Research Article

Biomarker Discovery based on Hybrid Firefly Optimization Algorithm and Hybrid Adaboost on ANNs Classifier

[1]M. Divyavani [2]Dr. G.Kalpana

[1]Ph.D. Research Scholar, [2]Associate Professor
[12]Department of Computer Science
[12]Sri Ramakrishna College of Arts & Science for Women, [12]Coimbatore
[1]divicse07@gmail.com [2]kalpanacs @srcw.ac.in

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 May 2021

Abstract:
Biomarker discovery is one of the biggest challenges in cancer research. In this paper, a new approach based on FFF and adaptive boosting on ANN classifier is proposed for finding genes that can classify the group of cancer correctly. In this approach, FFF as firefly wrapper-based feature selection is used to perform gene selection and Adaboost on ANN classifier with 7 cross fold cross validation is adopted as the classifier. The proposed approach is tested on five benchmark microarray gene expression profiles namely, Colon, SBRCT, Leukemia 1, Leukemia 2 and Lung. The experimental results show that our proposed method can select the most informative gene subsets by reducing the dimension of the data set and improve classification accuracy when compared to the existing algorithm. Our proposed algorithm shows the best classification accuracy compared with the FFF-SVM.

Keywords: Artificial Neural Network*; Adaboost;* Cancer Classification*;* Gene Selection*;* Firefly Optimization Algorithm; Fisher Score Based Filtering Algorithm.

I. INTRODUCTION

DNA microarray technology allows the investigators to investigate the expression levels of thousands of genes concurrently. The DNA microarray data typically contain small sample size and thousands of genes that most of them are proved to be uninformative and redundant. Hence, finding a large subset of important genes in microarray data in order to improve classification accuracy is an essential issue in gene expression analysis. Cancer is a frightening case for every human being that a thorough understanding of the classification of this disease is imperious. The traditional methods for diagnosing cancer highly depends on the doctor's expertise and their visual inspections. Human beings naturally commit errors due to their limitations, but human beings can without problems understand patterns.There is a substantial quantity of data with low quality and redundant information that even for medical experts, may be tough to acquire the accurate classification. Computer aided diagnostic tools are intended to help physicians for the accuracy of classification improvement [1,2].

Certain approaches and performances have been used for cancer classification. Specifically, N.N. Mohad Hasri et al. [3] used support vector machine (SVM) with recursive feature elimination (RFE) method for cancer classification. Xin Sun et.al.[4], suggested a robust and stable feature selection through the integrating ranking methods (IRM) and wrapper method in genetic data classification. The trials of the proposed approach on five cancerous microarray datasets include colon cancer. In this paper (Thanh Nguyen et al.) [5] presented a new technique for the selection of features based on a Modification of the Analytic Hierarchy Process (MAHP). Which is used different classifiers covering linear discriminant analysis, probabilistic neural network, KNN, multilayer perceptron, and SVM. Maolong Xi et al. [6] was proposed Binary Quantum-behaved Particle Swam Optimization (BQPSO) for feature selection from cancer datasets. The author uses five microarrays along with colon cancer. Aalaei et al. [7] applied for GA-based classifier with the ANN.

Several supervised learning algorithms such as linear discriminant analysis [8], Support Vector Machine (SVM) [9], decision tree (C4.5) [10], K-nearest neighbor [10], artificial neural network (ANN)

[4,5] have been used as fitness function of optimization algorithms for microarray data classification, successfully. In this paper (dong Ling Tong, 2011) [11] was developed a hybrid genetic algorithm (GA) - neural network model for feature selection on unpreprocessed microarray data. The fitness value GA is based on an accuracy of standard feed-forward artificial neural network (ANN). The goal of the genetic algorithm-neural network algorithm is to select highly informative genes by the calculation of the both GA fitness function and the ANN weights simultaneously. In (Li-Yeh Chuang, 2011),[12] Taguchi-GA and correlation-based feature selection is used as a hybrid method, and the K-nearest neighbor (K-NN) served as a classifier and then in this paper (Li-Yeh Chuang *et al*., 2011) [13] additionally based on Taguchi binary particle swarm optimization (PSO) conducted by the same authors. In the paper (Bing Liu, 2004) [14] a combinational feature selection method with ensemble neural networks is used for classification.

In the Paper (Emmanuel Martineza, 2010) [15] proposed an algorithm based on swarm intelligence feature selection method in which used for the initialization and update of only a subset of particles have happened in the swarm. The most frequent genes are evaluated by the GA/SVM again to obtain the most final relevant gene subset. In [16] a hybrid method based on information gain and GAs are proposed for gene selection in microarray data sets. The K-NN method with leave-one-out cross validation served as a classifier for evaluating the fitness function of this hybrid algorithm. In Jenny Önskog, 2011),[17] has developed the classification performance of five normalization methods and three gene selection methods as *t*-test, relief, paired distance, and eight machine learning methods as a decision tree with Gini index and information gain criterion, SVM classifier with different kernels and also neural network are compared with each other.

In the paper (Xiaosheng Wang, 2011),[18] uses single genes to create classification models and identified the most powerful genes for class discrimination. These kinds of classifiers, include diagonal LDA, K-NN, support vector machine and random forest and it can construct simple rules for cancer prediction by these single genes. In Makoto Takahashi a, 2010,[19] an unpaired *t*-tests with one of the supervised classifiers, ANNs was applied to schizophrenia gene expression data sets. In this paper (Khan javed, June 2001),[20] has developed by a method for classifying cancers using ANNs on small, round blue cell tumors as a model. *T*-test and PCA are used to reduction dimensionality of data sets. In (Nikhil R Pal, 2007),[21] a multilayer network with online gene selection ability and relational fuzzy clustering was used to identify a small set of biomarkers for accurate classification.

Finally, this study used genetic programming GP for cancer types classification by S.A.Ludwig et al. [22] presented a fuzzy decision tree algorithm in classifying gene expression data. The literature investigated supplied information, which helped for the evaluation of end result with the proposed method. Artificial Neural Network (ANNs) is a computer studying approach. The most prevalent type of ANNs is called feed-forward neural networks [23, 24-29]. Artificial Neural Networks (ANNs) is similarly valued in estimating complex target functions with target is determined according to the network architecture minus the linearity and conventions of limitation. These networks can observe without difficulty due to the overall performance of enhancing the pc systems, with complexity and a variety of areas to hire and assessment models.

Consequently, neural networks utilized the regression and discriminated analysis for conventional statistical methods in the current period. The illustration of a neural network designed as influential and adaptive nonlinear equation. They can provide information as regards to the multifaceted operational connections between the input and output data [30]. Because of these attributes, as soon as the output nodes accommodate actual values, they are successful of revolving into a regressor. Furthermore, a classifier can be described if the outputs are an integer or absolute. In overall, the function resolution techniques produce authentic values as outputs, and absolute values return via the classification algorithms. Hence, this study utilized neural networks for the estimation of the diagnostic of cancer classification from microarray gene expression profiles.

