Microarray Gene Expression and Multiclass Cancer Classification using Extreme Learning Machine (ELM) with Refined Group Search Optimizer (RGSO)

R. Balakrishnan and Thirunavu Karthikeyan

Abstract--- Microarrays can be utilized to determine the comparative amount of particular mRNAs in two or more tissue samples for thousands of genes concurrently. As the supremacy of this technique has been identified, various open queries arise about suitable examination of microarray data. The multicategory cancer classification is playing a vital role in the field of medical sciences. As the numbers of cancer victims are increasing steadily, the necessity of the cancer classification techniques has become indispensible. In this research, initially preprocessing and normalization process is carried out to select the best gene datasets. Then, a combination of Advanced Integer-Coded Genetic Algorithm (AICGA) and Extreme Learning Machine (ELM), with refined group search optimizer (RGSO) technique is used for gene selection and cancer classification. AICGA is used with RGSO Based ELM classifier to choose an optimal set of genes which results in an efficient hybrid algorithm that can handle sparse data and sample imbalance. The refined group search optimizer based extreme learning machine is used to carry out the classification process. In the proposed RGSO based ELM, the weights and bias to ELM are optimized using RGSO for better simplification and classification of large value of gene datasets. The performance of the proposed approach is evaluated and the results are compared with existing methods. The proposed approaches are applied for real time datasets and benchmark datasets taken from dataset repositories.

Keywords--- Extreme Learning Machine, Integer-Coded Genetic Algorithm, Gene Selection, Classification, Refined Group Search Optimizer

I. INTRODUCTION

Cancer detection and classification for diagnostic and projecting use is generally based on pathological analysis of tissue sections, resulting in subjective analysis of data [1]. The partial information gained from morphological analysis is often not enough to aid in cancer diagnosis and may result in expensive but ineffective treatment of cancer. To exactly identify cancer subtypes, recent studies have been carried out to identify genes that may cause cancer [2].

Some major issues related to cancer classification using microarray data are: robustness of gene selection and gene ranking, understanding of issues related to feature selection and performance evaluation of the selected genes [3]. Taxonomy, potential use, and variety of feature selection techniques are discussed in [2]. Recent literature [3] states that thousands of samples are required for robust gene selection, in order to have overlapping sets of genes. Our method of gene selection and classification shows that it is possible to get good classification results with a small set of samples. It can be analyzed that different sets of genes, which have few genes in common, can classify a wide variety of cancer types with high accuracy. This is made

R. Balakrishnan, Assistant Professor, Dr.NGP Arts and Science College, Coimbatore. E-mail:balakrishnan.scholar@yahoo.com

Thirunavu Karthikeyan, Associate Professor, P.S.G. College of Arts and Science, Coimbatore. E-mail:t.karthikeyan.gasc@gmail.com

feasible by the selection of genes that have high biased power and the exploit of a classifier that is robust enough to deal with the imbalances in the given data set. A biological analysis of the selected genes was made an effort to examine the functional nature of the genes, which may give explanation why different sets of genes are still able to effectively classify a wide variety of cancer types. Conventional gene selection methods are of two types, that is the filtering approach and the wrapper approach. In the filtering approach, selected genes are self-governing of the choice of classification methods, where in the wrapper approach; gene selection is influential on the choice of the classifier. A complete analysis of these methods has been presented in [4]. Generally, one uses the filtering approach on data, where large number of samples is presented. Recursive feature elimination with Support Vector Machine (SVM) was used in [6, 7], for cancer classification, by means of the Global Cancer Map (GCM) data set [5, 8]. A comparison of all the popular classification approaches for different data sets is analyzed by Statnikov et al. [9]. In [12] the use of a particle swarm optimization (PSO) and a genetic algorithm (GA) for the classification is compared for high dimensional microarray data. Both algorithms are employed for identifying small samples of informative genes between thousands of them. In [13] sparse principal component analysis (PCA) is employed to solve clustering and feature selection problems. Sparse PCA looks for sparse factors, or linear combinations of the data variables, clearing up a maximum amount of variance in the data while having only a limited number of nonzero coefficients. PCA is frequently used clustering technique and sparse factors allow them to interpret the clusters in terms of a reduced set of variables. Recently, Saras Saraswathi et al [15] proposed a novel combination of Integer-Coded Genetic Algorithm (ICGA) and Particle Swarm Optimization (PSO), coupled with the neural-network-based Extreme Learning Machine (ELM), is used for gene selection and cancer classification. ICGA is used with PSOELM to choose an optimal set of genes, which is then utilized to generate a classifier that handles sparse data and sample imbalance.

In this paper, a better gene selection and cancer classification technique is proposed for microarray data that is described by sample sparseness and imbalance. The microarray data includes several classes of cancers that are classified continuously as different to the existing traditional classification methods, where one class is exposed next to all the other classes. In this paper, an Advanced Integer-Coded Genetic Algorithm (ICGA) [10] is used for strong and healthy gene selection. Next, propose an Extreme Learning Machine (ELM), with refined group search optimizer (RGSO) [11] technique for managing the sparse/imbalanced data classification problem.

