International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-8, June 2020

An Improved Dragonfly Optimization Algorithm based Feature Selection in High Dimensional Gene Expression Analysis for Lung Cancer Recognition

F. Leena vinmalar, A. Kumar Kombaiya

Abstract: ***A microarray gene expression data is an efficient dataset for analyzing expression of thousands of genes and related disease. The more accurate analysis can be obtained by comparing Gene expression of disease tissues with normal tissues which helps to recognize the type of cancer. The processing of microarray datasets such as feature selection, sampling and classification is highly challenged due to its high dimensionality. Many recent researchers used various feature selection techniques for dimensionality reduction. Dragonfly optimization Algorithm (DA) was a feature selection technique used to reduce the dimensionality of lung cancer gene expression dataset. The dragonflies in DA are flying randomly based on the model developed by using the Levy Flight Mechanism (LFM). Because of huge searching steps, LFM has some drawbacks like interruption of arbitrary flights and overflowing of the search area. In fact, DA lacks an internal resemblance that record past potential solutions that can lead to its premature convergence into local optima. So, in this paper an Improved Dragonfly optimization Algorithm (IDA) is introduced which effectively reduces the dimensionality of the lung cancer gene expression dataset. In IDA, Brownian motion method is used to solve the issues of LFM and pbest and gbest idea of Particle Swarm Optimization (PSO) is used to direct the search method for finding potential candidate solutions to further refine the search space for avoiding premature convergence. The wrapper feature selection approach is followed by IDA to select optimal subset of features. The Random Sub space (RS), Artificial Neural Network (ANN) and Sequential Minimal Optimization (SMO) classifiers are utilized for feature selection of IDA and recognize Lung cancer subtypes. The accuracy of the classifier for selected features of Dragon flies in training instances is used as fitness value of Dragon flies in each iteration. Finally, the experimental results prove the effectiveness of the IDA in terms of accuracy, precision, recall and F-measure.***

Keywords: Lung cancer recognition; gene expression data; Dragonfly optimization Algorithm; Improved Dragonfly optimization Algorithm; Brownian motion method.

I. INTRODUCTION

Uncontrolled cell growth in tissues of the lung is called lung cancer disease. The origin of lung cancer has some unexplained causes for being a hereditary disease. Globally, the mortality rate due to lung cancer disease has risen approximately 25%.

Revised Manuscript Received on June 30, 2020.
***** Correspondence Author
F. Leena vinmalar*, Research Scholar, Department of Computer science, Chikkanna Government Arts College-Tirupur, India. E-mail: anuleena7@gmail.com
Dr. A. Kumar Kombaiya, Assistant Professor, Department of Computer Science Chikkanna Government Arts College-Tirupur, India.

It is higher than the other rising cancers. Effective treatment of lung cancer helps keep other cells from becoming metastatic. Normally, lung cancer is categorized as Small Cell Lung Cancer (SCLC) and Non-SCLC (NSCLC).

The successful detection of lung cancer type is highly critical for the health of patients and for reducing patients' toxicity. The breakthrough in molecular biology provides a positive approach to the use of gene profiles for the detection of lung cancer [2]. From the microarray technology, the information of DeoxyriboNucleic Acid (DNA) RiboNucleic Acid (RNA) and proteins are acquired to diagnosis the tumors in earlier stages. It will improve the probability of survival of cancer patients, particularly for patients with lung cancer. The classification of lung cancer must be precise and consistent to ensure the correct diagnosis and appropriate care for lung cancer cases. It can be achieved by using machine learning techniques [3], those examine the gene expression in cancer cells and identify the genes that are most likely to cause cancer. The application of enormous gene expression data in machine learning algorithm that generate the most effective candidate genes, which have a greater predictive value individually or as a part of complex method to evaluate the susceptibility of someone to cancer.

Gene expression data is a high dimensional data that contain vast numbers of duplicate genes in the retrieval of information related to the disease [4]. In order to cope this issue, the most significant features are selected, that eliminates irrelevant and redundant genes. A Nested Genetic Algorithm (NGA) [5] was selecting the most discriminative features for lung cancer diagnosis. It consisted of two genetic algorithms are Outer Genetic Algorithm (OGA) and Inner Genetic Algorithm (IGA). The OGA and IGA were worked on microarray gene expression dataset and DNA methylation datasets respectively. The selected features were processed in Random Sub space (RS), Artificial Neural Network (ANN) and Sequential Minimal Optimization (SMO) algorithms for lung cancer diagnosis. However, NGA has time complexity and random motion problems which can affect the lung cancer diagnosis accuracy.

In this paper, an Improved Dragonfly optimization Algorithm (IDA) is introduced for feature selection which solves time complexity and random motion problem in NGA-based feature selection for lung cancer diagnosis. The IDA is an improved version of DA which is influenced by the dynamic and static activities of dragonflies in nature. DA is developed with the Levy Flight Mechanism (LFM). When there is no neighborhood solution, the LFM models the search method for the best solution of dragonflies. LFM has internal memory problem and sometimes it leads to premature convergence. So, in IDA a Brownian motion is used instead of LFM. In Brownian motion, the movement of dragonflies takes the form of normal distribution.

In order to solve this problem and to improve the efficiency of feature selection, some theory of PSO is introduced in DA. The selected features by IDA are given as input to RS, ANN and SMO classifiers to predict the lung cancer effectively.

II. LITERATURE SURVEY

Hosseinzadeh et al. [6] proposed a diagnostic system according to the sequence-derived physicochemical and structural features for prediction of lung tumor types. Initially, the features in the collected data were extracted and then attribute weighting algorithm was applied to select the most appropriate features for lung cancer prediction. Finally, Naïve Bayes (NB), Artificial Neural Network (ANN) and Support Vector Machine (SVM) were applied with selected features for prediction of lung cancer. However, it has high computation overhead when huge number of attributes is used for lung cancer prediction.

Dass et al. [7] used a data mining technique to categorize the lung cancer types. A combined classification decision tree induction algorithm was processed in the proteomic and genomic datasets for lung cancer classification. The decision tree was built using J48 algorithm provided an advanced decision tree induction technique. With the help of this technique, the top classification rules were received using Apriori technique. It used for prediction of lung cancer. However, computation time is high because of using Apriori algorithm.

Azzawi et al. [8] proposed a Gene Expression Programming (GEP) model for forecasting the lung cancer from microarray datasets. An n-Top ranked gene selection method was applied on the collected dataset to extract the gene features and these features were given as input to the radial basis function, multi-layer perceptron and Support Vector Machine (SVM) for lung cancer prediction. However, it needs improvement in terms of accuracy.

Ramos-González et al. [9] proposed a framework with gradient boosting based feature selection method to classify the subtypes of lung cancer. However, the pathway analysis in the framework will be utilized in future for providing knowledge about the processes carried out in tumor cells.

Pati [10] proposed an eco-genomics approach for prediction of lung cancer. An information gain attribute ranking techniques was applied to compute the gene expression score which was used to select the most probable genes in the data. The most probable genes were processed in the Sequential Minimal Optimization (SMO), multi-layer perceptron and random subspace for classification of data as patient with presence of lung cancer and absence of lung cancer. However, it has high computational complexity problem.

