA-DBN...
DBN has achieved the accuracy, detection rate, and FAR of 0.9876, 0.9850, and 0.0596, and for the leukemia database, the proposed MSparse Kmeans + Chronological GOA-DBN has attained the accuracy, detection rate, and FAR of 0.9844, 0.9666, and 0.4.

V. CONCLUSION
This work presents an approach for cancer classification by employing MSparse Kmeans + Chronological GOA-DBN. Initially, the gene-expression data is provided to pre-processing to reduce the length of gene sequences using log transformation. After pre-processing, the gene selection is performed by using adaptive sparse K-means clustering algorithm for dimensionality reduction. Then, the selected genes are given to the classification step. Here, the classification is performed using the proposed Chronological-GOA, which uses the past histories, while updating the solution and this makes the selection to be optimal. Then, optimal gene data selected by proposed Chronological GOA is given for DBN network, and training is done accordingly. At last, the DBN classifier processes the gene expression data and categorizes it as a normal gene or abnormal gene. Experimentation is carried out using two databases, viz. Leukemia database, and the Colon database. The performance of the MSparse Kmeans + Chronological GOA-DBN is evaluated based on accuracy, detection rate, FAR. The proposed method generates the maximal accuracy of 0.9876, maximal detection rate 0.9893, and the minimal FAR of 0.0596, that specifies the significance of the proposed method.

REFERENCES