II. PROPOSED METHODOLOGY

The proposed methodology of the block diagram is shown in the figure 1 as follows

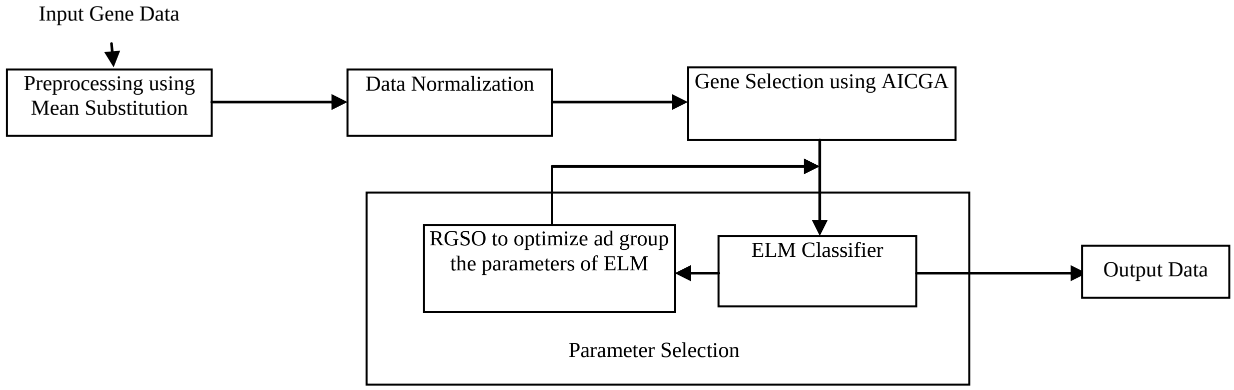

Input Gene Data
Preprocessing using Mean Substitution
Data Normalization
Gene Selection using AICGA
RGSO to optimize ad group the parameters of ELM
ELM Classifier
Output Data
Parameter Selection

Fig. 1: Proposed Block Diagram

Initially, the preprocessing process is carried out to find missing values of the datasets. After that the output of the preprocessing data is normalized to obtain the scaled dataset. Then the performance of ELM classifier is mainly based on the selected input genes. In order to minimize the computational aspect, an AICGA is used to choose and minimizes the number of genes, which can discriminate the cancer classes efficiently. Here extreme learning machine (ELM), with refined group search optimizer (RGSO) technique is proposed, to select the best parameters for better simplification and training of the classifier for gene data. Based on these chosen genes, ELM algorithm generates significant classifier by calculating weights of the genes. Initially, AICGA selects n independent genes from the available gene set. In the proposed RGSO based ELM, the weights and bias to ELM are optimized using RGSO for better simplification and classification. For the selected genes, RGSO will identify optimal parameters like number of hidden nodes and input weights such that the performance of the ELM multiclass classifier is improved. The best validation performance (η+) will be utilized as fitness for the AICGA evolution. The validation performance of ELM classifier (η) is used in RGSO for selection and grouping of ELM parameters.

A. Preprocessing of the Data

The data preprocessing approaches have a significant influence on the performance of machine learning algorithms. To produce quality mining results, data preprocessing is very important. The challenging problem in machine learning and data mining is missing values imputation [16]. High-quality database design and analysis can reduce the missing data problems. An appropriate technique should be selected to handle missing values depending on problem domain and the goal. In this paper, a mean substitution approach is used to impute missing values and data scaling algorithm to improve the accurateness of the classification performance of the entire system.

a. Mean Substitution

The imputation method is to fill in the missing data values is to use a variable's mean or median. The following algorithm explains the proposed form of mean substitution method [17]

```
Let
D = { A1, A2, A3, ….. An }
Where
D is the set of data with missing values
Ai – is the ith attribute column of values of D with missing values in some or all columns
n - is the number of attributes.
Function MeanSubstitution(D)
Begin
For i=1 to n {
$ai \leftarrow Ai \cap mi$
where
ai is the column of attributes without missing values
mi is the set of missing values in Ai (missing values denoted by a symbol)
Let i be the mean of ai
Replace all the missing elements of Ai with i
}
At last will have the imputed data set
End
```

b. Data Normalization

Normalization is a scaling down transformation of the samples. Within that sample there is frequently a large difference between the maximum and minimum values. When normalization is carried out the value magnitudes and scaled to significantly low values [18].

```
The Data Scaling Algorithm
Let
D = { A1, A2, A3, ….. An }
Where
```

```
D is the set of unnormalized data
Ai – is the ith attribute column of values of
m- is the member of rows (records)
n - is the number of attributes.
Function Normalize(D)
Begin
For i=1 to n {
$\mathrm{Maxi} \leftarrow \max(\mathrm{Ai})$
$\mathrm{Mini} \leftarrow \min(\mathrm{Ai})$
For r =1 to m {
$\mathrm{Air} \leftarrow \mathrm{Air} - \mathrm{Mini}$
$\mathrm{Air} \leftarrow \mathrm{Air} / \mathrm{Maxi}$
Where
Air is the element of Ai at row r
}
}
Finally will have the scaled data set
End y
```

B. RGSO Based ELM for Gene Classification

During recent years in medical analysis, artificial neural networks [19] participates a most important significant role for feature selection of the gene and solves image classification problem in various applications. Because of the quick convergence time and less number of data is required for training data in the classification. When compare to other classification methods the performance of ANN is high and less completion time. In earlier several number of the neural network algorithms [19] such as radial basis function neural network (RBFNN), probabilistic neural network (PNN), back propagation neural network (BPNN), and support vector machines (SVM) is used for the classification of medical and image data in efficient manner. The major issue occurs all of these methods are it requires more time to preprocess the data in the network structure and it is applicable for only less number of the training samples.

In order to overcome this problem and get better classification accuracy for neural networks algorithm with more number of the training data by using Extreme learning machine (ELM) [20]. The proposed extreme learning machine (ELM) classifier which holds the training for particular hidden layer feed forward neural networks (FFNN).

a. Basic Extreme Learning Machine (ELM) Classifier

Extreme Learning Machine (ELM) [20] meant for Single Hidden Layer Feed-Forward Neural Networks (SLFNs) will randomly select the input weights and systematically find out the output weights [21]. This algorithm tends to pay for the best performance at extremely fast learning speed.