The firefly method is natural-inspired global optimization algorithm inspired by the behavior and the main patterns for flashing light of firefly insects. Hence, it simulates the attraction behavior of fireflies. Firefly insects uses their flashing light pattern to attract other fireflies i.e. firefly from the opposite sex. The firefly optimization algorithm has been used for the feature selection on several small medical datasets. Compared with GA and PSO, BBHA has some attractive characteristics such as fast convergence rate and only one parameter for setting, as an ensemble algorithm, Adaboost on ANN tends to achieve high prediction accuracy by reinforcing training on misclassified samples [31]. In this study we proposed work aims to remove uninformative and redundant genes, increase classification accuracy and discovery of biomarkers, the FFF based-ANN-Adaboost approach which uses FFF for gene selection and Adaboost on ANN for samples classification is done.

First, Euclidean distance measure and K-nearest neighbor is used for measuring the important genes in the microarray. Then the topmost weighted genes are grouped and a new feature subset is produced. Thereafter FFF is employed to select informative genes from new feature subsets. In the process of gene selection, Adaboost on ANN with 7-fold cross validation is used to evaluate the selected gene subsets. In the next step, the most frequent genes (biomarkers) are selected and finally, ANN classifier is applied for weight and bias for reducing local minima and finding the relation between these biomarkers. The rest of the paper is organized as follows. Section I comprises of an introductory section about the experiment with a review of the related study. Section II discusses the existing methods that includes fisher score-based filtering method, FFF- firefly optimization algorithm, Adaboost on ANN classifier and concludes with the hybrid proposed approach. Section III presents the experimental results on five microarray gene expression profiles. Finally, section IV summarizes the work with its findings.

In the proposed approach the hybrid firefly wrapper-based algorithm is used for feature selection method and each weighting of gene expression is determined by adaptive boost on ANN classifier's accuracy. The 7-fold cross validation classification accuracy on the gene expression in the training and evaluation samples is the evaluation criteria. The group of gene subset with the highest 7-CV classification accuracy is considered as the gene subset. After the selection of the most frequent genes, it can be used for discrimination of blind test data to depict the response of evaluation hybrid system on these kinds of data. The main aim of the research is to increase the accuracy of classification problems by selecting the best parameters of the classifier in the training and testing phase without using any trial and errors of users. Hence, the usage of a appropriate mixture of optimization algorithms for feature selection and additionally choosing perfect classifier can enhance the classification results.

II. METHODS AND MATERIALS

In this section, we give a detailed description of our proposed algorithm. We can implement both of these algorithms in hybrid form to benefit the useful advantages of both of them and their problems too. In this paper, Adaboost (Adaptive Boosting) on basic ANNs is used as a classifier and fitness function-based hybrid firefly optimization algorithm and also using fisher score filter method for the data normalization from the original dataset after that we can applied for feature selection and then finally applied for classifier algorithm.

A. Feature Selection

Feature selection is a technique to minimize the variety of attributes and selects a subset from the accurate features. In data pre-processing the feature selection regularly utilized to classify appropriate feature that are often unknown earlier than and take away the noise and irrelevant or redundant features, which have zero significance in the classification task. The progress of classification accuracy is the main goals feature selection. This section illustrates the feature selection model and classification model. In this study we applied for Firefly method is natural-inspired global optimization algorithm inspired by the behavior and the main patterns for flashing light of firefly insects. Hence, it simulates the attraction behavior of fireflies [33]. Firefly insects uses their flashing light pattern to attract other fireflies i.e. firefly from the opposite sex.

However, firefly feature selection methods depend on three important rules: first rule, assumption that all fireflies are attracted to each other fireflies, based on their sex. Second rule, the attractiveness is related to their brightness. Therefore, their attractiveness will increase as the distance between they decrease. Thus, the less bright fireflies will always move towards the brighter ones. Third rule, a firefly's brightness is affected or determined by the objective function's form i.e. fitness function. Firefly feature selection technique is a population-based metaheuristic algorithm the place every firefly represents a feasible solution (Biomarker genes) in the search space. The main behavior of artificial firefly feature selection algorithm is presented and illustrated in our previous proposed algorithm FF-SVM [32].

Furthermore, in this section we will present briefly the main function and procedure of firefly feature selection method. Firefly algorithm considers two issues they are; the variation of the brightness intensity, as well as how attractiveness is formulated. The attractiveness determined by brightness in the standard Firefly bio-inspired method, which is mainly related with the objective function $f(\mathbf{x})$. Thus, the brightness of specific firefly I at specific location x can be formulated as ($I(\mathbf{x}) \propto f(\mathbf{x})$). The following ALGORITHM 1: FFF- Firefly wrapper-based feature selection is described.

```
Input:
Microarray dataset
Size of firefly i.e number of features
Define light absorption coefficient $\gamma$  [10.01, 100]
Define randomization parameter $\alpha =1$
Define attractiveness parameter $\beta0=1$
Set maximum number of iterations: maxGeneration=25
Output:
                                        The best firefly and its fitness

Algorithm:
Apply Fisher score filter method
Using filtered dataset;
Generate initial population of n fireflies randomly xi, i= 1,2,3,…..n;
Evaluate the fitness, each firefly using objective function f(x);
While (t<MaxGeneration);
For i=1 to n;
      For j=1 to n;
            If (f (xi) <f (xj));
            Move firefly I towards j using
            Calculate the distance r (applied for Euclidean distance measure formula
                                    and K-nearest   neighbour)
            Calculate the new position $x_i$ of the firefly i
            Update the fitness of firefly i
            End if;
            Evaluate new solutions and update light intensity;
            End for j;
         End for i;
         End while;
         Rank the fireflies and find best firefly;
```

ALGORITHM 1. FFF (FIREFLY WRAPPER-BASED FEATURE SELECTION)-PHASE I

B. Adaboost

Adaboost is an ensemble algorithm that works by making a highly accurate classifier by merging many relatively weak and inaccurate classifiers. Adaboost classifier is the concept of converting a weak learner to a strong learner. It is the process of combining all weak learners to form a single strong rule. Each time when the base learning algorithm is applied it generates weak prediction rules through an iteration process. After conducting several iterations, the boosting algorithm combines all weak rules to form a single strong prediction rule. There are the three steps used for Adaboost algorithm: Step 1: The base learner is applied to distribute and assign equal weight to each observation. Step 2: If any prediction error is observed then a higher attention is paid for observations having error. Now, the next base learning algorithm is applied. Step 3: Step 2 is repeated until higher accuracy is achieved by the base learning algorithm. At the end, all the output weak learners are clubbed to form a strong learner. Boosting concentrates more on the misclassified examples or to the examples that have higher prediction errors.

