

# **An Efficient Hybrid Kernel Extreme Learning Machine (HKELM) With Hybrid Fuzzy Glow Worm Swarm Optimization (HFGSO) For Micro Array Gene Expression And Multi Class Cancer Classification**

**R.Balakrishnan<sup>1</sup> and Thirunavu Karthikeyan<sup>2</sup>**

<sup>1</sup>*Assistant Professor, Dr.NGP Arts and Science College, Coimbatore,  
Email: balakrishnan.scholar@yahoo.com*

<sup>2</sup>*Associate Professor, P.S.G. College of Arts and Science, Coimbatore.  
Email: t.karthikeyan.gasc@gmail.com*

## **Abstract**

Microarrays can be employed to find the relative amount of mRNAs in more tissue samples for thousands number of genes at the same time. As the superiority of this method has been find out various open queries arise about appropriate assessment of microarray data. The multiclass cancer classification is playing an important role in the field of medical sciences. Because the numbers of cancer victims are growing steadily, the need of the cancer classification techniques has become indispensable. 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 Hybrid Kernel Extreme Learning Machine (HKELM), with Hybrid Fuzzy based Glow worm Swarm Optimization (HFGSO) technique is used for optimal gene selection and cancer classification. AICGA is employed with HFGSO Based HKELM classifier to select an optimal set of genes which results in increased performance of an efficient hybrid algorithm that can handle sparse data and sample imbalance. The HFGSO with HKELM is used to carry out the classification process. In the proposed HFGSO based HKELM, the weights and bias to HKELM are optimized using HFGSO for better simplification and classification of large value of gene datasets. The proposed approaches are applied for real time datasets and benchmark datasets taken from dataset repositories.

**Keywords----** Hybrid Kernel Extreme Learning Machine (HKELM), Integer-Coded Genetic Algorithm, Gene selection, classification, Hybrid Fuzzy Glow worm Swarm Optimization (HFGSO)

## **1. Introduction**

Diagnosis of Cancer detection and classification are based on pathological analysis of tissue sections, in accordance with the subjective analysis of data [1]. Only a small amount of information is attained from morphological analysis is frequently not enough in cancer diagnosis and may result in expensive but unsuccessful treatment of cancer. To precisely identify cancer subtypes, recent research have been carried out to identify genes that may cause cancer [2].

Researchers said that, several issues are there in microarray data cancer classification such as robustness of gene selection and gene ranking [3]. Taxonomy, probable use, and variety of feature selection techniques are presented in [2]. In [3] the author states that thousands of samples are needed for healthy gene selection, in order to have overlapping sets of genes. This method of gene selection and classification shows good classification results with a small set of samples. It can be examined that different sets of genes, which have few genes in general, can categorize a wide variety of cancer types with high accuracy. This is made 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.

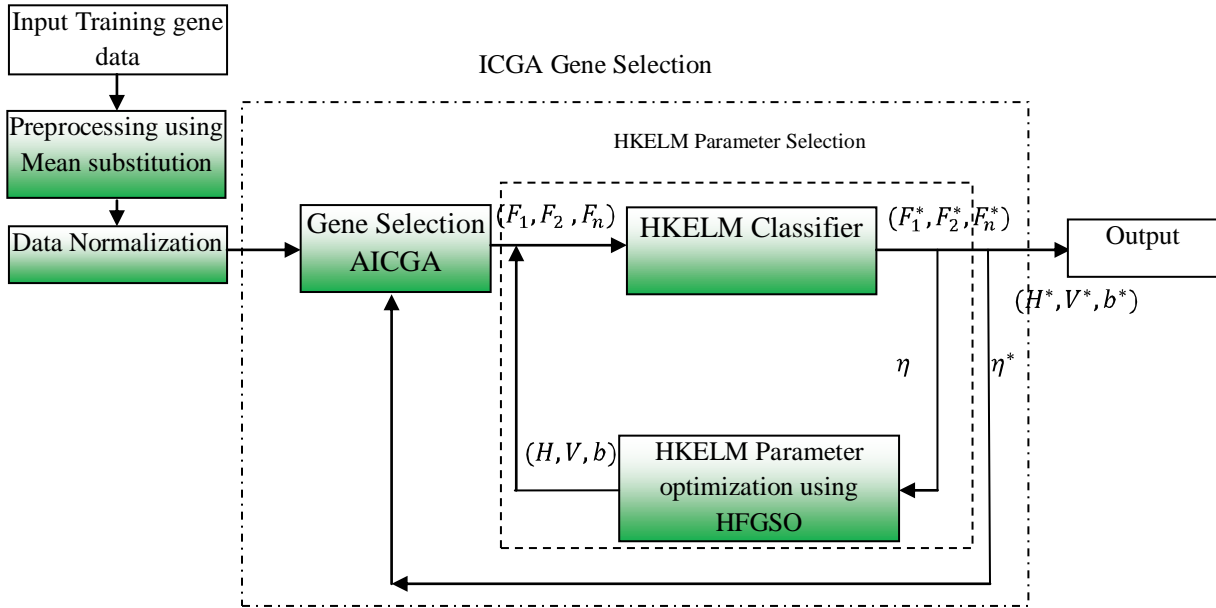
Conventional gene selection methods are of two types, that is the filtering approach and the wrapper approach. In the filtering approach, selected genes are self-governing of the choice of classification methods, where in the wrapper approach; gene selection is influential on the choice of the classifier. 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 [10] 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 [11] 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 [12] 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. This approach results in which the ELM classifier used in this approach can distinguish the cancer classes among the data denoting the chosen features quickly, but the performance of ELM classifier is based on the nature of the input data distribution. For sparse and highly imbalanced data set, the random input weight selection in ELM classifier affects the classification performance to a great extent. Also, the PSO algorithm has slow convergence in refined search stage and has weak local search ability.