ELM consists of an input layer, hidden layer and an output layer. Also, the ELM has many attention grapping and important features different from conventional popular learning algorithms for feed forward neural networks. These include the following:

- The training speed of ELM is extremely fast when put next to different classifier. The training process of ELM can be performed in seconds or less than seconds for numerous applications.
- The ELM will achieve the results directly with none difficulties. The ELM training algorithm is much easy than the other learning

The ELM algorithm which consists of three steps that can be summarized as

Step 1: Given a training set $\aleph = \{(X_i, t_i) | X_i \epsilon R^m, i = 1, \dots \dots, N\}$ activation function g(x), and hidden number node $\widetilde{N}$,

1) Give random hidden nodes through randomly generating parameters (a_i, b_i) according to any continuous sampling distribution, $i = 1, \dots., \widetilde{N}$

2) Calculate the hidden layer output matrix H.

3) Calculate the output weight β:$\tilde{\beta} = H^{+}T$

Then find the maximum repeated data in the whole processes. Where H^{+} is the Moore-Penrose generalized pseudo inverse of hidden layer output matrix.

b. Group Search Optimizer (GSO) Algorithm

The group search optimizer algorithm [22] is derived from the biological producer-scrounger (PS) model, which considers group members search either for “finding” (producer) or for “joining” (scrounger) chances. Animal scanning mechanisms are built-in to enlarge the GSO algorithm. GSO also uses “rangers” which perform random walks to keep away from entrapment in local minima. In GSO algorithm, the population is called a group and each individual in the population is called a member. The GSO algorithm is implemented for this work because of its nature of random walk in various directions. The movements of the members to find the solution are processed in a fast way by eliminating the less efficient members in the group. This results in precise and earlier convergence of the proposed algorithm.

There are three types of members in a GSO group they are producers, scroungers, and dispersed members. There is merely one producer and other members are either scroungers or dispersed members. Dispersed members carried out a random walks which are lesser significant. At each iteration, a group member is present to bestow the best fitness value, and is chosen as the producer. The other group members are chosen as scroungers or rangers by random select. After that, each scrounger takes a random move towards the producer, and each ranger takes a random move in the arbitrary direction.

In n-dimensional search space, the ith member at kth iteration has a current position $X_l^k \epsilon R^n$ and a head angle $\varphi_i^k = (\varphi_{i1}^k, \dots, \varphi_{i(n-1)}^k \epsilon R^{n-1}$. Search direction of ith member is a unit vector $D_i^k\left(\varphi_i^k\right) = (d_{i1}^k, \dots, d_{in}^k) \epsilon R^n$ that can be calculated from φ_i^k via a Polar to Cartesian coordinate transformation.

At kth iteration the producer X^p performs as follows.

1. The producer will scan at zero degree and after that scan laterally by randomly sampling three points in the scanning field: one point at zero degree,

$$X_z = X_p^k + r_1 l_{max}\, D_p^k(\varphi^k)$$

one point in the right hand side hypercube

$$X_r = X_p^k + r_1 l_{max}\, D_p^k(\varphi^k + \frac{r_2\theta_{max}}{2})$$

and one point in the left hand side hypercube

$$X_l = X_p^k + r_1 l_{max}\, D_p^k(\varphi^k - \frac{r_2\theta_{max}}{2})$$

$\theta_{max} \in R^1$ is maximum pursuit angle and $l_{max} \in R^1$ is maximum chase distance. $r_1 \in R^1$ is a normally distributed random number with mean 0 and standard deviation 1 and $r_2 \in R^{n-1}$ is a uniformly distributed random sequence in the range (0, 1).

The producer will then find the best point with the best resource (fitness value). If the best point has a better resource than its current position, then it will fly to this point or it will stay in its current position and turn its head to a new randomly generated angle. Let us, consider

$$\varphi^{k+1} = \varphi^k + r_2 \propto_{max}$$

Where $\propto_{max} \in R^1$ is the maximum turning angle.

2. If the producer cannot find a best area after iterations, it will return its head back to zero degree,

$$\varphi^{k+a} = \varphi^k$$

Where $a \in R^1$ is a constant.

At kth iteration, ith scrounger walks randomly towards the producer. Consider

$$X_i^{k+1} = X_i^k + r_3 o\, (X_p^k - X_i^k)$$

Where $r_3 \in R^n$ is a uniform random sequence in the range (0, 1). Operator o is the Hadamard product or the Schur product, which computes the entry wise product of the two vectors. If a scrounger identifies a better location when compared to the current producer and other scroungers, then it will change as producer in the next iteration.

The group members, who are less significant foragers than the dominant one, will be isolated from the group members. If the ith group member is isolated, it will process ranging. At the kth iteration, it generates a random head angle $\varphi_i$through (4); and then it chooses a random distance

$$l_i = ar_1 l_{max}$$

and moves to the new point

$$X_i^{k+1} = X_i^k + l_i D_i^k(\varphi^{k+1})$$

To maximize their chances of finding resources, the GSO algorithm makes use of fly-back mechanism to deal with the particular problem constraints. When the optimization process is initiated, the members of the group search for the solution in an efficient way. If any member moves into the inefficient area, it will be required to move back to the earlier position to guarantee a feasible solution.

c. Developed Refined GSO (RGSO)

In the traditional GSO, 75% rest members will perform scrounging and the remaining 25% of members will process for ranging. In this research, the ranging operation for the remaining 25% of members will not be completed as like the original GSO. As an alternative, these members for ranging operation will study from the “worst” member in its group. This improvement of learning from “worst” member leads to finding better solution regions in complex optimization search spaces.