C. *The Proposed Hybrid FFF-ANN-Adaboost Approach*

ANNs is a branch of computational intelligence that exploits a variety of optimization and implement with layers of computing nodes that have incredible processing data features. ANN is an information processing system that became knowledge from human brain. It does data processing by provided that insignificant processor that are parallel interconnected with each other to form a network to solve a problem. Neural networks which are used to implement complex functions in several grounds, including pattern recognition, identification, classification, speech and image processing, and control systems. Subsequently training the neural network, each particular input has a specific response. A neural network contains of mechanisms as layers and weights. Network behavior is associated to the connections between its members. In overall, the neural network has three layers of neurons such as an input layer, hidden layers, and output layer.

The input layer takes raw data and feature vectors. Performance of the hidden layers is resolute by inputs and weighted vectors between input and hidden layers. The weights between input and hidden units consume to be strong-minded when a hidden unit is been active. The output layers have determined by the weights between the hidden and output units. In other types of multi-layer perceptron networks or feed-forward networks, each layer may be determined by their parameter matrices and the network can form by a combination of nonlinear operators. The objective is finding and estimation of the representing their function between input and output spaces. Estimation of appropriate network is based on a minimization of the error between the desired output and network's output and each layer, activation functions can be nonlinear in each layer and additionally can be distinctive from every other. In these networks, there are two kinds of weight matrices, such as an intermediate layer or hidden layers and output layer weight matrix. These matrixes’ sizes rely on the range of neurons in hidden layers and output layer's neurons. So how the network works is as follows

$$u(n) = W^h \times x(n), \; h(n) = \varnothing\,(u(n))$$

$$v(n) = W^y \times h(n), \; y(n) = \varphi\,(v(n))$$

In summary, we can write:

$$y\,(n) = \varphi\left(W^y \times \varnothing\left(W^h \times x(n)\right)\right)$$

Training neural networks means choosing the best model of network by the finest parameters such as weights, number of neurons based on the cost function. The task of pattern classification in ANN is to assign an input pattern as gene expression profile represented by feature vector to one of the presented classes such as normal or cancer. After providing the best network based on feature vectors and parameters, our model can be able to predict the class of new data based on training for each dataset [34].

ANN Algorithm: General steps

Input: dataset D, learning rate, network. Output: a trained neural network. Step 1: receive the input. Step 2: weight the input. Each input sent to network must be weighted i.e. multiplied by some random value between -1 and +1. Step 3: sum all the weighted input. Step 4: generate output. The output of network is produced by passing that sum through the activation function.

Procedure of the ANN

1. Assign random weights to all the linkages to start the algorithm.

2. Using the inputs and the (input → hidden node) linkages find the activation of hidden nodes.
3. Using the activation rate of hidden nodes and linkages to output, find the activation rate of output nodes.
4. Find the error rate at the output node and recalibrate all the linkages between hidden nodes and output nodes.
5. Using the weights and error found at output nodes, cascade down the error to hidden nodes.
6. Recalibrate the weights between hidden node and the input nodes.
7. Repeat the process till the convergence's criterion is met
8. Using the final linkage weight score the activation rate of the output nodes. The detailed description of our proposed algorithm describes in ALGORITHM 2: Proposed Hybrid ANN-Adaboost approach is as follows:

```
Input:
Microarray dataset D, learning rate, network. Output: a trained neural network.
Size of the genes i.e size =25
Set selected genes from each dataset = 1000
Define the weighting value for each feature
Set maximum number of iterations: 100 MaxGeneration =25
Output: Prediction value for large number of gene set

Algorithm:
Apply fisher score filter method
Using filtered dataset;
Generate initial population of n weights of each input randomly xi, i= 1,2,3,4,…n;
TS: training set, TS=ui (i=1,2,…,n), labels vi εV
Assign TS sample (u1,v1),…,(un, vn); ui ε U, vi ε {-1, +1}
Initialize the weights of Di(i) = 1/N, i=1…N
    For t= 1,…T
Train weak learner using distribution Di
Get weak hypothesis ht;
Update distribution Di:
     Next t that, t+1
Output the final hypothesis: H(u);
W1← Weight vector for hidden layer
W2← Weight vector for output layer
NumberCorrect = 0
for I < Number of Training Iterations do
        for j< Size ( Train set ) do
                Input ← Trainset (j)
                HiddenOutput ← f(Bias; W1, Input)
                Output ← g(Bias; W2, HiddenOutput)
                Prediction ← argmax(Output)
                If Prediction = Train label(j) then
                        NumberCorrect+=1
                end if
                D= weighting distance calculated (Applied by Euclidean distance measure)
                K= nearest neighbour ← Input ← hiddenOuput (Applied by K-nearest neighbour)
                End if
                $\text{delta}_1$ ← (Output—trainlabel (j) )* ( 1—$\text{Output}^2$)
                $\text{delta}_2$ ← (W2*$\text{delta}_1$ ) * ( 1—$\text{HiddenOutput}^2$)
                W1← W1—alpha* ( Input * $\text{delta'}_2$ )
                W2← W2—alpha* ( hiddenOutput * $\text{delta'}_1$ )
        End for

        if W1>0 then
                W1← 1
        else
                W1← 0
        end if
```

ALGORITHM 2. THE PROPOSED HYBRID FFF-ANN-ADABOOST APPROACH-PHASE I

In this study, we developed a new hybrid gene selection algorithm to select the most informative genes that cause cancer using Microarray gene expression profile. In the subsequent section we will describe our proposed algorithm and the used fitness function and weight function also applied Euclidean distance measure and KNN for finding distance between two neurons. The proposed algorithm consists of three main phases: filtering phase, gene selections phase and classification phase. The proposed method is called Hybrid approach for Basic ANNs classifier with Adaboost. The detailed steps for hybrid approach ANNs-Adaboost are as follows:

1) Phase One: Filtering Phase

In this phase, fisher-score filter method was applied to reduce data dimensionality and lower the search space complexity and also to be remove the noise data, redundant data and irrelevant features. Therefore, the most statistically relative genes are selected and used as input to the second phase (gene selection phase). Different number of relative genes is selected for each dataset. Hence, we applied fisher-score filter method to select 100, 200, 300, 400, 500 and 600, 700, 800,900, 1000 features for all the datasets. Therefore, for each dataset we obtained five different filtered datasets. Then, the filtered datasets are classified using Adaboost on ANN classifier, the accuracy is evaluated using incremental weighting value. The datasets that has the best accuracy with the minimum number of selected features are then moved to the next phase. *Fisher Score* F_i [40]. It is mostly useful in gene selection as a filter. The fisher score value of each gene represents its relevance to the dataset; a higher fisher score means that the gene contributes more information. This data assistances to measure the grade of separability of the classes through a given gene $g_{i.}$

2) Phase Two: Gene Selection Phase

In this phase, we applied the firefly wrapper feature selection method which is based on an evolutionary bio-inspired algorithm. The aim of this phase is to find the more predictive genes that maximize the ANN classification performance using the filtered dataset resulted from phase one. The basic methodology of firefly feature selection method is to compare each firefly on the swarm to every other firefly and based on the firefly’s brightness, which is represent the fitness value in our problem, then only one best firefly will be chosen and returned as solution. The steps this phase can be described as follows:

Step 1 (Initialized Firefly Population): The firefly feature selection method initiate the swarm by create a population of n fireflies x_i, i = 1,2,3,.....,n where n is the swarm’s size. Initially, the fireflies take their position randomly in the search space. Each firefly x_i in the population represent a set of predefined number of features i.e. the solution to biomarker gene discovery problem.