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 (AICGA) [13] is used for strong and healthy gene selection. Next, proposes a Hybrid Kernel Extreme Learning Machine (HKELM) [14], with Hybrid Fuzzy based Glow worm Swarm Optimization (HFGSO) [15] technique for managing the sparse/imbalanced data classification problem.

## **2. Proposed Methodology**

The proposed methodology of the block diagram is shown in the figure 1. 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 HKELM 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 Hybrid Kernel Extreme Learning Machine (HKELM) with Hybrid Fuzzy based Glow worm Swarm Optimization (HFGSO) technique is proposed, to select the best parameters for better simplification and training of the classifier for gene data. Based on these chosen genes, HKELM algorithm generates significant classifier by calculating weights of the genes. Initially, AICGA selects  $n$  independent genes from the available gene set. In the proposed HFGSO based HKELM, the weights and bias to HKELM are optimized using HFGSO for better simplification and classification. For the selected genes, HFGSO will identify optimal parameters like number of hidden nodes and input weights such that the performance of the HKELM multiclass classifier is improved. The best validation performance ( $\eta^+$ ) will be utilized as fitness for the AICGA evolution. The validation performance of HKELM classifier ( $\eta$ ) is used in HFGSO for selection and grouping of HKELM parameters.



**Figure 1: Proposed Flow Diagram**

### 2.1. 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

$$D = \{ A_1, A_2, A_3, \dots, A_n \}$$

Where

D is the set of data with missing values

$A_i$  – is the  $i$ th 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 {

$a_i \leftarrow A_i \cap m_i$

where

$a_i$  is the column of attributes without missing values

$m_i$  is the set of missing values in  $A_i$  (missing values denoted by a symbol)

Let  $\bar{a}_i$  be the mean of  $a_i$

Replace all the missing elements of  $A_i$  with  $\bar{a}_i$

}

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 are scaled to significantly low values [18].

The Data Scaling Algorithm

Let

$D = \{ A_1, A_2, A_3, \dots, A_n \}$

Where

$D$  is the set of unnormalized data

$A_i$  – is the  $i$ th 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  $n$  {

$Max_i \leftarrow \max(A_i)$

$Mini \leftarrow \min(A_i)$

    For  $r=1$  to  $m$  {

$A_{ir} \leftarrow A_{ir} - Mini$

$A_{ir} \leftarrow A_{ir} / Max_i$

    Where

$A_{ir}$  is the element of  $A_i$  at row  $r$

    }

    }

Finally will have the scaled data set

End

### **2.2. Improved Integer-Coded Genetic Algorithm (AICGA) for gene selection**

Genetic algorithms, which are related to evolutionary search algorithms [19], were presented to give details about the adaptive processes of natural systems and helps to formulate artificial systems based upon these natural systems. Genetic algorithms are 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 [20]. 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 Advanced 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.

#### **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 results. 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 HGSO Based HKELM fivefold cross-validation test are used as fitness criterion, i.e., for the selected features, HGSO 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 HKELM parameters. The features returning the best validation efficiency eventually are chosen as representative of the full data set:

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

The best solution (for the selected set of genes and HKELM 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.

### 2.3. HFGSO Based HKELM for Gene Classification

After selecting the best genes, the selected genes are classified using Hybrid Glowworm Swarm Optimization (HGSO) based Hybrid Kernel Extreme Learning Machine (HKELM).

#### A. Formation of Hybrid Fuzzy based Glowworm Swarm Optimization (HFGSO)

In this section, the algorithms like glowworm swarm optimization and the fuzzy logic approach used for test case optimization is explained here briefly.