While compared to the original GSO [22], RGSO algorithm searches for best regions to find the global optimum solution. The difference between GSO and RGSO is that the differential operator is functional to only accept the fundamental GSO generating new better solution for each krill in preference to accepting all the krill updating adopted in krill herd (KH). This is similar to greedy approach. The original GSO is well-organized and influential but greatly tends to premature convergence. As a result, to avoid premature convergence and additional improvement of the original GSO, a differential direction is employed to valve the valuable information in all krill individuals to update the position of a particular krill individual. Equation (18) shows the differential mechanism.

$$Z_i - Z_j = (z_{i1}\ z_{i2}\ \ z_{i3}\ \ \ \dots\ z_{in}) - \left(z_{p1}\ z_{p2}\ \ z_{p3}\ \ \ \dots\ z_{pn}\right)$$

whereas z_{i1} is the primary element in the n dimension vector Z_i. z_{in} is the nth element in the n dimension vector Z_i. z_{p1} is the primary element in the n dimension vector Z_p. ρ is the random integer generated individually for each z, between 1 and n, but $p \neq i$.

Therefore, the refined GSO is shown in Algorithm 1. As the problem of interest in this research is difficult in nature, the above refined GSO will be able to discover better genes to train the ELM for gene classification.

d. Proposed RGSO Based ELM for gene Classification

This proposed method combines the idea of RGSO for optimizing the weights in ELM neural network. The refined GSO combined with ELM facilitates the selection of input weights to increase the simplification performance and the training of the single layer feed forward neural network. RGSO based ELM and AICGA based gene selection approach is proposed in this research, which can minimize the size through gene (feature) selection and use the chosen relevant genes for accurate classification of a sparse and imbalanced data set. The proposed ELM classifier can differentiate the cancer classes amongst the data indicate the chosen features in fast manner. The proposed classifier in which the proposed RGSO algorithm is employed to find the optimal input weights such that ELM classifier can differentiate the cancer classes considerably, that is, the performance of the ELM classifier is improved. The data are separated into training and predicting sets in this research. Based on the input and output weights obtained by training data, the data can be estimate directly by the established ELM.

The steps of the proposed approach are as follows.

Step 1: Initialize the gene positions and head angles with a set of input weights and hidden biases of the gene data as given below

$$[W_{11}, W_{12}, \dots, W_{1n}, \dots, W_{21}, W_{22}, \dots, W_{2n}, \dots, W_{H1}, W_{H2}, \dots, W_{Hn}, b_1, b_2, \dots b_H]$$

These will be randomly initialized within the range of $[-1, 1]$ on D dimensions in the search space.

Step 2: For each member in the group, the particular output weights are calculated at ELM as given in (9).

Step 3: At this instant invoke refined GSO based on AICGA

Step 4: Then the fitness of each member, is measured

Step 5: Find the producer of the group based on the fitness of the data.

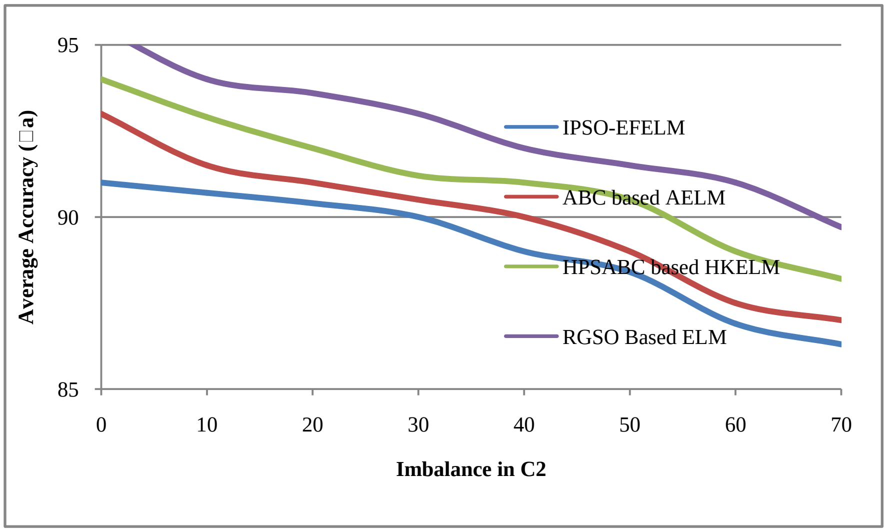
Step 6: Update the position of each member as given in (13) to (17).

Step 7: Stopping criteria: the algorithm does again the Steps 2–6 until the stopping criteria are met, along with hard threshold value as maximum number of iterations. On meeting the stopping condition, the algorithm returns the optimal weights with optimal fitness as its solution. Thus refined GSO (RGSO) with ELM finds the best optimal weights W and bias b so that the fitness reaches the minimum to achieve better generalization performance, with minimum number of hidden neurons, considering both the advantages of both ELM and RGSO. In the process of selecting the input weights, the refined GSO considers not only the best genes on validation set but also the norm of the output weights [23]. The proposed RGSO based ELM will combine the feature of RGSO into ELM to compute the optimal weights and bias to make the best genes minimal.

C. Analysis on Imbalance Data

The sample imbalance handling capacity of RGSO Based ELM classifier is based on the technique in [16]. The number of samples in one of the class was reduced and performance of the classifier was examined for different imbalance criteria. A similar examination was conducted for the proposed classifier and the average (η_a), overall $(\eta_0)$and individual (η_2)classification efficiencies obtained are shown in Fig. 2.