Wu et al. [11] build a model to find the probability of getting NSCLC which helps to diagnosis lung cancer. This model played a significant role to find the probability of NSCLC in different phase. In each phase of the model, the maximum effect was determined with the significance diagnosis and top three high decision data through adopting the selecting effective parameters with big data. However, it is just an assistant program that does not substitute for physicians with correct assessment in NSCLC.

ALzubi et al. [12] proposed an ensemble method for lung cancer diagnosis. During the feature selection process in the ensemble method, a combined method was applied for selection of most discriminative features which reduced the classification time. After the selection of optimal features, a boosted method was applied for classifying the patients. While the ensemble method has been validated with different datasets, a massive amount of data points will be checked with the ensemble method which leads to high computation time for lung cancer diagnosis.

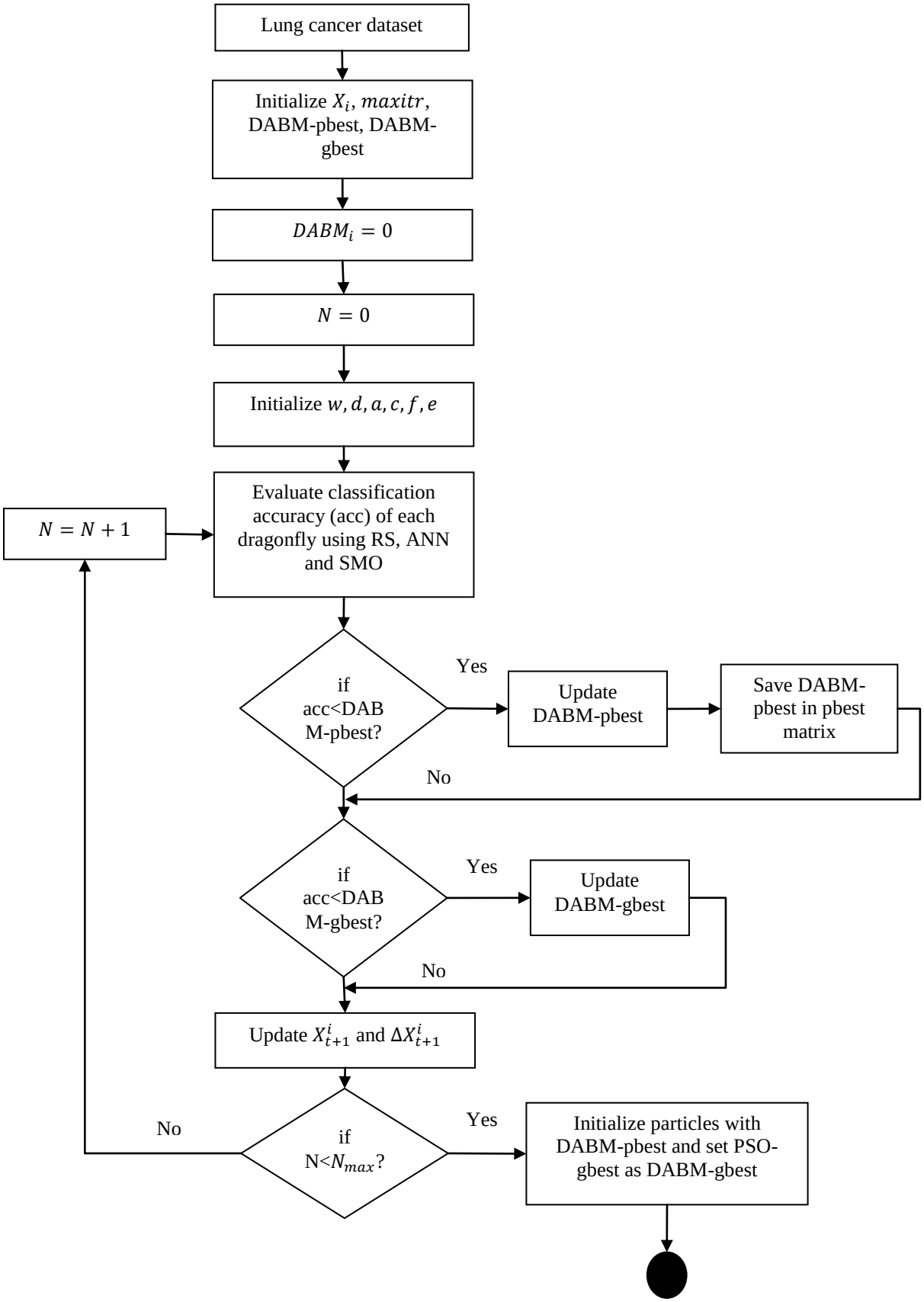
III. PROPOSED METHODOLOGY

Here, the proposed IDA with different classifiers for ling cancer diagnosis is described in detail. Initially, gene expression data is collected and then the feature selection techniques such as DA and IDA are applied to select the most significant features. The selected features are given as input to RS, ANN and SMO for lung cancer diagnosis. The workflow of this proposed work is given in Fig. 1.

A. Feature Selection using Dragonfly optimization Algorithm

The collected gene expression data is given as input to DA for the selection of most discriminative features. DA is a meta-heuristic algorithm based on swarm intelligence. It is influenced by dragonfly's stagnant and complex behavior. Dragonflies are known as little hunters killing virtually all other little insects in the wild. Some aquatic insects and even tiny fish are predated by nymph dragonflies. The two mechanisms of swarm dragonflies are hunting and migration. The first mechanism is called as static (feeding) swarm and the second mechanism is called as dynamic (migratory) swarm.

Dragonflies join small groups in static swarm, which move through a confined area and attract other travelling preys such as butterflies and mosquitoes. The key features of a static swarm are spatial motions and sudden shifts in the moving direction. However, the swarm travelled one direction for long periods due to a huge number of dragonflies in dynamic swarms.


Lung cancer dataset
Initialize X_i, $maxitr$, DABM-pbest, DABM-gbest
$DABM_i = 0$
$N = 0$
Initialize w, d, a, c, f, e
Evaluate classification accuracy (acc) of each dragonfly using RS, ANN and SMO
$N = N + 1$
if acc<DABM-pbest?
Yes
Update DABM-pbest
Save DABM-pbest in pbest matrix
No
if acc<DABM-gbest?
Yes
Update DABM-gbest
No
Update X_{t+1}^i and ΔX_{t+1}^i
if N<N_{max}?
No
Yes
Initialize particles with DABM-pbest and set PSO-gbest as DABM-gbest