##### i. *Glowworm swarm optimization (GSO) algorithm*

Krishnanand proposed a Glowworm swarm optimization (GSO) algorithm in 2005 was explained in [21]. In this algorithm, glowworms light emission property was implemented which offered them with peer or prey attraction ability. Production of light of glowworms is done by a chemical named Luciferin.

Glowworm algorithm contains a greatest ability in solving problems, such as finding some local optima of multi modal functions simultaneously [22] searching spaces of higher dimensions and orienting multiple sources. In GSO, each glowworm distributes the objective function in definite space [23]. These glowworms carry their own luciferin values and have the relevant field of idea called local-decision range. Their brightness concerns with in the position of objective function value. Brighter the glow, the best is the position that is to say has the good target value. As the glow seeks for the neighbor set in the local-decision range, in the set, a brighter glow has a higher attraction to attract this glow toward this traverse, and the flight direction each time different will change along with the choice neighbor. Furthermore, the local-decision range size will be influenced by the neighbor quantity, when the neighbor density will be low, glow's policy-making radius will enlarge favors seeks for more neighbors, otherwise, the policy-making radius reduces. Finally, the majority of glowworm return gathers at the multiple optima of the given objective function.

Each glowworm  $i$  encodes the object function value  $J(x_i(t))$  at its current location  $x_i(t)$  into a luciferin value  $l_i$  and broadcasts the same inside its neighbourhood. The set of neighbours ( $N_i(t)$ ) of glowworm  $i$  comprises of those glowworms that have a relatively higher luciferin value and that are located within a dynamic decision domain and updating by formula (1) at each iteration.

Local-decision range update:

$$r_d^i(t+1) = \min \left\{ r_s, \max \{ 0, r_d^i(t) + \beta(n_t - |N_i(t)|) \} \right\}; \quad (1)$$

and  $r_d^i(t+1)$  is the glowworm  $i$ 's local-decision range at the  $t+1$  iteration,  $r_s$  is the sensor range,  $n_t$  is the neighbourhood threshold,  $\beta$  which affects the rate of change of the neighbourhood range. The number of glow in local-decision range:

$$N_i(t) = \{j: \|x_j(t) - x_i(t)\| < r_d^i; l_j(t) < l_i(t)\}; \quad (2)$$

and,  $x_j(t)$  is the glowworm  $i$ 's position at the  $t$  iteration,  $l_i(t)$  is the glowworm  $i$ 's luciferin at the  $t$  iteration.; the set of neighbours of glowworm  $i$  consists of those glowworms that have a relatively higher luciferin value and that are located within a dynamic decision domain whose range  $r_d^i$  is bounded above by a circular sensor range  $r_s$  ( $0 < r_d^i < r_s$ ). Each glowworm  $i$  selects a neighbour  $j$  with a probability  $p_{ij}(t)$  and process toward it. These movements that are based only on local information, enable the glowworms to partition into disjoint subgroups, reveal an instantaneous taxis-behaviour in the direction and eventually co-locate at the multiple optima of the given objective function. Probability distribution used to choose a neighbour:

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)} \quad (3)$$

Movement update:

$$x_i(t+1) = x_i(t) + s \left( \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|} \right); \quad (4)$$

Luciferin-update:

$$l_i(t) = (1 - \rho)l_i(t-1) + \gamma J(x_i(t)); \quad (5)$$

and  $l_i(t)$  is a luciferin value of glowworm  $i$  at the  $t$  iteration,  $\rho \in (0,1)$  lead to the reflection of the cumulative kindness of the path followed by the glowworms in their current Luciferin values, the parameter  $\gamma$  only scales the function fitness values,  $J(x_i(t))$  is the value of test function.

## ii. *Fuzzy Logic*

Fuzzy logic building is in accordance with fuzzy sets theory. This theory is a general idea of conventional sets theory in mathematics. In conventional sets theory, a component fit into the set or doesn't. In to the fact that, membership of elements is consequent to a two value pattern. Other than that, fuzzy sets theory widens this model and applies membership degree. As a result, an element can be a member of the set not fully too some extent. Fuzzy set is defined as:

$$\{(x, \mu_A(x)) \mid x \in X\} \quad (6)$$

where, the membership of set's members is identified by membership function  $\mu(x)$  that  $x$  represents a specific element and  $\mu$  is a fuzzy function which decide the membership degree of  $x$  in equivalent set and it takes a value between zero and one.