It is observed that the average and overall classification efficiency of proposed classifier is almost constant up to 50% sample imbalance in class 2 data. By proper selection of the input weights and bias value, a better performance can be attained. If careful observation is not taken then the classification performance of RGSO Based ELM classifier falls drastically with sample imbalance.



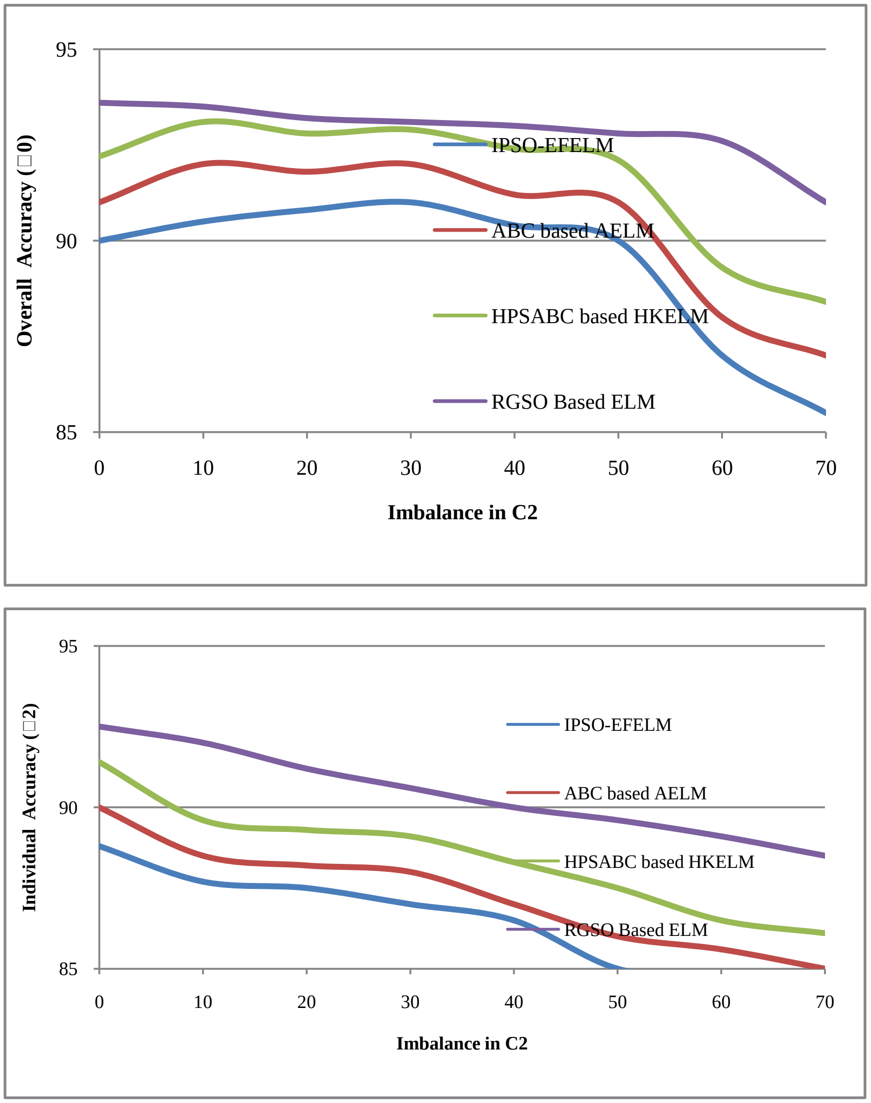


Fig. 2: Properties of the Imbalances in Data are depicted here; also the Performance of the RGSO Based ELM Classifier was Analyzed for Different Imbalance Conditions

D. Improved Integer-Coded Genetic Algorithm (AICGA)

Genetic algorithms, which are based on evolutionary search techniques [24], were developed in an attempt to explain the adaptive processes of natural systems and to design artificial systems based upon these natural systems. Genetic algorithms are widely used to solve complex optimization problems, where the number of parameters and constraints are large and analytical solutions are difficult to obtain. In recent years, many schemes for combining genetic algorithms and neural networks have been proposed and tested for feature selection. The complete survey on evolving neural networks using genetic algorithms can be found in [24]. The components of the GA consist of String Representation, Selection Function, Genetic Operators, and the Fitness Function. Detailed information on selection function (ranking method) and genetic operators (hybrid crossover and mutation) are described.

This paper presents an improved integer coded genetic algorithm (AICGA) to select the genes from the database. AICGA technique reduces the size of chromosomes and computation time significantly. Here, the proposed AICGA has been enhanced by using perturbation operator.

a. Chromosome Definition

Each chromosome consists of NG genes corresponding to NG units. The schedule for each unit can be demonstrated by a 5 digit string so that each digit shows the period of time that the unit remains in up or down state. Positive/negative numbers indicate up/down state. Due to the limitations of plants in startup and shutdown it seems to be rational to restrict 5 transitions for each unit during 24 hours of a day. Figure 1 shows the concept clearly and Equation (13) describes it.

$$\sum_{c=1}^{5} T_i^c = 24 \qquad i \in \{Gen.units\}$$

b. Perturbation Operator

Perturbation operator is a special case of proposed mutation. While in mutation, any selected digit can be replaced by any other acceptable digit; in perturbation, randomly selected digit will be added with 1. It means we decide to increase or decrease the previous on or off time. To meet (13) one another digit (e.g., adjacent integers) should be changed. This operator is applied to the best chromosomes with a suitable rate.

c. String Representation

In this paper, an AICGA is used for selecting the N best independent features from the given set. The characteristic string, which represents N independent features, is given as

$$S = [F_1, F_i, F_j, L, F_N]$$

Where the selected features belong to the set S and they are independent.

d. Fitness

The main aim of feature selection is to determine the features (search nodes) that best illustrate the input output characteristics of the data. The results of the RGSO Based ELM fivefold cross-validation test are used as fitness criterion, i.e., for the selected features, RGSO will identify the best hidden neurons, input weights, and biases values, and return the validation efficiency obtained by the proposed algorithm along with the best ELM parameters. The features returning the best validation efficiency eventually are chosen as representative of the full data set:

$$F_i = \eta^+ \qquad (10)$$

The best solution (for the selected set of genes and ELM parameters) obtained after a given number of generations is used to develop a classifier using the complete training set. This classifier is then used to classify the testing samples.