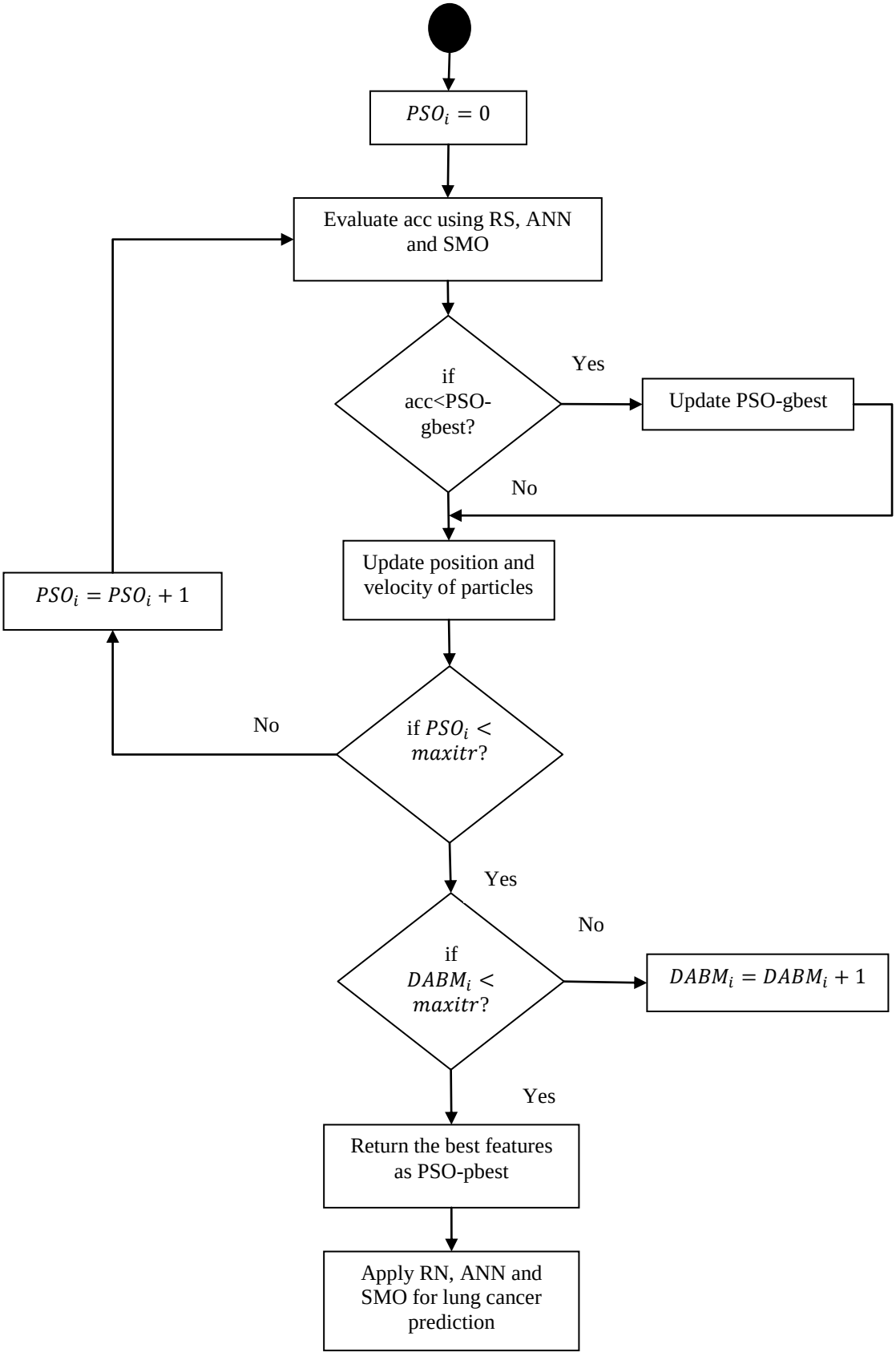


Fig. 1.Workflow of IDA with different classifier for lung cancer prediction

The dynamic and static swarming activities represent discovery and manipulation quite closely to two main phases of meta-heuristic optimization. The static swarm is complimentary in the exploration phase while the dynamic swarm is complimentary in the exploitation phase. The exploration and exploitation are replicated either statistically or dynamically probing for optimal features or remove the irrelevant features in the gene expression data.

The survival is the main intention of any swarm, therefore all of the individuals should attracted towards the dragonflies which has high classification accuracy and distracted outwards the dragonflies which has irrelevant features. Each of the behaviors of dragonflies is mathematically expressed as follows:

$$D_i = -\sum_{n=1}^{N} X - X_n \quad (1)$$

In (1), D_i is the separation of ith individual, X is the current individual, X_n shows the position jth neighboring individual, and N is the number of neighboring individuals.

$$A_i = \frac{\sum_{n=1}^{N} Vel_n}{N} \quad (2)$$

In (2), A_i is the alignment of ih individual, Vel_n is the velocity of nth neighboring individual.

$$Coh_i = \frac{\sum_{n=1}^{N} X_n}{N} - X \quad (3)$$

In (3), Coh_i is the cohesion of ith individual. The following equation shows the attraction of dragonflies towards a food source.

$$Food_i = X^+ - X \quad (4)$$

In (4), $Food_i$ is the food source of the ith individual, X^+ represents the position of the food source (classification accuracy) and X represents the position of the current individual. Distraction outwards irrelevant feature is calculated as follows:

$$Ene_i = X^- + X \quad (5)$$

In (5), Ene_i is the position of the irrelevant features and X^- is the position of irrelevant features. In order to update the position of artificial dragonflies in the search space and recreate their motions, two vectors such as step (ΔX) and position (X) are considered. ΔX represents the direction of the dragonfly motions which is calculated as,

$$\Delta X_{t+1} = (dD_i + aA_i + cCoh_i + fFood_i + eEnei + w\Delta Xt \quad (6)$$

In (6), d denotes the separation weight, f is the food factor, a is the alignment weight, e is the enemy factor, c is the cohesion weight, w is the inertia weight and t is the iteration counter.

The position vectors are calculated after the step vector calculation, which is given as follows:

$$X_{t+1} = X_t + \Delta X_{t+1} \quad (7)$$

Dragonflies desired to organize their movement in the dynamic swarm. Dragonfly arrangement is incredibly sluggish in the static motion, although it is very highly capable of fighting the opponent. Therefore, the arrangement coefficient is high and the harmony coefficient is very low in discovery, whereas the arrangement coefficient is small in the extraction cycle and the harmony coefficient is large.

If there is no neighborhood solution to improve the discovery of artificial dragonflies, a randomness solution is achieved using LFM. As a result, the dragonfly's position is updated as,

$$X_{t+1} = X_t + Levy(d) \times X_t \quad (8)$$

In (8), d represents the size of the position vector.

$$Levy(x) = 0.01 \times \frac{r_1 \times \sigma}{|r_2|^{1/\beta}} \quad (9)$$

In (9), r_1 and r_2 are random numbers which is ranges from 0 to 1 and β is the constant value. σ is calculated as,

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}} \right) \quad (10)$$

In (10), $\Gamma(x) = (x-1)!$.

Drangonfly algorithm for feature selection

Initialize the population of dragonflies $X_i (i = 1,2, \ldots h)$
Initialize X_i with features of gene expression data
while the end condition is not satisfied
Call RS/ANN/SMO to find the classification accuracy
Revise the food source and enemy
Revise w, d, a, c, f and e
Compute $D, A, Coh, Food$ and Ene using (1) to (5)
Revise neighboring radius
if a dragonfly has at least one neighboring dragonfly
    Revise velocity vector using (6)
Revise position vector using (7)
else
    Revise position vector using (8)
end if

According to the variable boundaries check and correct the new positions
end while

B. Feature selection using Improved Dragonfly optimization Algorithm

To some point, the LEW enhances DA's productivity but depending on the specifications of the process, is adverse in the very lengthy steps. Such key steps in the algorithm are evaluated in two modes. One of the modes is to keep individual out of the search space and to create a new step vector while taking a long step. However, it is not clear whether this solution will always deliver the correct results. The new step to be produced may take the common reversal back. Another mode is to consider 1% of the step size or a particular percentage according to the variable range in the lung cancer diagnosis. This is a different alternative than the former mode. Hover, the step size regulation goes against nature of LFM. So, in the IDA another random motion mechanism called Brownian motion is used instead of LFM for feature selection.