In addition, it can be represented as  $(x)$  is a map from values  $x$  to possible numbers between zero and one.  $(x)$  may be a set of discrete or continuous values. The

properties expressed to find out fuzzy set's members are fuzzy and are not accurate. Afterward, it is feasible to use different membership functions to exemplify a fuzzy set. Approximately, those functions are employed which have a simple mathematical representation and are flexible by a partial number of parameters. Though, membership functions are divided into point, linear and nonlinear functions. General form of linear one is motivated from polygonal shapes, like trapezoidal membership function and general form of nonlinear case is owing to bell shapes, like Gaussian membership function. Fuzzy logic is applied widely in solving a variety of problems. Many researchers make use of fuzzy logic to develop optimization algorithms efficiency [24].

### **B. Hybrid Kernel Extreme Learning Machine (HKELM)**

During recent years in medical analysis, artificial neural networks [25] 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 [25] 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 Hybrid Kernel Extreme Learning Machine (HKELM). The proposed Hybrid Kernel Extreme Learning Machine (HKELM) classifier which holds the training for particular hidden layer feed forward neural networks (FFNN).

### **iii. Kernel Function**

Here the low dimensional linear space inseparable mode by nonlinear mapping to high-dimensional feature space can attain linear separable. However, suppose this technology is used in high-dimensional space for classification or regression, then there will be a problem on formative the form and parameters of the nonlinear mapping functions and feature space dimension. In the high dimensional feature space operation the major difficulty is the survival of "dimensions of disaster". The kernel function approach can used to solve these problems effectively.

In 1964, Aizermann presented a kernel function theory to the field of machine learning and it is developed for the function of neural network algorithm representation. Kernel function method uses an arbitrary random vector  $X$  in the  $n$  dimensional vector space mapped to a high dimensional feature space  $F : x \rightarrow \Phi(x) \in F$  with a nonlinear transformation and can attain a high dimensional feature space linear classification. In high dimensional feature space  $F$ , the interaction

between each coordinate element is inadequate to the inner product linear learning algorithm, does not need specific forms of nonlinear transformation, providing the kernel function to replace the inner product in the linear algorithm persuade the Mercer condition can obtain the original input space equivalent to nonlinear algorithm.

Generally most of the kernel function is mainly belongs to 3 main categories, they are:

- 1) Polynomial kernel function

$$K(x, x_i) = (\gamma(x \cdot x_i) + r)^d, \gamma > 0 \quad (3)$$

- 2) Perceptron kernel function

$$K(x, x_i) = \tan h (v(x \cdot x_i) + c) \quad (4)$$

- 3) Gauss RBF nuclear function

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \quad (5)$$

- 4) Sigmoid function

$$K(x) = \frac{1}{1 + \exp(-x)} \quad (6)$$

In the above formula (4) to (6), d,c,σ are real constant parameters.

#### **iv. Hybrid Kernel Function**

There are several types of kernel function summed up with each other, which can be separated into two major types of local kernel function and global nuclear function. For instance, RBF function is a general local kernel function, the Perceptron kernel function and polynomial kernel function are two characteristic global nuclear function. In test point the local kernel functions is closer to the region class has an impact on the data points and global nuclear function permits the kernels away from the test input data point values also influences were studied in [25]. Since the local kernel has high learning ability but the performance is poor, by this, the global nuclear function has finest generalization performance however the training ability is not good. Therefore, these two types of nuclear function combined to form a hybrid kernel function.

The traditional approach of building kernel functions is based on the thought of Mercer's theorem which was proposed in [25].

Simple kernel function build complex hybrid kernel function, that is the hybrid kernel function still satisfies the Mercer theorem of the kernel function.

Set  $K_1, K_2$  is defined in the kernel function on  $X \times X$ ,  $f$  is real-valued functions on  $X: \Phi(x): X \rightarrow \mathbb{R}^N, K_3$  is a kernel function on  $\mathbb{R}^N \times \mathbb{R}^N, a \in \mathbb{R}^+, B$  is a  $n \times n$  dimensional positive semi definite symmetric matrix, followed by Mercer theorem, the functions as follows are the kernel function:

1.  $K(x, z) = K_1(x, z) + K_2(x, z)$
2.  $K(x, z) = f(x)f(z)$
3.  $K(x, z) = aK_1(x, z)$
4.  $K(x, z) = K_1(x, z)K_2(x, z)$
5.  $K(x, z) = K_3(\Phi(x), \Phi(z))$
6.  $K(x, z) = x^T B_z$

**Proof:** Set matrix  $K_1, K_2$  which is defined on a finite set of points  $\{x_1, \dots, x_n\}$ , for any vector  $\alpha \in \mathbb{R}^n$ ,  $K$  is a positive semidefinite matrix and the compulsory and satisfactory condition is all  $\alpha$  should satisfy  $\alpha' K \alpha \geq 0$ . It can be said that  $\alpha' (K_1 + K_2) \alpha = \alpha' K_1 \alpha + \alpha' K_2 \alpha \geq 0$ , then  $K_1 + K_2$  is positive semidefinite, that is to say  $K_1 + K_2$  meet Mercer theorem, so it is kernel function,  $K(x, z) = K_1(x, z) + K_2(x, z)$  is kernel function.