III. EXPERIMENTAL RESULTS

In this section, the performance of the proposed approach is compared with other methods based on Global Circulation Models (GCM) data set [25], in two steps. Initially, with the GCM data set the preprocessing process is carried out to find the missing values to change those values into feasible values. Then the classification process is carried out and the results are compared with other classifiers therefore the results for gene selection are compared with other existing results for gene selection. The samples in each class are tiny with high sample imbalance in GCM data set, that is, large number of classes with high dimensionality requires attention for selection of samples to training and testing. In these experiments, the data set is dividing into training and testing data.

A. Global Cancer Map Data

The GCM data is the collection of six different medical institutions around 14 different types of malevolent tumors. It consists of 190 primary complete tumor samples and 8 samples are not used here called metastasis. Each sample contains the virtual expression of 16,063 genes (take for granted a one-to-one mapping from gene to probe set ID). From 190 samples, 144 samples are utilized for gene selection and classifier growth and the left behind 46 samples are used for assessment of the generalization performance. The amount of training samples per class varies from 8 to 24 which are sparse and imbalanced. Based on these notes, the GCM data set is sparse in environment with a high sample imbalance and a high-dimensional feature space for huge number of genes. The main objective is to select sets of genes from the 16,063-dimensional space and recognize the smallest number of genes needed to concurrently categorize every tumor types with greater accuracy.

B. Results on Preprocessing Process

In preprocessing process, it helps to change the missing values with various feasible values for further processing.

Table 1: Missing values Results

Datasets	KNN (%)	Enhanced KNN(%)	Proposed Mean Substitution (%)
GCM Dataset	86	93	95

Table 1 shows the results comparison of the pre processing step. It is clearly observed from the table that the proposed Mean Substitution approach provides 95 % better results when compared with the KNN and enhance KNN approach. The whole GCM dataset are taken for consideration, the proposed Mean Substitution approach provides better results. Thus, the proposed Mean Substitution approach outperforms the existing KNN and Enhanced KNN approach.

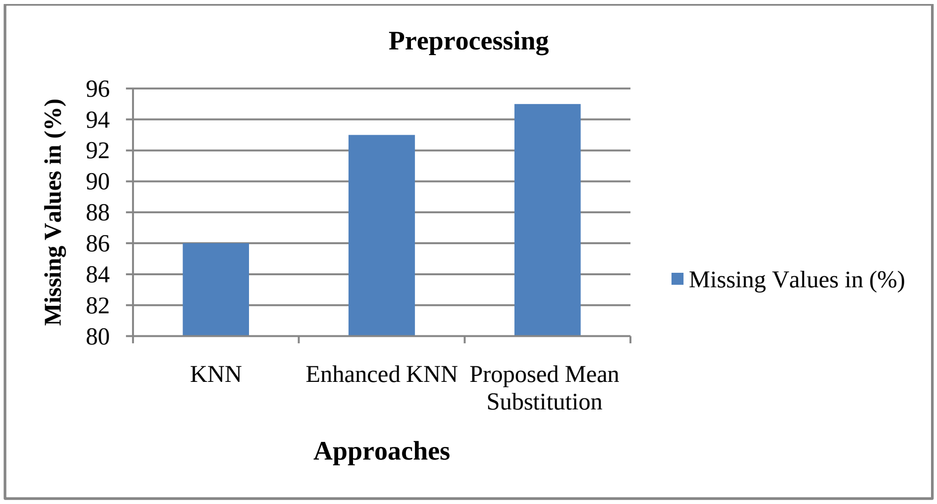


Fig. 3: Results on Preprocessing

In turn to calculate the classifier performance for sparse and imbalance data set, the results obtained by the proposed RGSO Based ELM classifier for a given number of genes is compared them with the existing classifiers. Here, 98 genes as selected in [5] as the source for the classifier performance comparison. The RGSO Based ELM classifier is ruined to recognize the paramount number of hidden neurons, input weights, and bias by means of 144 training data. With the use of best ELM parameters, an ELM classifier is developed by means of the complete training data and the resultant classifier is tested on the remaining 46 samples. This study is experimented for a variety of random combinations of 144 training and 46 testing samples set, and the results are account in Table 2.

Table 2: Comparative Analysis on Classification Methods for GCM Data Set Using 98 Genes Selected

Various Methods	ns	Training		Testing	
		Mean	Std_Deviation	Mean	Std_Deviation
SVM [26]	106	96.50	1.85	73.78	5.10
ELM[27]	50	92.30	2.25	79.43	6.23
PSO_ELM [15]	36	94.91	1.42	85.13	4.88
IPSO_E-FELM	30	93.14	1.23	88.45	3.94
ABC based AELM	26	92.85	1.10	89.74	3.24
HPSABC based HKELM	21	90.12	1.01	91.35	3.02
Proposed RGSO Based ELM	18	89.4	1.00	93.55	2.84

From the table 2, examine that the RGSO Based ELM classifier gives better performance than the existing IPSO_E-FELM classifier, ABC based AELM and HPSABC based HKELM for 98 genes selected in [5].