The Brownian motion is influenced by the movement of free gas or liquid molecules. The Brownian motion is the spontaneous displacement of particles in a substance which is disrupted through collisions with rapidly travelling molecules in the fluid. It is assumed as Markov process with Gaussian process and a continuous path that takes place continually over time.

$$X_{t+1} = X_t + W_t \quad (11)$$

In (11), W_t is a random vector created from known probability distribution. If W_t is generated from the Gaussian distribution, random walking is isotropic.

Here, the motion takes the form of regular distribution called Brownian motion. With square root-scaling, the expected step size can be modeled as follows,

$$R(t) \propto \sqrt{t} \qquad (12)$$

The diffusion becomes abnormal when the steps W_t are obtained from LFM. Here, the expected step size becomes,

$$R(t) \propto t^q, q > 0 \qquad (13)$$

The IDA by the Brownian motion is given as follows:

$$X_{t+1} = X_t + o \times rand() \times K_g \qquad (14)$$

where,

$$o = \sqrt{\frac{M}{S}} \qquad (15)$$

In (15), M denotes the motion time period in seconds of a dragonfly and S denotes the number of sudden motions for the same dragonfly in proportion to time.

$$S = 100 \times M \qquad (16)$$

$$K_g = \frac{1}{o\sqrt{2\pi}} exp\left(-\frac{(dimension - dragonfly)^2}{2o^2}\right) \qquad (17)$$

In the Brownian motion, steps are selected based on Gaussian distribution. The daily movement of the dragonflies has been dispersed over time and abrupt leaps and random movements are produced as vibrations.

Although DA is tradeoff between the global and local search capabilities, the DA to be deficient in internal memory that records previous potential solutions. In the meantime, DA rejects all the fitness values go beyond the global best and rarely records potential solutions that will leads to global optima. It affects the output of DA, which leads to very slow convergence and stagnates at local optima. So, two features are included in DA to avoid this problem. One of the features is an internal memory that monitors possible solutions that convergence to global optima and another feature iteration level which implements on this set of saved solutions. Each dragonfly can monitor its co-ordinates in the hyperspace of problems that are related to the accuracy of classifiers. It is achieved by including internal memory feature with DA.

At every iteration, the classification accuracy of each search agnet in current population is compared with the best fitness value in that iteration. The best solutions are stored and DABM-pbest matrix is formed using DA with Brownian motion. Dragonflies are also made to monitors best value obtained so far any dragonfly in the neighborhood which is same as gbest concept of PSO and is saved as DABM-gbest. It improves the exploitation capability of DA and it also helps to escape from local optima.

In order to find better features of gene expression data for lung cancer diagnosis, PSO is initialized with DABM-pbset and DABM-gbest is assigned as the gbest of PSO. The position and velocity update in IDA is given as follows:

$$\Delta X_{t+1}^i = w\Delta X_t^i + C_1 r_1 \left(DABM_{pbest_t^i} - X_t^i\right) + C_2 r_2 \left(DABM_{gbest_t^i} - X_t^i\right) \qquad (18)$$

$$X_{t+1}^i = X_t^i + \Delta X_{t+1}^i \qquad (19)$$

In (18), $DABM_{pbest_t^i}$ is the pbest of ith particle of PSO and $DABM_{gbest_t^i}$ is the gbest of ith particle of PSO. Therefore, IDA combines DA's exploration features and PSO's exploitation features in order to attain maximal solutions globally.

Hence, the IDA integrates the exploration features of DA in initial stage and exploitation capabilities of PSO in final stage to achieve global optimal solutions. The overall process of IDA based feature selection for lung cancer diagnosis is given in the following algorithm.

Improved Drangonfly algorithm for feature selection

Initialize the dragonflies population $X_i (i = 1,2, \dots h)$, maximum iteration $maxitr$, $t = 0$, number of search agents N, maximum number of search agents N_{max}

Initialize X_i with features of gene expression data

while the end condition is not satisfied

for each dragonfly

Call RS/ANN/SMO to find the classification accuracy

if classification accuracy<DABM-pbest

Migrate the current value to DABM-pbestmatrix

end **if**

if classification accuracy<DABM-gbest

Assign current value as DABM-gbest

end **if**

end **for**

For each dragonfly

Revise the food source and enemy

Revise w, d, a, c, f and e

Compute $D, A, Coh, Food$ and Ene using (1) to (5)

Revise neighboring radius

if a dragonfly has at least one neighboring dragonfly

Revise velocity vector using (6)

Revise position vector using (7)

else

Revise position vector using (14)

end if

According to the variable boundaries check and correct the new positions

end while

For each particle

Initialize particles with DABM-pbestmatrix

Assign PSO-gbest as DA-gbest

end

while $maxitr$ is not attained

For each particle

Call RS/ANN/SMO to find the classification accuracy

if classification accuracy<PSO-pbest in history

Assign current value as the new PSO-pbest

end **if**

end **for**

Select the particle with the best fitness value of all the particles as the PSO-gbest

For each particle

Compute the particle velocity using (18)

Update particle position using (19)

end **for**

end **while**

best-fitness=PSO-gbest

end **while**

C. Lung Canncer Diagnosis using Random SubSpace Classifier

RS is well suited for gene expression data which process the selected features by DA and IDA to diagnosis the lung cancer. This adds up the resulting base classifiers trained at different subsets of levels of genetic expression and can increase diversity among the simple learners while rejecting the available gene expression information with no substantial loss of accuracy. The process of RS is explained as follows:

1. For a d-dimensional dataset $D = \{x_i, y_i\}, (i = 1,2,\ldots,T$, where xi is the features selected by DA or IDA, y_i is the lung cancer diagnosis class and T is the number of training objects, choose $dim_i (dim_i < D)$ as the number of input variable for each classifier. There exists one value of dim_i for all individual classifiers.
2. Generate training set by selecting dim_i features from D without replacement. The features are selected without replacement for each classifier I.
3. The results of individual classifiers are integrated by majority voting to diagnosis lung cancer.

D. Lung Canncer Diagnosis using Artificial Neural Network Classifier

The selected features by DA and IDA are given to train the ANN classifier for diagnosis of lung cancer. ANN consists of three layers namely input, hidden and output layer. The features of gene expression data are denoted as $f(x) = x$ are given to the input layer of neurons. The hidden layer of ANN is defined as tan-sigmoid function.

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \qquad (20)$$

Each input has its own weight values as $w_1, w_2, \ldots, w_n$ and weighted sum of the inputs is done by the adder function as follows,

$$u = \sum_{i=1}^{n} w_i x_i \qquad (21)$$

The output layer of ANN is described as follows,

$$y = f(\sum_{i=1}^{n} w_i x_i + b_i) \qquad (22)$$

In (22), y is the output neuron value, $f(x)$ is the transfer function, w_i refers the weight values, x_i is the selected features of gene expression data. Based on the output neuron values, the lung cancer is diagnosed. The error rate of the output is calculated as,

$$\varepsilon_o = \frac{W_i X_i + b}{\varepsilon_H} \qquad (23)$$

In (23), ε_o is the error rate on output nodes and ε_H denotes the error rate at hidden layer nodes.