Step 2 (Fitness and Objective Function Calculation): The second step, calculate the fitness function $f(x_i)$ of the initial swarm which is represent the light intensity (detailed in phase 3). A sample of the initial population is representing Fig 3, x_i represent a firefly i.e. one possible solution with its fitness f(xi). Each firefly contains D number of biomarker genes.

Step 3 (Finding the Best Firefly in Each Iteration): In the third step, our proposed algorithm search for the best firefly in each iteration that maximize the performance of classification accuracy while keep the maximum number of selected predictive genes. The purpose of this step is to compare each firefly in the population with every other firefly in the same iteration. The proposed algorithm starts the comparison by given a fixed number of generation (each generation represent an iteration). If the fitness of the firefly *i* is less than the fitness of the firefly *j*, then firefly *i* will move toward firefly *j*. equation 4 represents the movement of the firefly. Since the position of the firefly *i* is updated, its fitness must also update. Then the algorithm follows the same procedure for subsequent iteration.

Step 4 (Ranking and Return the Best Firefly): In this step, our proposed algorithm wills ranking the resulted best fireflies (i.e. candidate solutions) in each iteration. After the comparison, it returns the best firefly in search space.

3) Phase Three: Classification Phase

This phase calculates the fitness function for the hybrid FFF- ANN-Adaboost. The fitness function aims to maximize classification accuracy performance while keeping small number of selected predictive biomarker genes. The classification accuracy is generated using incremental weighting value. Applied K-nearest neighbour and Euclidian distance measure for finding distance between two neurons (node). Next is the pseudo code for proposed hybrid FFF-ANN-Adaboost as classifier algorithm.

III. EXPERIMENTAL RESULT AND ANALYSIS

To evaluate the effectiveness of our proposed method, the experiments are done on five bench mark microarrays (leukemia_1, leukemia_2, SRBCT, lung, colon), which are obtained from [21-25].

A. Microarray Gene Expression Profiles

Five-bench mark microarray cancer gene expression dataset were used to evaluate the proposed algorithm FFF-ANN. In this section, we introduce the gene expression datasets which were used in this paper and also propose the modified hybrid algorithm. Five datasets are used to test our proposed algorithm. The first data include 72 samples in two type of classes as acute lymphoblastic leukemia _2 (ALL) and acute myeloid leukemia (AML). The original size of genes in this dataset is 7129. These two categories of cancer are quite similar at the microscopic level and have a same behavior over the years. This dataset is generated by Golub in (Golub T, 1999) using 25 AML and 47 ALL samples. The second are generated by (Alon U, 1999) for colon cancer categories. These data have 22 samples for normal class and 40 samples for tumor class. The size of genes in this dataset is 2000. The third data include 10 normal samples in two class of lung cancer samples 86 are placed in their class and original size of genes in these data is 7129. The fourth dataset SRBCT [12] which data is contained 29 EWS and 18 NB, 11 BL and 25 RMS then it is used four different classes they are placed each of them their class. The original size of features or genes in these data is 2308. The last dataset leukemia_1 is containing three types of classes and 28 AML, 24 ALL, and 20 MLL. The original size of genes in this dataset is 7129. The evaluated datasets are of binary and multi-class namely, leukemia_2 dataset, SRBCT dataset, Lung cancer dataset, leukemia_1 dataset and colon cancer dataset. Detailed description of these datasets presented in Table 1.

TABLE 1. MICROARRAY GENE EXPRESSION PROFILES DESCRIPTION

Dataset	**Tissue**	**Sample**	**No. of Classes**	**Samples Per each Classes**	**Classes**	**Genes**
Armstrong [2002] [35]	Leukemia_2	72	3	28, 24, 20	AML, ALL, MLL	7129
Khan [2001] [36]	SRBCT	83	4	29,18,11,25	EWS, NB, BL, RMS	2308
Beer [2002] [37]	Lung Cancer	96	2	86,10	Cancer, Normal	7129
Golub [1999] [38]	Leukemia_1	72	2	25,47	AML, ALL	7129
Alon [1999] [39]	Colon Cancer	62	2	40,22	Cancer, Normal	2000

In the following discussion, we introduce the proposed algorithm which is used on gene expression profiles. ANN and FFF are two optimization algorithms which have many advantages in these kinds of problems. They are computational optimization method that search all part of the solution space with a different kind of solution or a group of feature subsets to find the best answer in each iteration. In ANN, the searching process only needs to determine the input value of the Activation function (hidden layer) at different weight points and also, added Ada boosting algorithm which is used reduced the bias and variance. The most important ANN with combination of Ada boosting are gives best solutions.

In the following, 10% of data must belong randomly as a blind test data, and also remaining 90% of data can be entering to training and evaluation phase of the algorithm by 7-fold cross validation. The value of parameters such as, size of population, individual length, size of the genes, number of iterations, number of features, incremental coefficient (W), selected genes per each dataset, training factors (learning factors), and maximum velocity is mentioned in Table 2. We have measured the performance our model through the following best parameters is accuracy, sensitivity, specificity those are mentioned in Table 3.

TABLE 2. PARAMETERS IN FFF-ANN

S.No	**FFF-ANN parameters**	**Leukemia_1**	**Leukemia_2**	**SRBCT**	**Lung**	**Colon**

1	Population	25,47	28, 24, 20	29,18,11,25	86,10	40,22
2	Individual length	72	72	83	96	62
3	Gene size	25	25	25	25	25
4	Number of class	2	2	2	2	2
5	Selected genes per each dataset	1000	1000	1000	1000	1000
6	Number of features	25	25	25	25	25
7	Number of iterations	100	100	100	100	100

In addition of creating initial position (X_{id}), should be determined randomly in the population. This stage is related to making the initial population, at first the population with N genes create randomly. The length of genes can be explained as, adding number of features which has been selected based on statistical method and 10 additional genes which have been used for determination of optimum parameters of classifier by hybrid algorithm. Primary random and binary initialization are taken place first, in such a way that 1 shows the existence of the feature in training system and 0 is meaning of not existing of that feature. Now, each gene is a word of bits in two main parts. First part is equal to feature dimensions size (segment 1), and the second part is used for determining and designing classifier parameters. The fitness values for all particles have to be calculated in order to determine functionality of each particle, which is so-called validation of particles. It is important to note that in genetic operators, there is no discussion in speed changes or the best memory of offspring; hence, we have determined the best memory of offspring based on the best memory of parents which have the best fitness value. After this step, this is the time for running hybrid FFF, from the solutions which are presented by the ANN with Ada boost. At the end of the progress, the best features with the best parameters of classifier are selected, so we have applied these features and parameters to blind test that has no interference in the training and validation phase at all. Determine the occurrence frequency of each feature in the whole process. On average, biomarkers that have been repeated >2 times in the best locations are reported.