By the function  $K = \lambda K_1 + (1 - \lambda)K_2, \lambda \in (0,1)$  is a mixed function which satisfies the condition of Mercer algorithm. In this paper, the RBF kernel function  $K(x, x_i) = \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)$  is combined with the Perceptron kernel function  $K(x, x_i) = \tan h(v(x, x_i) + c)$  to form a hybrid kernel function. With the help of hybrid function, the learning ability of the kernel function and its generalization ability are improved. Then apply it to the ELM therefore achieve good classification performance.

Note that to ensure the proposed hybrid kernel does not varies the rationality in the original mapping space and then the proportion coefficients sum of two kernel functions is 1. Based on this, a hybrid kernel function is introduced.

$$K(x, x_i) = \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) + (1 - \lambda) \tan h(v(x, x_i) + c) \quad (7)$$

**v. ELM based on Hybrid Kernel Function (HKELM)**

In the paper, construct an extreme learning machine representation based on hybrid kernel function called HKELM. In HKELM, the hybrid kernel function combines well-built learning capability of local kernel function and generalization ability of global nuclear function, by means of improved learning performance. The steps are as follows

- Step 1: Describe network. Define the number of hidden layer nodes and arbitrarily allocate to input weights and indirect bias.
- Step 2: Construct a Model. Establish the hybrid kernel function to build ELM learning model.
- Step 3: Samples training to train the data.

Step 4: Performance testing of HKELM generalization performance and learning performance.

#### **2.4. Proposed HFGSO Based HKELM for gene Classification**

In this approach, Hybrid Fuzzy Based Glow worm Algorithm (HFGSO) is used for optimizing the weights in HKELM neural network. In standard GSO algorithm, glowworms have to be familiar with their neighbors before moving step. In this algorithm, a glowworm which needs to alter position toward other glowworms is considered as reference glowworm. Afterward, other glowworms which their Euclidean distance from this reference glowworm is less than reference glowworm's decision field are measured as reference glowworm's neighbors. This approach has some complexity. If a glowworm is just a little further than the effect of reference glowworm, it will not present in neighbours set. In addition, it is promising that no other glowworm is in a neighborhood of a glowworm and in moving step; the position of the glowworm with no neighbors doesn't change. It causes glowworms not to converge to a global optimum with an appropriate rate and to be far away from each other, and even in several cases, a few glowworms might not move until the end of algorithm iterations. According to this case, in order to take into account the effect of more glowworms on each glowworm, the degree of neighborhood of glowworms to each other is firm by a fuzzy membership function. So neighborhood degree is restored with former idea of crisp neighbors set.

This HFGSO combined with HKELM facilitates the selection of input weights to increase the simplification performance and the training of the single layer feed forward neural network. HFGSO based HKELM 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 HKELM classifier can differentiate the cancer classes amongst the data indicate the chosen features in fast manner. The proposed classifier in which the proposed HFGSO algorithm is employed to find the optimal input weights such that HKELM classifier can differentiate the cancer classes considerably, that is, the performance of the HKELM 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 HKELM.

In the proposed methodology each weight would symbolize a Luciferin value of the glowworm and the objective of this method is to find a best glowworms' Luciferin that refers to the test cases with maximum coverage.

The best glowworms' Luciferin value corresponds to a potential solution of the optimization problem and the nectar amount match to the fitness of the associated solution.

The steps of the proposed approach are as follows.

1. Initialize the genes position as a glowworm  
Set number of dimension =  $m$

Set number of glowworm = n  
 Initialize the test case which is to be performed by the glowworm  
 Let s be the step size  
 Let  $x_i(t)$  be the location of glowworm i at time t  
 deploy\_agent\_randomly,  
 For  $i=1$  to n do  $l_i(0) = l_0$   
 Set maximum iteration number=iter\_max  
 Set  $t=1$ ;  
 While ( $t \leq \text{iter\_max}$ ) do:  
 {  
 for each glowworm i do : % Luciferin-update phase  
 $l_i(t) = (1 - \rho)l_i(t - 1) + \gamma J(x_i(t))$ ,  
 for each glowworm i do : %Movement -phase  
 {  
 $N_i(t) = \{j: \|x_j(t) - x_i(t)\| < r_d^i(t); l_i(t)l_j(t)\}$ ,  
 Where  $\|\vec{x}\|$  is the norm of  $\vec{x}$   
 $\mu(\vec{x}) = \exp\left(-\left(\frac{x - m}{\sigma}\right)^2\right)$   
 For each glowworm  $j \in N_i(t)$  do:  