Artificial Neural Network Algorithm

1. Initialize all weights and biases in network
2. Calculate weighted sum of the inputs using (21).
3. Obtain the results of each node using activation function as in (22).
4. Compute the error rate of output using (23).
5. Update weight and bias values to reduce the error rate.

E. Lung Canncer Diagnosis using Sequential Minimal Operator

Sequential Minimal Optimization (SMO) classifier decomposes the quadratic programming problem into quadratic programming sub-problems by choosing a set of only two-points as the working set which is the smallest amount due to the following condition:

$$\sum_{x=1}^{n} \alpha_i y_i = 0 \qquad (24)$$

In (24), α_i is the Lagrange multiplier and y_i is the lung cancer diagnosis class. Each iteration of SMO uses only few operations to overall increase the speed of some orders of magnitude. It provides improvement in efficiency, and the SMO speed-up Support Vector Machine (SVM) algorithm for diagnosis of lung cancer. Some of the indexset Ind are defined as follows which denotes the training data pattern:

$$Ind_0 = \{ind: y_{ind} = 1, 0 < a_{ind} < c\} \cup \{ind: y_{ind} = -1, 0 < a_{ind} < c\} \qquad (25)$$

$$Ind_1 = \{ind: y_{ind} = 1, a_{ind} = 0\} \quad \text{(Positive Non-Support Vectors)} \qquad (26)$$

$$Ind_2 = \{ind: y_{ind} = -1, a_{ind} = c\} \quad \text{(Bound Negative Support Vectors)} \qquad (27)$$

$$Ind_3 = \{ind: y_{ind} = 1, a_{ind} = c\} \quad \text{(Bound Positive Support Vectors)} \qquad (28)$$

$$Ind_4 = \{ind: y_{ind} = -1, a_{ind} = 0\} \quad \text{(Negative Non-Support Vectors)} \qquad (29)$$

In above equations, c denotes the correction parameter. The bias b_{up} and b_{low} are defined with their associated indices

$$b_{up} = \min\{f_{ind}: ind \in Ind_0 \cup Ind_1 \cup Ind_2\} \qquad (30)$$

$$Ind_{up} = \arg\min_{ind} f_{ind} \qquad (31)$$

$$b_{low} = \max\{f_{ind}: ind \in Ind_0 \cup Ind_3 \cup Ind_4\} \qquad (32)$$

$$Ind_{low} = \arg\max_{ind} f_{ind} \qquad (33)$$

The optimality conditions are tracked through the vector:

$$f_{ind} = \sum_{j=1}^{l} a_j y_j K(X_j, X_{ind}) - y_{ind} \quad (34)$$

In (34), K is the kernel function and X_i is training data points. SMO optimize two a_i corresponding to b_{up} and b_{low} according to the following:

$$a_2^{new} = a_2^{old} - y_2(f_1^{old} - f_2^{old})/\eta \quad (35)$$

In (35), $\eta = 2k(X_1, X_2) - k(X_1, X_1) - k(X_2, X_2)$.

$$a_1^{new} = a_1^{old} - s(a_2^{old} - a_2^{new}) \quad (36)$$

After optimizing a_1 and a_2, f_{ind} which denotes the error of indth training data is updated according to the following:

$$f_{ind}^{new} = f_{ind}^{old} + (a_1^{new} - a_1^{old})y_1 k(X_1, X_{ind}) + (a_2^{new} - a_2^{old})y_2 k(X_2, X_{ind}) \quad (37)$$

The weight vector is updated as,

$$\vec{w}^{new} = \vec{w} + y_1(a_1^{new} - a)\vec{x}_1 + y_2(a_2^{new,clipped} - a2x \quad (38)$$

The optimality of the solution for lung cancer diagnosis is checked by calculating the optimality gap that is the gap between b_{up} and b_{low}. SMO algorithm is terminated when $b_{low} \leq b_{up} + 2\tau$.

SMO algorithm

1. Initialize $\alpha_{ind} = 0, f_{ind} = -y_{ind}$
2. Calculate $b_{up}, b_{low}, Ind_{up}, Ind_{low}$ using Eq. (30) to Eq. (33)
3. Repeat
4. Update f_{ind} using (37)
5. Calculate $b_{up}, b_{low}, Ind_{up}, Ind_{low}$
6. Update α_{high} and α_{low}
7. Until $b_{low} \leq b_{up} + 2\tau$
8. Update the bias b
9. Store the new α_1 and α_2 values
10. Update weight vector w using (38)
11. Apply w, α and b in SVM to predict the lung cancer.

IV. RESULT AND DISCUSSON

Here, the efficiency of NGA, DA and IDA with different classifiers is tested in terms of accuracy, precision, recall and F-measure. A lung cancer dataset is used for experimental purpose. The lung cancer dataset consists of 12,625 genes and 56 samples. The instance in lung cancer dataset ranges from AD2 to AD384. The lung cancer dataset is available in the http://grafia.cs.ucsb.edu/autodecoder/dataset.html.

A. Accuracy

It measures the fraction of correct lung cancer predictions to the total number of instances evaluated. It is calculated as,

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + False\ Negative\ (FN)}$$

The accuracy value of NGA, DA and IDA with different classifiers for lung cancer diagnosis is given in Table 1.

TABLE I. Evaluation Of Accuracy

	NGA	DA	IDA
RS	0.733	0.822	0.857
ANN	0.80	0.832	0.867
SMO	0.857	0.896	0.93

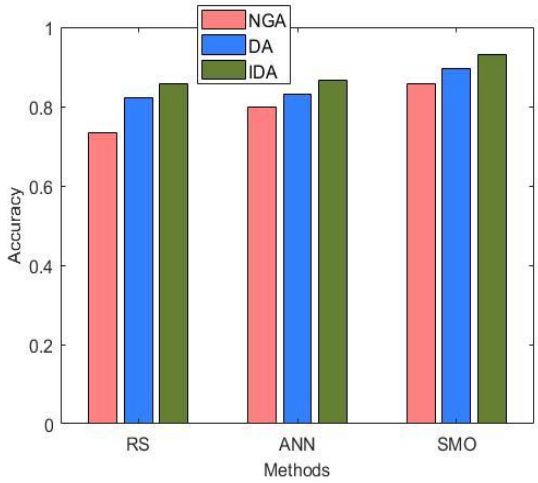


Fig. 2.Comparison of Accuracy

The accuracy of NGA, DA and IDA with RS, ANN and SMO classifiers for lung cancer diagnosis is shown in Fig. 2. The methods are taken in X axis and the accuracy value is taken in Y axis. The accuracy of IDA-SMO is 8.52% and 3.79% greater than IDA-RS and IDA-ANN method for lung cancer diagnosis. From Fig. 2, it is proved that the IDA-SMO has better lung cancer diagnosis accuracy than other methods.

B. Precision

It is used to measure the positive patterns (i.e., presence of lung cancer) that are correctly predicted from the total predicted patterns in the positive class. It is calculated as,

$$Precision = \frac{TP}{TP + FP}$$

The precision value of NGA, DA and IDA with different classifiers for lung cancer diagnosis is shown in Table 2.