Finally, the decision tree's rules can be found from the best-extracted features. By applying the proposed algorithm to 5 cancer databases, the amounts of accuracy, sensitivity, precision, specificity were computed. These values are statistical indicators for the evaluation of a binary classification. Our goal is to find the best performance and comparison of this modified algorithm with the others methods. Table 3 shows the result of the applying algorithm to databases. Relative to the number of samples in each database, we select genes (500–1000), and also increase gene size up to 25 then we apply them to a hybrid algorithm. Following an argument, we present the biomarkers which obtained by a hybrid algorithm, then we applied Euclidean distance and KNN also added adaptive boosting for ANN classifier which are achieved by the biomarkers. The results indicate the good performance of our proposed algorithm in finding small subset of features with high accuracy. Furthermore, the results show the good similarities between our biomarkers and the biomarkers that have been showed by existing biomarkers in the existing algorithm. From the results, we can understand that the hybrids algorithm with this classifier have a better result rather than the result which obtained from existing Hybrid FFF-SVM with LOOCV classifier algorithm. Furthermore, we can improve the accuracy of classification by determining its parameters automatically during the feature selection stage with small feature of subsets.

TABLE 3. THE PERFORMANCE RESULTS OF PROPOSED HYBRID FFF-ANN-ADABOOST APPROACH AND EXISTING HYBRID FFF- SVM WITH LOOCV APPROACH

Datasets	**Hybrid FFF-ANN –Adaboost Approach**	**Hybrid FFF-SVM With LOOCV Approach**

	Accuracy	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**
Leukemia_1	92%	94%	93%	91%	92%	91%
Leukemia_2	92%	95%	94%	90%	93%	92%
SRBCT	95%	95%	94%	93%	93%	93%
Lung	95%	95%	94%	92%	91%	92%
Colon	**97**%	97%	97%	**94**%	94%	95%

Classification accuracy of microarrays without any gene selection by Adaboost on ANNs classifier is 92%, 92% ,95%, 95%, 97% for leukemia_1, leukemia_2, SRBCT, lung, colon respectively. The number of selected top genes for all datasets is 25. Classification accuracy of microarrays for selected above 25 top genes on leukemia_1, leukemia_2, SRBCT, lung, colon cancer is 92%, 92% ,95%, 95%, 97% respectively. The highest classification achieved colon dataset by proposed approach compared with existing approach. Because of more emphasizing on presented FFF/adaptive boost on ANN hybrid algorithm, we do further check with more details on these results. Fig 1: shows the overall performance of our proposed model comparison with existing model. Occurrence frequency of genes by hybrid Firefly optimization algorithm/artificial neural network algorithm with 7-fold cross validation for give different microarray gene expression profiles, leukemia _1, leukemia_2 and SRBCT, lung, colon is respectively. For more details, we use a bar graph for individual each dataset showing on discovered biomarkers percentage using proposed model and the results are compared with existing model. In Fig 2: images show the bar graph of cancer in types, leukemia _1, leukemia_2 and SRBCT, lung, colon is respectively. In these bar graph, use orange, sky blue, violet is representing for accuracy, specificity and sensitivity.

Fig 1. Overall Comparison Of The Classification Accuracy Between The Two Approaches Namely, Fig 1.1. Proposed Hybrid FFF- ANN-Adaboost Approach, And Fig 1.2. Existing Hybrid FFF-SVM with LOOCV Approach)

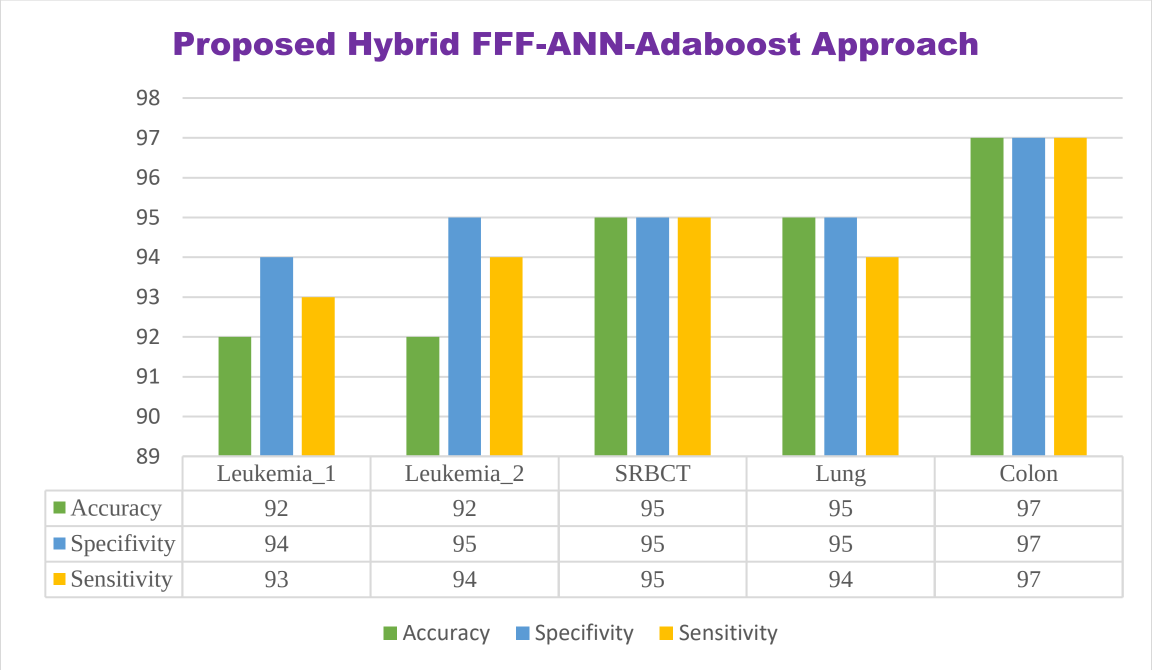
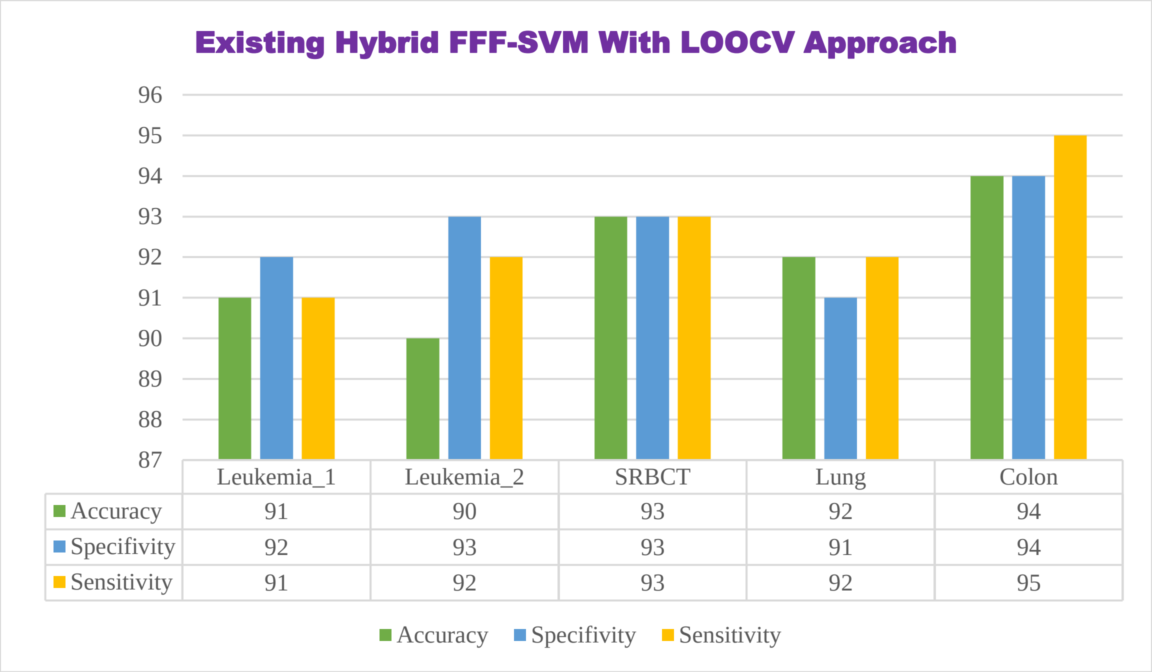