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)}$$
  
 $p_{ij} \rightarrow \mu_{ij} p_{ij}$   
 $j = \text{select\_glowworm}(\vec{p})$   

$$x_i(t + 1) = x_i(t) + s \left( \frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|} \right)$$
  

$$r_d^i(t)(t + 1) = \min \{r_s, \max \{0, r_d^i(t) + \beta n t - N_{it}\}$$
  
 }  
 $t \leftarrow t + 1$ ;  
 }

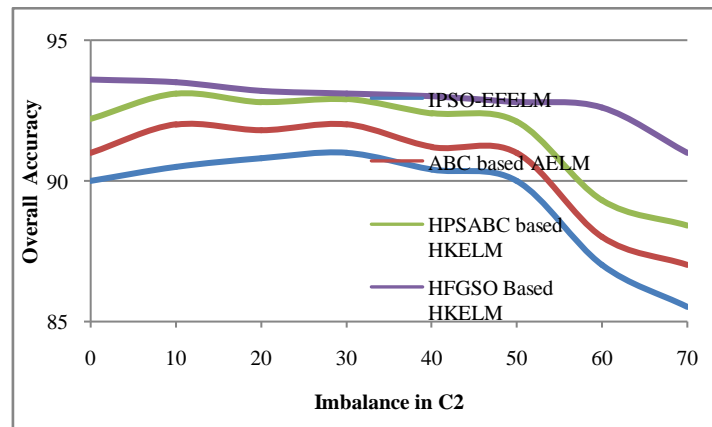
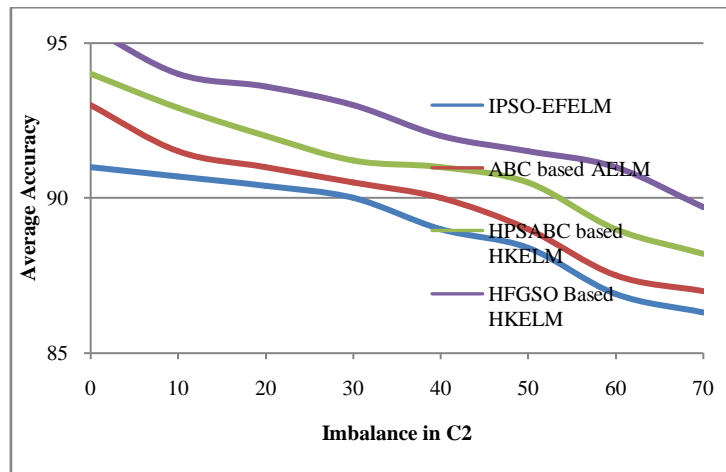
2. Ultimately, each glowworm is ensured for its objective values. If a gene is denoted by best glowworm, then optimization criterion is confirmed and it is stored. If not, other bees in the population are checked. The stopping criterion for this process is either run-on until to assign the optimal genes or until the end of predetermined loop counter.
3. For each member in the group, the particular output weights are calculated at HKELM.
4. At this instant invoke HFGSO based on AICGA
5. Then the objective function of each member, is measured
6. Find the producer of the group based on the fitness of the data.
7. Update the position of each member.
8. Stopping criteria

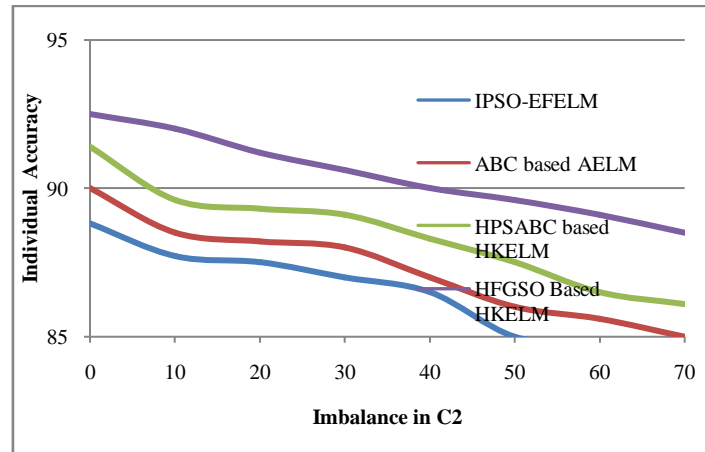
This increases local search property and using fuzzy neighborhood increases convergence rate algorithm speed generally. HFGSO iterates till its stopping criterion is met which determines the maximum number of paths covered and faults covered.

**2.5. Analysis on imbalance data**

The sample imbalance handling capacity of HFGSO Based HKELM 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 ( $\eta_a$ ), overall ( $\eta_o$ ) and individual ( $\eta_2$ ) 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 HFGSO Based HKELM classifier falls drastically with sample imbalance.