C. HPSABC based HKELM with ICGA Based Gene Selection and Classification Results

The proposed approach is called to select 14, 28, 42, 56, 70, 84, and 98 genes from the original 16,063 genes using a 10-fold cross-validation method on the 144 training samples. The unexploited testing set (46samples) is worn to assess the generalization performance. RGSO Based ELM with AICGA is identified best genes for each set. In this experiments, create that the best genes are chosen throughout different runs do not share any common genes. The overlap between the best genes sets (14-98) chosen by proposed approach is insignificant, but their ability to differentiate the cancer classes is more or less similar. These results show that there be real subsets of genes that can discriminate or differentiate the cancer classes efficiently

Table 3: Performance of Proposed Classifier for the Best Set of Features Selected by RGSO Based ELM with AICGA Gene Selection Approach

Genes	Training Efficiency %			Testing Efficiency		
	Avg	Max	Std_dev	Avg	Max	Std_dev
14	94	98	2	74	82	6
28	94	96	2	72	86	6
42	92	95	1	75	98	4
56	92	95	1	88	97	3
70	95	98	2	90	97	4
84	95	98	2	93	97	4
98	94	98	2	94	99	4

The performance of the proposed classifier by creating 100 random trials on the training and testing data sets is done by the best gene sets selected as above. It helps us to predict the classifier sensitivity to data variation. The average, maximum, and standard deviations of training and testing performances are given in Table 3 and the selected genes are listed in Table 4.

Table 4: Genes Selected from GCM Data Set That Were Used for Classification by RGSO Based ELM with ICGA

GCM 42 Genes							
Gene	Accession ID	Gene #	Accession ID	Gene #	Accession ID	Gene #	Accession ID
572	D79987_at	1882	M27891_at	7870	AA232836_at	13781	RC_AA403162_at
5836	HG3342-HT3519_s_at	6868	M68519_rnal_at	8034	AA278243_at	13964	RC_AA416963_at
917	HG3432-HT3618 _at	6765	M96132_at	8107	AA287840_at	14565	RC_AA446943_at
5882	HG417-HT417_s_at	3467	U59752_at	8231	AA320369_s_at	14793	RC_AA453437_at
1119	J04611_at	3804	U80017_rna2	8975	AB002337_at	11421	X05978_at
1137	J05068_at	6154	V00565_s_at	9546	H44262_at	476	D50678_at
9731	L13738_at-2	11443	X52056_at-2	9833	M21121_s_at		
1383	L20320_at	4629	X79510_at	10322	R74226_at		
9781	L40904_at	4781	X90872_at	12020	RC_AA053660_at		
5319	L46353_at	4944	Y00815_at	12182	RC_AA100719_s_at		
1655	L77563_at	11606	Z30425_at-2	12717	RC_AA233126_at		
1791	M20530_at	7284	AA036900_at	13541	RC_AA347973_at		

D. Performance Comparison of Proposed RGSO Based ELM with AICGA Classifier with Existing Methods

The proposed approach for the GCM data set results is compared with other existing methods. Table 4 shows the minimum number of genes needed by each method to attain maximum generalization performance. From the table 4, the proposed RGSO Based ELM with AICGA selects a minimum 42 genes with a high average testing accuracy. GA/SVM, selects a minimum of 26 genes which gives results close to RGSO Based ELM with ICGA performance. It was seen that genes chosen in a variety of runs for any given subset do not have major overlaps also there is no any overlap of genes between any two subsets. Until now, the classifiers improved by means of these sets of selected genes make similar classification performance and were experiential to have the same discriminatory power to classify various cancer classes.

The RGSO Based ELM with AICGA gene selection and classifier was used to select the minimum number of genes necessary for accurate classification. The average classification accuracies are given in Table 5 and 6.

Table 5: Minimum Number of Genes Required by Various Methods to Achieve Maximum Generalization Performance

Data Set	Gene selection method	Genes	Avg. Testing Accuracy %
GCM	Proposed RGSO Based ELM	42	95
		98	97.2
	HPSABC based HKELM with ICGA	42	93.6
		98	96.12
	ABC based AELM with ICGA	42	92
		98	95
	ICGA_IPSO_E-FELM	42	90
		98	94
	ICGA_PSO_ELM	42	88
		98	91
	GA/SVM	26	85

Table 6: Results for Gene Selection and Classification by RGSO Based ELM with AICGA for Different Data Sets

Data set	#Classes	#Genes	Testing Accuracy %	
			Average	Best
Lymphoma	6	12	100	100
CNS	2	12	100	100
Breast Cancer-B	4	12	95	100

IV. CONCLUSION

In this paper, initially preprocessing process is carried out using a Mean substitution and normalization approach is proposed to find missing values of datasets and the scaled datasets. Then an accurate gene selection and sparse data classification for microarray data is done using HPSABC based RGSO based ELM gene selection for multiclass cancer classification is proposed. Advanced ICGA selected genes included with optimal input weights and bias values selected by RGSO and used by the ELM classifier, to deal with higher sample imbalance and sparse data conditions resourcefully. Hence, AICGA gene selection approach is incorporated with the RGSO based ELM classifier to identify a dense set of genes that can discriminate cancer types efficiently resulting in enhanced classification results. Thus the experimental result shows that the proposed approach provides better result when compared with other approaches. The application is to develop this algorithms based on these computing techniques for diagnostic science applications and hence provide a better framework for development of emerging medical systems, enabling the better delivery of healthcare.