TABLE II. Evaluation of Precision

	NGA	DA	IDA
RS	0.721	0.764	0.854
ANN	0.804	0.846	0.875
SMO	0.89	0.90	0.937

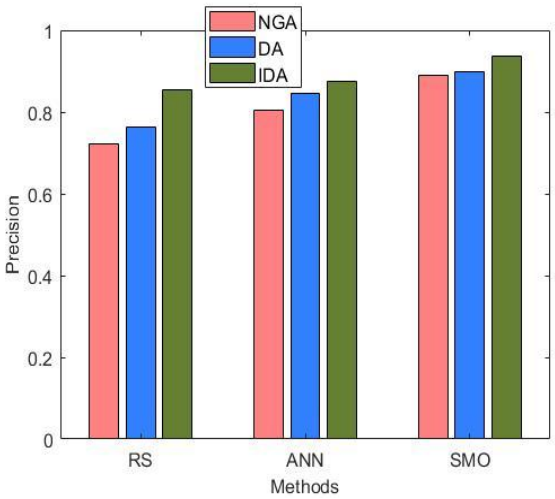


Fig. 3.Comparison of Precision

The precision of NGA, DA and IDA with RS, ANN and SMO classifiers for lung cancer diagnosis is shown in Fig. 3. The methods are taken in X axis and the precision is taken in Y axis. The precision of IDA-SMO is 5.28% and 4.11% greater than IDA-RS and IDA-ANN method for lung cancer diagnosis. From Fig. 3, it is proved that the IDA-SMO has better lung cancer diagnosis precision than other methods.

C. Recall

It is used to measure the ratio of positive patterns that are correctly classified. It is calculated as,

$$Recall = \frac{TP}{TP + TN}$$

The recall value of NGA, DA and IDA with different classifiers for lung cancer diagnosis is shown in Table 3.

TABLE III. Evaluation of Recall

	NGA	DA	IDA
RS	0.73	0.78	0.83
ANN	0.816	0.876	0.90
SMO	0.833	0.923	0.937

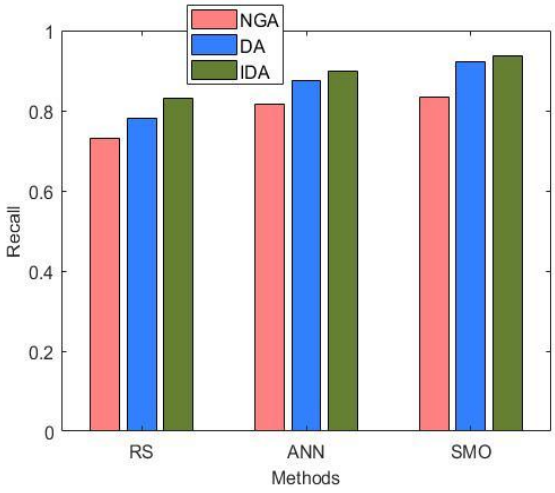


Fig. 4.Comparison of Recall

The recall of NGA, DA and IDA with RS, ANN and SMO classifiers for lung cancer diagnosis is shown in Fig. 4. The methods are taken in X axis and the recall is taken in Y axis. The recall of IDA-SMO is 12.48% and 1.52% greater than IDA-RS and IDA-ANN method for lung cancer diagnosis. From Fig. 4, it is proved that the IDA-SMO has better lung cancer diagnosis recall than other methods.

D. F-measure

It is the mean between the precision and recall. It is calculated as,

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

The F-measure value of NGA, DA and IDA with different classifiers for lung cancer diagnosis is given in Table 4.

TABLE IV. Evaluation of F-measure

	NGA	DA	IDA
RS	0.727	0.76	0.84
ANN	0.81	0.86	0.895
SMO	0.895	0.91	0.93

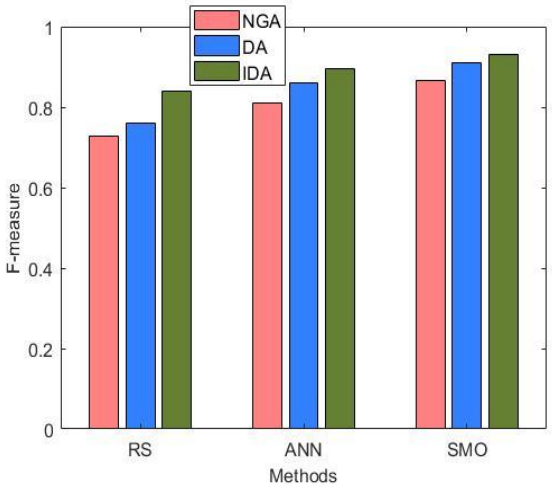


Fig. 5.Comparison of F-measure

F-measure of NGA, DA and IDA with RS, ANN and SMO classifiers for lung cancer diagnosis is shown in Fig. 5. The methods are taken in X axis and the F-measure is taken in Y axis. The F-measure of IDA-SMO is 7.51% and 2.2% greater than IDA-RS and IDA-ANN method for lung cancer diagnosis. From Fig. 5, it is proved that the IDA-SMO has better lung cancer diagnosis F-measure than other methods.

V. CONCLUSION

In this article, IDA is proposed to select the most discriminative features of gene expression data for lung cancer diagnosis. It solves lack of internal memory problem, premature convergence problem and random motion problem. A Brownian motion is used in DA instead of LFM to update the movement of the dragonfly when there is no neighborhood solution. It enhanced the significance of neighborhood radius used in DA, both while facilitating the movement of dragonflies and saving time. Moreover, the particle best and global best idea of PSO is included to DA to direct search space for potential candidate solutions for lung cancer diagnosis and PSO is then initialized with particle best of DA with Brownian motion to further exploit the search space.

After that, the PSO process is carried out to select the most discriminative features based on the classification accuracy of classifiers such as RS, ANN and SMO. Finally, the selected features are applied in RS, ANN and SMO classifiers to predict the lung cancer effectively. The experimental results show that the proposed IDA-SMO method has better accuracy, precision, recall and F-measure for lung cancer diagnosis than other methods.

REFERENCES

1. H. Azzawi, J. Hou, R. Alanni, Y. Xiang, R. Abdu-Aljabar and A. Azzawi, "Multiclass lung cancer diagnosis by gene expression programming and microarray datasets," In Int. Conf. Adv. Data Min. Appl., Springer, Cham, pp. 541-553, 2017.
2. H. Azzawi, J. Hou, Y. Xiang, and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," IET syst. boil., vol. 10, no. 5, pp. 168-178, 2016.
3. J. Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach," IEEE Access, vol. 7, pp. 4232-4238, 2018.
4. A. Wahid, D. M. Khan, N. Iqbal, S. A. Khan, A. Ali, M. Khan, and Z. Khan, "Feature selection and classification for gene expression data using novel correlation based overlapping score method via Chou's 5-steps rule," Chemom. Intell. Lab. Syst., vol. 199, pp. 1-24, 2020.
5. S. Sayed, M. Nassef, A. Badr, and I. Farag, "A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets," Expert Syst. Appl., vol. 121, pp. 233-243, 2019.
6. F. Hosseinzadeh, A. H. KayvanJoo, M. Ebrahimi, and B. Goliaei, "Prediction of lung tumor types based on protein attributes by machine learning algorithms," SpringerPlus, vol. 2, no. 1, pp. 1-14, 2013.
7. M. V. Dass, M. A. Rasheed, and M. M. Ali, "Classification of lung cancer subtypes by data mining technique," In Proc. IEEE Int. Conf. Control, Instrum., Energy Commun., pp. 558-562, 2014.
8. H. Azzawi, J. Hou, Y. Xiang, and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," IET syst. boil., vol. 10, no. 5, pp. 168-178, 2016.
9. J. Ramos-González, D. López-Sánchez, J. A. Castellanos-Garzón, J. F. de Paz, and J. M. Corchado, "A CBR framework with gradient boosting based feature selection for lung cancer subtype classification," Comput. boil. med., vol. 86, pp. 98-106, 2017.
10. J. Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach," IEEE Access, vol. 7, pp. 4232-4238, 2018.
11. J. Wu, P. Guan, and Y. Tan, "Diagnosis and data probability decision based on Non-small cell lung cancer in medical system," IEEE Access, vol. 7, pp. 44851-44861, 2019.
12. J. A. ALzubi, B. Bharathikannan, S. Tanwar, R. Manikandan, A. Khanna, and C. Thaventhiran, "Boosted neural network ensemble classification for lung cancer disease diagnosis," Appl. Soft Comput., vol. 80, pp. 579-591, 2019.