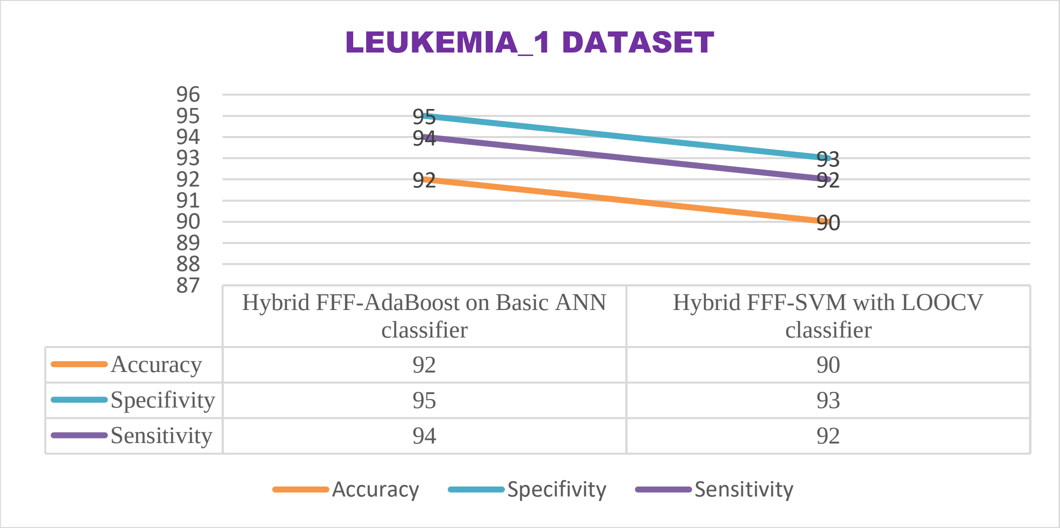
Fig 1.1. Proposed Hybrid FFF- ANN-Adaboost Approach



	Leukemia_1	Leukemia_2	SRBCT	Lung	Colon
🟩 Accuracy	91	90	93	92	94
🟦 Specifivity	92	93	93	91	94
🟨 Sensitivity	91	92	93	92	95

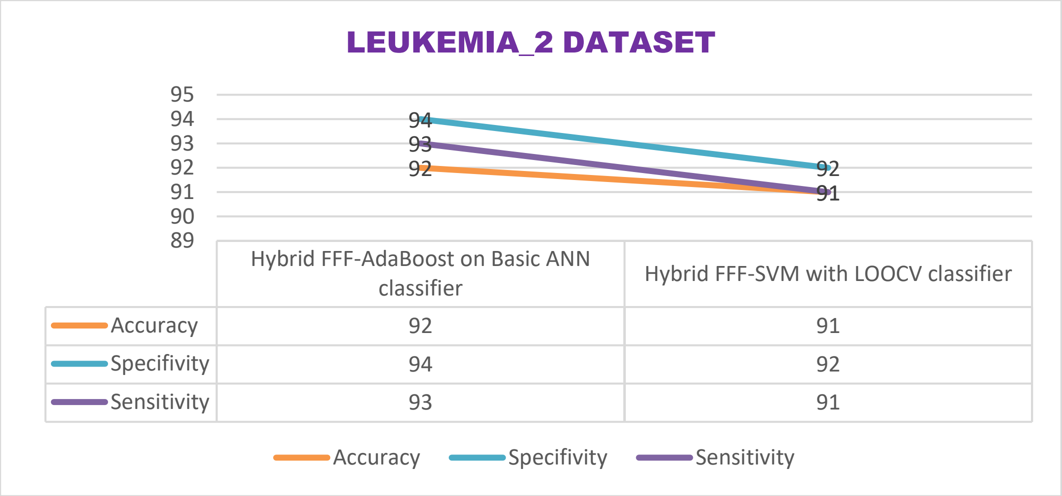
Fig 1.2. Existing Hybrid FFF-SVM with LOOCV approach

Fig 2. Individual Comparison of Proposed Model to Existing Model for five microarray gene expression profiles (2.1) Leukemia_1, (2.2) Leukemia_2, (2.3) SRBCT, (2.4) Lung, (2.5) Colon