**Fig. 2: Properties of the imbalances in data are depicted here; also the performance of the HFGSO based HKELM classifier was analyzed for different imbalance conditions.**

### 3. Experimental Results

In this section, the performance of the proposed approach is compared with other methods based on Global Circulation Models (GCM) data set [12], 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.

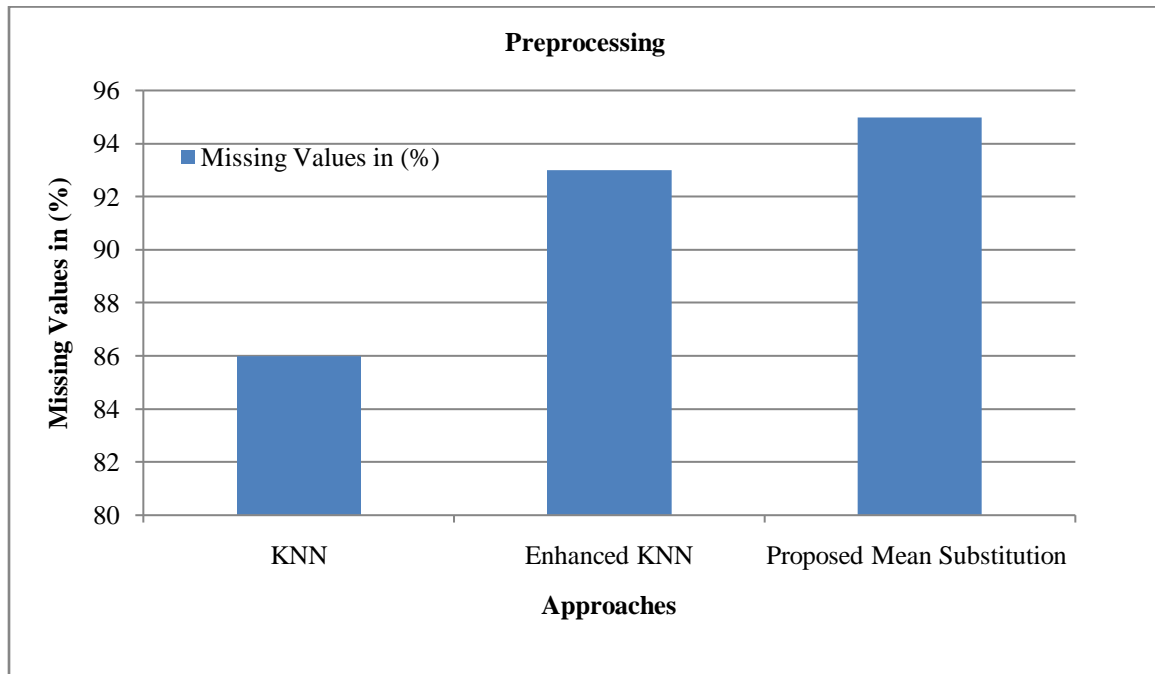
### B. Results on Preprocessing Process

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 and Figure 3 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.



**Figure 3: Results on Preprocessing**

In turn to calculate the classifier performance for sparse and imbalance data set, the results obtained by the proposed HFGSO Based HKELM 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 HFGSO Based HKELM 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

HKELM parameters, an HKELM 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 [12]	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 HFGSO Based HKELM	18	88.73	1.00	94.65	2.79

From the table 2, examine that the HFGSO Based HKELM classifier gives better performance than the existing IPSO\_E-FELM classifier, ABC based AELM and HPSABC based HKELM for 98 genes selected in [12].

### ***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. HFGSO 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 HFGSO 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 HFGSO 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 HFGSO based HKELM 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 HFGSO based HKELM 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 HFGSO based HKELM with AICGA 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 HFGSO Based HKELM 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 HFGSO based HKELM	42	95.5
		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 HFGSO 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

#### 4. 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 HFGSO based HKELM gene selection for multiclass cancer classification is proposed. Advanced ICGA selected genes included with optimal input weights and bias values selected by HFGSO and used by the HKELM classifier, to deal with higher sample imbalance and sparse data conditions resourcefully. Hence, AICGA gene selection approach is incorporated with the

HFGSO based HKELM 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..