AUTHOR PROFILE

Name : **A. Kumar Kombaiya**
Designation : Assistant Professor
Address : 4SG, Jain Vengateswara Apartment
Ramanujam Nagar, Uppalipalayam (PO),Coimbatore – 641 015.
Contact Number : 98 94 83 80 90
Email Id : kumar_kombaiya@rediffmail.com
Date of Joining in Collegiate Education : 09/07/2009
Date of Joining in the Present College : 16/07/2010

Academic Profile:

Degree	Institute /College	University	Period
B.Sc	Aditanar college , Tiruchendur	Madurai Kamaraj University	1986
MCA	Alagappa University	Alagappa University	1990
M.Phil	Bharathidasan University	Bharathidasan University	2005
Ph.D	Alagappa University	Alagappa University	2016

Teaching Experience

i) **Total** : 28years, 5months.
ii) **UG** : 28years, 5months.
iii) **PG** : 25years, 1month.

Name of the College	Position held	Period
Govt Arts College for Women Krishnagiri-1	Assistant Professor	2009-2010
Chikkanna Govt Arts College-Tirupur	Assistant Professor	2010-till date

Honors and Research Awards:

Field of Interest

i) **Teaching** : Computer Science
ii) **Research** : Data Mining
iii) **Proficiency in Instrumentation** :

Research Guidance
Guidance Number : M. Phil (15681/A2/2005-13/1/2006)
Ph. D (15/12/A2/2017-13/02/2017)

(If have more than one university, given them against University name)

S. No	M. Phil/Ph. D	Name of the Student	Thesis Title	Completed/Ongoing
1.	M. Phil	R.ANITHA	Efficient sub tree based data stream cluster and outliers	Completed
2.	M. Phil	R.SATHISH KUMAR	Interpreting the public sentiment variations on twitter	Completed
3.	M. Phil	S.GNANASOUNDARI	Efficient routing in delay tolerant network based on secure fuzzy spray decision algorithm	Completed

4.	M. Phil	P. MUTHUKUMAR	Prediction and analysis of stack market trading using enhance support vector machine classifier	Completed
5.	M. Phil	C. VADIVELMURUGAN	Location prediction and pupation in mobility communication network	Completed
6.	M. Phil	F. LEENA VINMALAR	A novel cancer gene search model and classification using GPSO- BPNN	Completed
7.	M. Phil	A. GOMATHI JAYAM	Enhance ADA-Boost approach for classification of cancer gene expression	Completed
8.	Ph. D	GANESHAN A		Ongoing
9.	Ph. D	PREMA K		Ongoing
10.	Ph. D	VIDYA B		Ongoing
11.	Ph. D	LEENA VINMALAR F		Ongoing
12	Ph. D	PRAKASH K		Ongoing

Funded Projects

Membership in Professional Bodies

S. No	Name of the Professional Body	Member ship Detail with Number
1.	Board of studies in Computer Science(PG) Bharathiar University	Chairman (PG)
2.	Board of studies in Computer Science(UG) Bharathiar University	Member (UG)
3.	Ramakrishna College of Arts and Science Coimbatore	Member (BCA)

Research Publications

i) Research Papers : 16 Ref: ANNEXURE I

National and International Conferences :

i) Participated & Paper Presented : 4 Ref: ANNEXURE II

ii) Poster Presented : Nil

Conference/ Seminars Organized

Workshop attended : Nil

Resource Person/Invited Lectures :

Faculty Development Programs Attended

Course	University/Institute	Subject	Period
Orientation Course	Madras University	-	09/07/2009-05/08/2009
Refresher Course	Bharathiar University	Computer Science	15/09/2010-05/10/2010
Refresher Course	Bharadhidasan University	Computer Science	13/11/2015-03/12/2015

Academic Activities

i) **Subject Handled** : Computer Science Subjects

ii) **Class Advisor** : UG & PG

iii) Special Coaching :

iv) Student Community Beneficial Activities :

v) Co- curricular and extracurricular activities : Computer Learning Programme.

Professional Activities

i) Reviewer :

ii) **Board of Studies/UR** : Chairman - Board of studies in computer science (PG), Bharathiar University.

iii) **Examiners/ Scrutiny** : Bharathiar, Bharathidasan, Madurai Kamarajar, Alagappa, Madras Universities

iv) Senates/Syndicate :

National / International Collaborators

ANNEXURE I

1. Dr.A.Kumar Kombaiya- IEEE International Conference on Computational Intelligence and Computing Research ,pp.700-704-ISBN:9781-4799."**Efficient Analysis of Pharmaceutical Compound Structure Based on Pattern Matching Enhanced Algorithm in Data Mining Techniques"**-on December 2014.
2. Dr.A.Kumar Kombaiya-International Journal of Technology Enhancements And Emerging Engineering Research-ISSN:2347-4289."**Efficient Routing in Delay Tolerant Network Based Secure Fuzzy Spray Decision Algorithm"**-on 2014.
3. Dr.A.Kumar Kombaiya-International Journal of Innovative Research in Computer Communication Engineering(IJRCCE),Vol.3-ISSN:3362-3367."**Efficient Analysis of Pharmaceutical Compound Structure Based on Enhanced K-means Clustering Algorithm"**-on April 2015.
4. Dr.A.Kumar Kombaiya-International Journal of Computer Sciences and Engineering(IJCSE)Vol.3.-pp58-63."**Designing a Knowledge Discovery of Clustering Techniques in Pharmaceutical Compounds"**-on April 2015.
5. VI,J.Published Paper on "**A Novel Cancer Gene Search Model And Classification using GPSO-BPNN**"On ,June-2016.