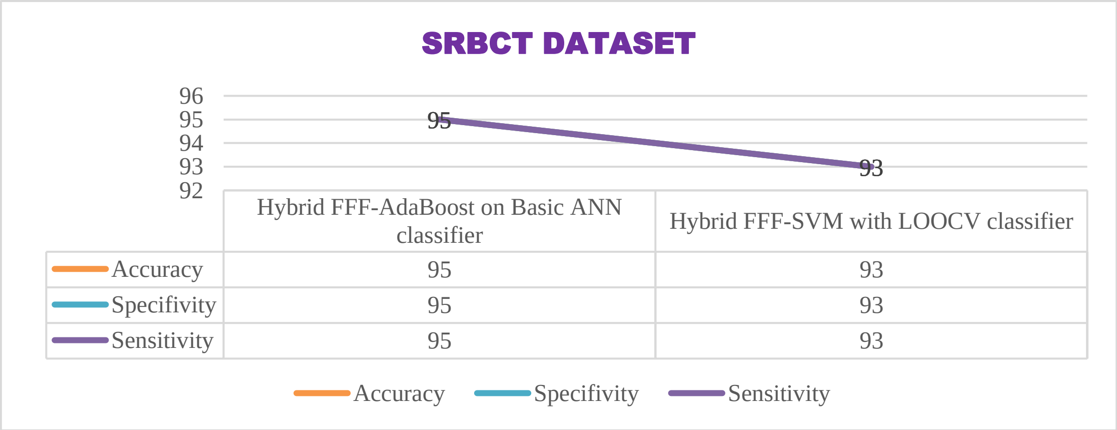
	Hybrid FFF-AdaBoost on Basic ANN classifier	Hybrid FFF-SVM with LOOCV classifier
Accuracy	92	90
Specifivity	95	93
Sensitivity	94	92

Fig 2.1. Leukemia_1 Dataset


LEUKEMIA_2 DATASET

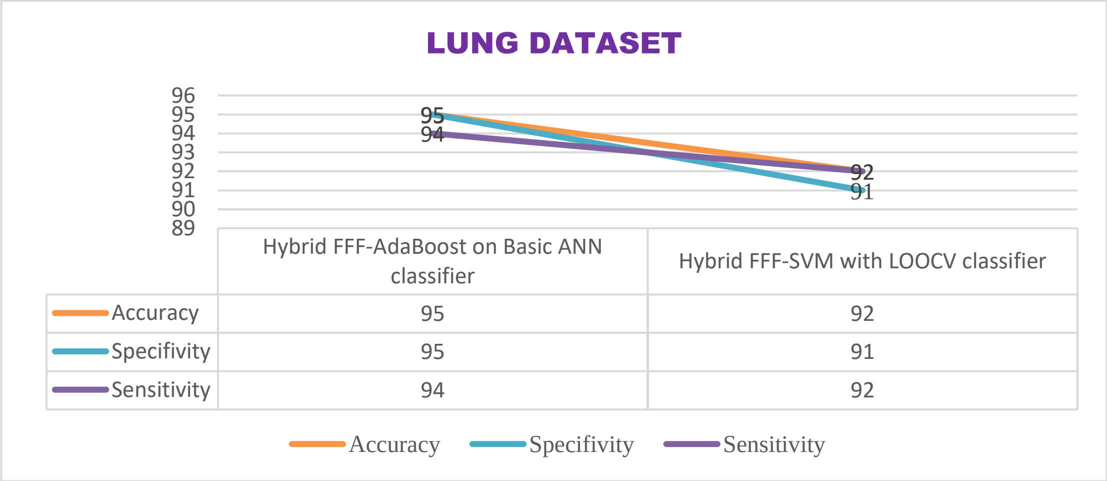
	Hybrid FFF-AdaBoost on Basic ANN classifier	Hybrid FFF-SVM with LOOCV classifier
Accuracy	92	91
Specifivity	94	92
Sensitivity	93	91

Fig 2.2. Leukemia_2 Dataset


SRBCT DATASET

	Hybrid FFF-AdaBoost on Basic ANN classifier	Hybrid FFF-SVM with LOOCV classifier
Accuracy	95	93
Specifivity	95	93
Sensitivity	95	93

Fig 2.3. SRBCT Dataset


LUNG DATASET

	Hybrid FFF-AdaBoost on Basic ANN classifier	Hybrid FFF-SVM with LOOCV classifier
Accuracy	95	92
Specifivity	95	91
Sensitivity	94	92

Fig 2.4. Lung Dataset

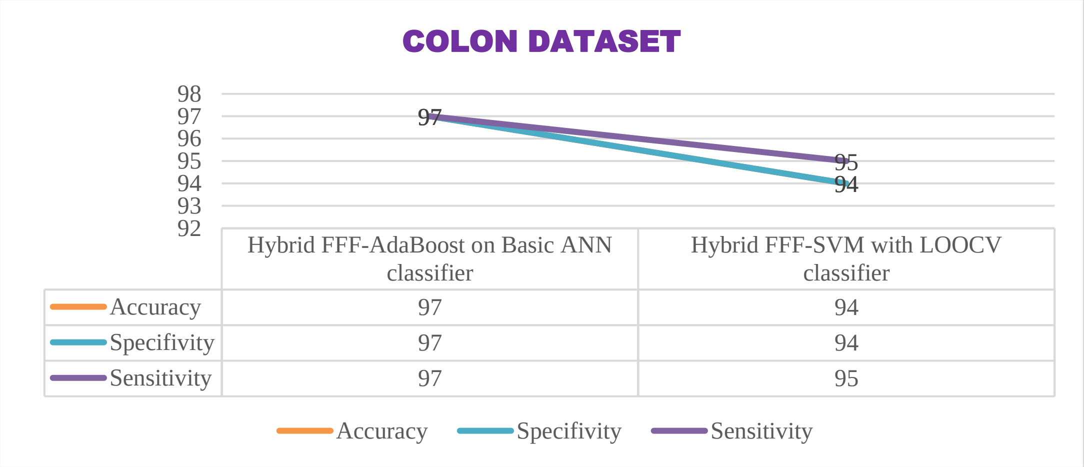


Fig 2.5. Colon Dataset

In the following, we can have some comparison on proposed algorithm with the others works. The first comparison is based on an accuracy of classification which is shown in Table 4. The tables show that overall performance of our algorithm can achieve to best accuracy for the classification problem than existing method. In addition, the procedures of modeling in references papers are test with different parameters by users, but our algorithm was running without any user's interference and any trial and errors.