REFERENCES

[1] S. Peng, Q. Xu, X.B. Ling, X. Peng, W. Dua, and L. Chen, "Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithms and Support Vector Machine," FEBS Letters, vol. 555, no. 2, pp. 358-362, 2003.

[2] Y. Saeys, I. Inza, and P. Larran˜ aga, "A Review of Feature Selection Techniques in Bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, Oct. 2007.

[3] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of Samples Are Needed to Generate a Robust Gene List for Predicting Outcome in Cancer," Proc. Nat'l Academy of Sciences USA, vol. 103, no. 15, pp. 5923-5928, Apr. 2006.

[4] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result Analysis of the NIPS 2004 Feature Selection Challenge," Proc. Conf. Advances in Neural Information Processing Systems (NIPS), vol. 17, pp. 545-552, 2004.

[5] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub, "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures," Proc. Nat'l Academy of Sciences USA, vol. 98, no. 26, pp. 15149-15154, Dec. 2001.

[6] X. Zhou and D. Tuck, "MSVM-RFE: Extensions of SVM-RFE for Multiclass Gene Selection on DNA Microarray Data," Bioinformatics, vol. 23, no. 9, pp. 1106-1114, 2007.

[7] C.H. Ooi and P. Tan, "Genetic Algorithms Applied to Multi-Class Prediction for the Analysis of Gene Expression Data," Bioinformatics,vol. 19, no. 1, pp. 37-44, Jan. 2003.

[8] N. Yukinawa, S. Oba, K. Kato, and S. Ishii, "Optimal Aggregationof Binary Classifiers for Multiclass Cancer Diagnosis Using Gene Expression Profiles," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 6, no. 2, pp. 333-343, Apr.-June 2009.

[9] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis," Bioinformatics, vol. 21, no. 5, pp. 631-643, 2005.

[10] S. Golestani, M. Raoofat and E. Farjah, "An Improved Integer Coded Genetic Algorithm for

Security Constrained Unit Commitment", The Pacific Journal of Science and Technology, Volume 12. Number 1. May 2011.

[11] K. Kothavari, B. Arunadevi, and S. N. Deepa, "A Hybrid DE-RGSO-ELM for Brain Tumor Tissue Categorization in 3D Magnetic Resonance Images", Hindawi Publishing Corporation, 2014.

[12] Enrique Alba, Jos ´ eG arc ´ õa-Nieto, Laetitia Jourdan, El-Ghazali Talbi, "Gene Selection in Cancer Classification using PSO/SVM andGA/SVM Hybrid Algorithms ",IEEE Congress on Evolutionary Computation - CEC , pp. 284-290, 2007

[13] Ronny Luss, Alexandre d'Aspremont, "Clustering and Feature Selection using Sparse Principal Component Analysis", July 4, 2007.

[14] Yanwei Huang, Liqing Zhang, "Gene Selection for Classifications Using Multiple PCA with Sparsity", TSINGHUA SCIENCE AND TECHNOLOGY, ISSNll1007-0214ll06/10llpp659-665, Volume 17, Number 6, December 2012.

[15] Saras Saraswathi, Suresh Sundaram, Narasimhan Sundararajan, Michael Zimmermann, and Marit Nilsen-Hamilton, "ICGA-PSO-ELM Approach for Accurate Multiclass Cancer Classification Resulting in Reduced Gene Sets in Which Genes Encoding Secreted Proteins Are Highly Represented", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 8, NO. 2, MARCH/APRIL 2011

[16] Brian D. Ripley (1996), Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge.

[17] R.S. Somasundaram and R. Nedunchezhian, "Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications Issn-09758887, 2011.

[18] Angeline Christobel. Y, P. Sivaprakasam "Improving the Performance of K-Nearest Neighbor Algorithm for the Classification of Diabetes Dataset with missing values", International Journal of Computer Engineering and Technology" (IJCET), Volume 3, Issue 3, October -December (2012), pp. 155-167

[19] S. N. Sivanandam, S. Sumathi, and S. N. Deepa, Introduction to Neural Networks Using Matlab 6.0, Tata McGraw Hill, New Delhi, India, 2006.

[20] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," Neurocomputing, vol. 70, no. 1–3, pp. 489–501, 2006.

[21] G. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," International Journal of Machine Learning and Cybernetics, vol. 2, no. 2, pp. 107–122, 2011.

[22] S.He, Q. H. Wu, and J. R. Saunders, "Anovel group search optimizer inspired by animal Behavioral ecology," in Proceedings of the IEEE Congress on Evolutionary Computation (CEC '06), pp. 1272–1278, Sheraton Vancouver Wall Center, Vancouver, Canada, July 2006.

[23] F. Han, H.-F. Yao, and Q.-H. Ling, "An improved Extreme learning machine based on particle swarm optimization," in Proceedings of the International Conference on Intelligent Computing, pp. 699–704, 2012.

[24] Y. Zhang and L. Wu, "An MR brain images classifier via principal component analysis and kernel support vector machine," Progress in Electromagnetics Research, vol. 130, pp. 369–388, 2012.

[25] Z. Michalewicz, Genetic Algorithm + Data Structures = Evolution Programs, third ed., pp. 18-22. Springer-Verlag, 1994.

[26] S. Suresh, N. Sundarajan, and P. Saratchandran, "A Sequential Multi-Category Classifier Using Radial Basis Function Networks," Neurocomputing, vol. 71, nos. 7-9, pp. 1345-1358, 2008.

[27] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory Classification Using an Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis," IEEE/ ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 3, pp. 485-495, July-Sept. 2007.