6. Prema.K and Dr.A.Kumar kombaiya ,IJARSE International Journal of Advance Research in Science and Engineering ISSN(O):2319-8354,ISSN(P):2319-8346 Published Paper on "**A Survey on use of Data Mining Methods Techniques and Applications**"on December 2017 .
7. A.Ganesan and Dr.A.Kumar kombaiya ,JETIR –IJSART . ISSN:2349-5162, Volume 4,Issue 1,ISSN:2395-1052, Published Paper on "**Classification Model for Intrusion Detection** "on January 2018.
8. F.Leena vinmalar and Dr.A.Kumar Kombaiya –International Journal On Computer Science and Engineering(SSRG-IJCSE)-Special Issue ICCREST-April 2018. –IJETS.ISSN(P):2348-8387,Volume 56,February 2018-Published Paper on "**Cancer Gene Expression Model An Classification Using Fuzzy Clustering And TSVM-SSRG**"On ,February-2018.
9. F.Leena vinmalar and Dr.A.Kumar Kombaiya –International Journal On Engineering Technology and Sciences –IJETS.ISSN(P):2349-3988,ISSN(O):2349-3976.Volume III,Issue F.Leena vinmalar and Dr.A.Kumar Kombaiya –International Journal On Engineering Technology and Sciences –IJETS.ISSN(P):2241-5381,ISSN(O):Volume 56,February 2018-Published Paper on "**Application of Data Mining Techniques in Early Detection of Breast Cancer**"On ,February-2018.
10. A.Ganesan and Dr.A.Kumar kombaiya ,JETIR –Journal of Emerging Technologies and Innovative Research . ISSN:2349-5162, Volume5,Issue 8,ISSN(P):2319-8346 Published Paper on "**Comprehensive study of Classification model for intrusion detection systems: Review** "on August 2018.
11. V.Shilpa and Dr.A.Kumar Kombaiya –International Journal of Scientific Research and Development-ISSN:2321-0613.Volume 6.Issue 9,"**Monitoring Air Pollution Tree-Based Routing Scheme For Wireless Sensor Nodes**"on November 2018.
12. Prema.K and Dr.A.Kumar Kombaiya,Published online in journal of computational Information Systems ISSN:1553-9105.Survey on"**An Analytical Investication and Feature selection methods on Microarray Leukemia Data**"on Jan 2019 in UGC journal.
13. F. Leena Vinmalar and Dr. A. Kumar Kombaiya, Published in The International Conference on Artificial Intelligence, Smart Grid and Smart City Applications,(ASGSC 2019)," **A Morphological and Hybrid Moth Flame Optimization Based Feed Forward Neural Network for Skin Cancer Detection**", January 2019.
14. B.Vidhya and Dr.A.Kumar Kombaiya,IJRAR,Volume6,Issue 1,ISSN(E):2348-1269,ISSN(P):2349-5138."**A Perspective Study on Acute Lymphoblastic Leukemia Detection using Various Image Processing Techniques**-on March 2019.
15. F.Leena vinmalar and Dr.A.Kumar Kombaiya –Oxidation Communication-Book2,ISSN:0209-4541,Published Paper on "**Lung Cancer Diagnosis Using Hybrid Dragonfly Optimization and Radial Basis Neural Network Classification**"- June 2019.
16. Prema.K and Dr.A.Kumar Kombaiya ,Published in The International journal of analytical and experimental model analysis,UGC CARE Approved group-A Journal ISO:7021-2008 Certified IJEMA journal Volume XI,Issue X,October ISSN NO:0886-9367, "**Microarray based Cancer Classification using Clustering Algorithm for Feature Selection Approach**"on October 2019.

ANNEXURE II

1. Prema.K, International Conference on Computing Intelligence and Applications ICCIA 2K17 ISBN 978-1-941505-98-4 Presented paper on "**Artificial Neural Networks: Applications, Issues & Trends in Data Mining**" on 29th July 2017 in Dr.SNS Rajalakshmi College of Arts and Science.
2. Prema.K and Dr.A.Kumar Kombaiya, 4th International Conference on Computer Applications & Information Technology CAIT 2017 ISBN 978-81-932858-9-3 Published a Paper on "**A Survey on Image Compression Techniques**" on 8 August 2017 in Hindusthan College of Arts and Science.
3. Prema.K and Dr.A.Kumar Kombaiya, International Conference on Intelligent Computing and Technology ICICT 2017 Presented paper "**A Survey on use of Data Mining Methods Techniques and Applications**" on 8 December 2017 in Sri Ramakrishna College of Arts and Science.
4. Prema.K and Dr.A.Kumar Kombaiya, International Conference on Research and Innovations in computational Intelligence(RICI 2019) presented paper "**Leukemia data analysis using an K-nearest neighbour classifier with Map reduce framework in Microarray**" on 4 February 2019 in Sri Ramakrishna College of Arts Science for Women.

F. Leena Vinmalar
Ph. D Scholar
Department Of Computer Science
Chikkanna Government Arts College
Tiruppur, Tamil Nadu, India- 641 602.
Mobile: 91 8220806312
Mail:anuleena7@gmail.com
Date of Birth: 17/05/1990

Qualifications:

Degree	Year of Passing	Institution
B. Sc (Computer Science)	2010	Tirupurkumaran college for women-Tirupur
M. B. A(Finance/HR)	2012	Vivekanandha Institute Of Information And Management Studies-Tiruchencode.
M. SC(Computer Science)	2014	St'Joseph College For Women –Tirupur
B. ed	2015	St.Peter's College Of Education-Karumathampatti.
M. Phil (Computer Science)	2016	Chikkanna Government Arts College-Tirupur.

Teaching Experience:

Period	Position	Employer
2016-2017	Assistant Professor	Park's College-Tiruppur.

Research Publications,Conferences,Workshops,Seminars:

S.NO	DD/MM/YY	EVENT TYPE	TITLE	INSTITUTION	NO. OF. DAYS
1.	04/02/2016-05/02/2016	National seminar	Research Trends And challenges In It (NSRTCIT'16)	Kongu Arts And Science College-Erode	2 Days
2.	18/03/2016	National Conference (3rd National Conference on "Innovative Intelligence in Computer Technology-NCICT2K'16)	(A Novel Cancer Gene search Model and Classification Using GPSO-BPNN)	ADITHYA Institute Of Technology-Coimbatore	1 Day
3.	07/10/2016	National Seminar	Research Publication and Evaluation: writing Research Papers,Citation Analysis and Plagiarism.	ChikkannaGovt Arts College-Tirupur	1 Day
4.	24/08/2017	International Seminar	Applications and Challenges in Data Mining	Bharathidasan College of Arts and Science-Coimbatore	1 Day
5.	30/04/2018	International Conference(Current Research in Engineering Science and Technology-ICCREST-2018)	Cancer Gene Expression Model and Classification sing Fuzzy Clustering	Jayaram College Of Engineering and Technology-Trichy.	1 Day
6.	30/08/2018	National Level FDP	Underwater Computing & Communications	Dr.G.R. Damodaran College of Science(Autonomos)-Coimbatore	1 Day
7.	24/01/2019-25/01/2019	National Level Workshop	Virtual Reality Development Using nity-3D Game Engine & C#	Sri Ramakrishna college of Arts & Science	2 Days
8.	30/01/2019-01/02/2019	FDP	Machine Learning	Bharathiar University-Coimbatore	2 Days
9.	14/02/2019-15/02/2019	National Workshop	Data Analytics and its Applications in Biotechnology	Dr .N .G . P Arts And Science College (Autonomos)-Coimbatore	2 Days
10.	22/02/2019	National Seminar	Data Analytics	SRI VASAVI COLLEGE ,ERODE	1 Day
11.	28/02/2019	National Workshop	Medical Image Processing using MATLAB/GUI:A Research Perspective	KONGU Engineering College	1 Day