TABLE 4. OVERALL COMPARISON ACCURACY OF PROPOSED MODEL TO EXISTING MODEL

	Model	**Leukemia_1**	**Leukemia_2**	**SRBCT**	**Lung**	**Colon**
Accuracy	Proposed Model	92	92	95	95	**97**
	Existing Model	91	90	93	92	**94**

IV. CONCLUSION

In this paper, we used hybrid combination of firefly optimization algorithm and Adaptive boost on ANN classifier and determining the classifier's parameters such as number of layers and number of neurons in each layer. The main comparison is based on accuracy of classification that is, shown in Table 4. In this proposed model is obtained a good result with this algorithm. The best accuracy has achieved for each dataset, leukemia_1, leukemia_2, SRBCT, lung, colon, respectively. This result is better than the individual use of Hybrid Firefly optimization algorithm and adaptive boost on ANN also the ability of algorithm in determining the training parameters and small feature subsets in databases perfectly to predict large number genes and which is not predictable from the existing methods. It can only predict small number of genes with high accuracy. In this experiment we have achieved the best accuracy level for predicting large number of genes and measured model performance in best parameters namely sensitivity and specificity.

REFERENCES

[1]. S. Hengpraprohm, P. Chongstitvatana, "Feature Selection by Weighted-SNR for cancer Microarray Data Classification," International Journal of Innovative Computing, Information and Control, 2008.

[2]. X. Zhang, "Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis," European Journal of Human Genetics, 2005.

[3]. J.M. Sorace, M.Zhan, " A data review and re-assessment of ovarian cancer serum proteomic profiling, "BMC Bioinformatics, 2003.

[4]. TR. Golub, et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, 1999.

[5]. T. Nguyen, A. Khosravi, D. Creighton, and s. Nahavandi, "A novel aggregate gene selection method for microarray data classification," pattern recognit.lett, vol 60.-61, 2015.

[6]. M.Xi, J. Sun, L., Liu, F. Fan, and X. Wu, "Cancer feature selection and classification using a Binary Quantum-Behaved particle Swam Optimization and Support Vector Machine, " Comput.Math. Methods Med., vol 2016, 2016.

[7]. S. Aalaei, H. Shahraki, A. Rowhanimanesh, and S. Eslami, "Feature Selection using genetic algorithm for breast cancer diagnosis: an experiment on three different datasets., " Iran.J.Basic Med.Sci., vol. 19, 2016.

[8]. D. Koller and M. Sahami, "Toward optimal feature selection," ICML 96 Proc. Thirteen.Int.Conf. Learn., 1996.

[9]. L. Sarkar et al., "Characteristic attributes in cancer microarrays, " J. Biomed. Inform., vol.35, 2002.

[10]. K. Islam, G. Mujtaba R. R…. ICE2T), 2017 International, and undefined 2017, "Elevator button and floor number recognition through hybrid image classification approach for navigation of service robot in buildings, "researchgate.net.

[11]. Tong DL, Schierz AC. Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data. Artif Intell Med. 2011.

[12]. Li-Yeh Chuang, Cheng-San Yang, Kuo-Chuan Wu, Cheng-Hong Yang. Gene selection and classification using Taguchi chaotic binary particle swarm optimization. Expert Syst Appl. 2011.

[13]. Chuang LY1, Yang CH, Wu KC, Yang CH. A hybrid feature selection method for DNA microarray data. Comput Biol Med. 2011.

[14]. Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics. 2004.

[15]. Martineza E, Alvarezb MM, Trevino V. Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. J Comput Biol Chem. 2010.

[16]. Yang CH. A hybrid filter/wrapper method for feature selection of microarray data. J Med Biol Eng. 2009.

[17]. Önskog J, Freyhult E, Landfors M, Rydén P, Hvidsten TR. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. BMC Bioinformatics. 2011.

[18]. Wang X, Simon R. Microarray-based cancer prediction using single genes. BMC Bioinformatics. 2011.

[19]. 18. Takahashi M, Hayashi H, Watanabe Y, Sawamura K, Fukui N, Watanabe J, Kitajima T, Yamanouchi Y, Iwata N, Mizukami K, Hori T, Shimoda K, et al. Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures. Schizophr Res. 2010.

[20]. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001.

[21]. Nikhil R Pal, Kripamoy Aguan, Animesh Sharma, Shun-ichi Amari. Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. BMC Bioinformatics. 2007.

[22]. Önskog J, Freyhult E, Landfors M, Rydén P, Hvidsten TR. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. BMC Bioinformatics. 2011.

[23]. Wang X, Simon R. Microarray-based cancer prediction using single genes. BMC Bioinformatics. 2011.

[24]. Kao YT, Zahara E. A hybrid genetic algorithm and particle swarm optimization for multimodal functions. Appl Soft Comput. 2008.

[25]. Du SW, Cao LK. 2006. A Learning Algorithm of Artificial Neural Network Based on GA. PSO Proceedings of the 6th World Congress on Intelligent Control and Automation – Dalian, China.

[26]. Juang CF. A hybrid of genetic algorithm and particle swarm optimization for recurrent network design. IEEE Trans Syst Man Cybern B Cybern. 2004

[27]. Robinson JS, Sinton RS. San Antonio: IEEE Antennas and Propagation Society International Symposium [s. n.]; 2002. Yahya Particle Swarm, Genetic Algorithm, and their Hybrids: Optimization of a Profiled Corrugated Horn Antenna.

[28]. Kennedy J, Eberhart R. In: Proceeding of IEEE International Conference on Neural Networks. 1997. A Discrete Binary Version of the Particle Swarm Algorithm. IEEE Service Center – Piscataway.

[29]. Li S, Wu X, Tan M. Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput. 2008.

[30]. Abdi MJ, Hosseini SM. A novel weighted support vector machine based on particle swarm optimization for gene selection and tumor classification. Comput Math Methods Med. 2012.

[31]. V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, ‘‘A review of microarray datasets and applied feature selection methods,’’ *Inf. Sci.*, vol. 282, pp. 111–135, Mar. 2014.

[32]. N. Almugren and H. Alshamlan, ‘‘FF-SVM: New FireFly-based gene selection algorithm for microarray cancer classification,’’ in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Jul. 2019.

[33]. V. Bolon-canedo, et al., “A review of microarray datasets and applied feature selection methods,” Information Sciences, pp.111-135, 2014.

[34]. Niloofar Yousefi Moteghaed, Keivan Maghooli, Shiva Pirhadi, and Masoud Garshasbi[1] , "Biomarker Discovery Based on Hybrid Optimization Algorithm and Artificial Neural Networks on Microarray Data for Cancer Classification”, 2015.

[35]. S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, ‘‘MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia,’’ *Nature Genet.*, vol. 30, no. 1 , pp. 41–47, 2002.

[36]. J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, ‘‘Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,’’ *Nature Med.*, vol. 7, no. 6, pp. 673–679, 2001.

[37]. D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash, ‘‘Gene-expression profiles predict survival of patients with lungadenocarcinoma,’’*NatureMed.*,vol.8,no.8,pp.816–824,Aug.2002.

[38]. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, ‘‘Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,’’ *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[39]. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, ‘‘Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,’’ *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[40]. Nikhil R Pal, Kripamoy Aguan, Animesh Sharma, Shun-ichi Amari. Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. BMC Bioinformatics. 2007.