## Reference

- [1] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C.H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov J.P, Poggio T, Gerald W, Loda M, Lander E.S and Golub T.R 2001, "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures", Proc. Nat'l Academy of Sciences USA, vol. 98, no. 26, pp. 15149-15154.
- [2] Peng S, Xu Q, Ling X.B, Peng X, Dua W and Chen L 2003, "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.
- [3] Saeys Y, Inza I and Larrañaga P 2007, "A Review of Feature Selection Techniques in Bioinformatics", Bioinformatics, vol. 23, no. 19, pp. 2507-2517.
- [4] Guyon, Gunn S, Ben-Hur A and Dror G 2004, "Result Analysis of the NIPS 2004 Feature Selection Challenge," Proc. Conf. Advances in Neural Information Processing Systems (NIPS), vol. 17, pp. 545-552.
- [5] Guyon I, Weston J, Barnhill S and Vapnik V 2002, "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, vol. 46, nos. 1-3, pp. 389-422.
- [6] Zhou X and Tuck D 2007, "MSVM-RFE: Extensions of SVM-RFE for Multiclass Gene Selection on DNA Microarray Data," Bioinformatics, vol. 23, no. 9, pp. 1106-1114.
- [7] Ooi CH and Tan P 2003, "Genetic Algorithms Applied to Multi-Class Prediction for the Analysis of Gene Expression Data," Bioinformatics, vol. 19, no. 1, pp. 37-44.
- [8] Yukinawa N, Oba S, Kato K and Ishii S 2009, "Optimal Aggregation of Binary Classifiers for Multiclass Cancer Diagnosis Using Gene Expression Profiles," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 6, no. 2, pp. 333-343.
- [9] Statnikov A, Aliferis C.F, Tsamardinos I, Hardin D and Levy S 2005, "A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis," Bioinformatics, vol. 21, no. 5, pp. 631-643.

- [10] Enrique Alba, José García-Nieto, Laetitia Jourdan, El-Ghazali Talbi 2007, “Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms”, IEEE Congress on Evolutionary Computation - CEC, pp. 284-290.
- [11] Ronny Luss, Alexandre d’Aspremont 2007, “Clustering and Feature Selection using Sparse Principal Component Analysis”.
- [12] Saras Saraswathi, Suresh Sundaram, Narasimhan Sundararajan, Michael Zimmermann, and Marit Nilsen-Hamilton 2011, “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.
- [13] Golestani S, Raofat M and Farjah E 2011, “An Improved Integer Coded Genetic Algorithm for Security Constrained Unit Commitment”, The Pacific Journal of Science and Technology, Vol 12. No 1.
- [14] Shifei Ding, Yanan Zhang, Xinzheng Xu, and Lina Bao, 2013. “A Novel Extreme Learning Machine Based on Hybrid Kernel Function”, Journal of Computers, 8(8).
- [15] Abraham Kiran Joseph and Dr. G. Radhamani, “Optimization of Test Cases using Fuzzy based Glowworm Swarm Algorithm”, Int. J. on Recent Trends in Engineering and Technology, Vol. 11, No. 1, July 2014.
- [16] Brian D. Ripley 1996, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge.
- [17] Somasundaram RS and Nedunchezian R 2011, "Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications Issn-09758887.
- [18] Angeline Christobel Y, Sivaprakasam P 2012, “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, pp. 155-167
- [19] Goldberg, D.E. 1989. “Genetic Algorithms in Search, Optimization, and Machine Learning”, Addison-Wesley: Reading, MA.
- [20] Zhang Y and Wu L 2012, “An MR brain images classifier via principal component analysis and kernel support vector machine,” Progress in Electromagnetics Research, vol. 130, pp. 369–388.
- [21] Krishnanand, K. N., Ghose, D 2005, "Detection of multiple source locations using a glowworm metaphor with applications to collective robotics", In Proceedings of IEEE swarm intelligence symposium, Piscataway IEEE Press, pp. 84–89.
- [22] Krishnanand, K. N., Ghose D 2009, "Glowworm swarm optimisation: a new method for optimising multi-modal functions", Int. J. Computational Intelligence Studies, Vol. 1, No. 1.
- [23] Jiakun Liu, Yongquan Zhou, Kai Huang, zhe Ouyang, Yingjiu Wang 2011, “A Glowworm Swarm Optimization Algorithm Based on Definite Updating

- Search Domains” *Journal of Computational Information Systems* 7: 10, pp.3698-3705.
- [24] Abdelbar A, Abdelshahid, S 2005, "Fuzzy Pso: A Generalization of Particle Swarm Optimization", *Proceeding of International Joint Conference on Neural Network*, Canada.
- [25] Sivanandam S. N, Sumathi S and Deepa S. N 2006, *Introduction to Neural Networks Using Matlab 6.0*, Tata McGraw Hill, New Delhi, India..
- [26] Suresh S, Sundarajan N and Saratchandran P 2008, "A Sequential Multi-Category Classifier Using Radial Basis Function Networks," *Neurocomputing*, vol. 71, nos. 7-9, pp. 1345-1358.
- [27] Zhang R, Huang G.B, Sundararajan N and Saratchandran P 2007, "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.