Neural Computing and Applications (2020) 32:8599–8616
https://doi.org/10.1007/s00521-019-04355-x

ORIGINAL ARTICLE

A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm

Santos Kumar Baliarsingh[1] · Swati Vipsita[1] · Bodhisattva Dash[1]

Received: 3 May 2018 / Accepted: 18 July 2019 / Published online: 27 July 2019

Abstract
In microarray experiments, the sample size is considerably smaller than that of the feature size, thereby imposing the curse of dimensionality problem. To resolve this issue, evolutionary algorithms are often utilized. In this paper, a novel framework for feature selection and classification of the microarray data is presented. Initially, a statistical filter, namely ANOVA, is used to select the relevant genes (features) from the original set of genes. Then, an evolutionary wrapper-based approach utilizing the principles of enhanced Jaya (EJaya) algorithm and forest optimization algorithm (FOA) is proposed to find the optimal set of genes from the previously selected genes. The main objective of using EJaya is to tune the two important parameters, namely local seeding changes and global seeding changes of FOA. During the selection of the optimal set of genes, support vector machine is employed as a classifier to classify the microarray data. To perform a comprehensive experimental study, the proposed method is tested on both binary-class and multi-class microarray datasets. From the extensive result analysis, it has been observed that the proposed technique achieves better classification accuracy with considerably less number of features than that of the benchmark schemes.

Keywords Microarray · ANOVA · Jaya · Forest optimization algorithm (FOA)

1 Introduction

Gene expression profiling based on microarray has appeared as one of the powerful means of cancer treatment and diagnosis [17]. Recently, the DNA microarray mechanism provides opportunities for predicting the significant genes those cause cancer. However, the high dimension of the dataset is one of the primary disadvantages that lie in the study of microarray data. This obstructs the utility of the dataset information and thereby increases the computational overhead. This anomaly also has an adverse effect on the overall performance of the classifier. To overcome these issues, the irrelevant genes need to be discarded using some feature selection (FS) techniques. In fact, it is expected that selecting the relevant genes reduces the size of the gene expression data and enhances the classification accuracy (CA).

Gene selection methods are devised to determine the useful genes present in the microarray data [34]. To identify the significant genes, FS must be integrated within the classifier. Based on the type of integration, FS methods are classified into two distinct groups, namely filter-based and wrapper-based techniques. Filter-based techniques identify the optimal features from the original feature set based on the ranks. Despite the simplicity and computational efficiency, filter techniques are incapable of exploiting the relationship among the features, thereby reducing the overall accuracy. On the contrary, the wrapper-based techniques make use of the knowledge of the classifier to determine the optimal feature subset. The wrapper-based techniques use evolutionary algorithms to identify the optimal solutions by analyzing the search area from a set of solutions (population). The evolutionary algorithms such as genetic algorithm (GA) [12, 30], ant colony optimization (ACO) [25], salp swarm optimization algorithm (SSA) [7], bacterial foraging optimization (BFO) [13], dolphin swarm algorithm (DSA) [54], and particle swarm optimization

✉ Santos Kumar Baliarsingh
c115011@iiit-bh.ac.in

[1] DST-FIST Bioinformatics Laboratory, Department of Computer Science and Engineering, International Institute of Information Technology, Bhubaneswar 751003, India

(PSO) [22, 58] have been successfully utilized for solving various FS problems. These techniques are capable of finding the association between the genes and, hence, lead to better CA. However, the computational overhead of these techniques is very high. Hence, to balance the relationship between the CA and the computational overhead, a hybrid technique needs to be developed.

The proposed algorithm integrates both the filter and wrapper-based techniques to develop a hybrid FS method. The main motivation of this work is to enhance the overall CA by selecting the most relevant genes from the high-dimensional microarray data. The major contributions of the proposed work are highlighted as follows:

a. A statistical filter, namely ANOVA, is used to select the most relevant genes from the original set of genes.
b. To further select the optimal genes and to improve the CA, a recently proposed FOA [24, 47]-based wrapper technique is employed. However, from the literature, it is noticed that the performance of FOA mainly depends on two important parameters, namely local seeding changes (LSC) and global seeding changes (GSC).
c. To select the optimal values of these above-mentioned parameters, an enhanced version of Jaya [31, 52], namely (EJaya) technique, is proposed.
d. To demonstrate the robustness of the proposed method, extensive experiments are performed on seven benchmark datasets. In addition to this, the proposed EJaya-based FOA (EJFOA) approach is compared with some other state-of-the-art methods in terms of CA and the number of features selected.

The remainder of the paper is organized as follows. In Sect. 2, the existing research related to the scope of our work has been briefly described. Section 3 describes the proposed hybrid FS methodology, namely EJFOA. In Sect. 4, experimental design and evaluation procedure are presented. Result analysis is shown in Sect. 5. Lastly, Sect. 6 presents the concluding remark along with the scope for future work.

2 State of the art

Ample literature is available about hybrid FS and classification of microarray data. Some researchers have investigated the advantages of blending filter techniques based on feature ranking with search methods to develop a hybrid FS algorithm. For instance, Hernandez et al. [29] proposed the ranking of features based on Fisher discriminant criterion followed by support vector machine (SVM) wrapped GA. The effectiveness of the method is tested using three benchmark datasets giving highly competitive results. The limitation of this approach is that it is tested only on binary-class datasets. Alshamlan et al. [3, 4] presented two hybrid gene selection methods employing mRMR filter followed by artificial bee colony (ABC) and genetic bee colony (GBC) algorithms. The algorithms are applied on six binary and multi-class datasets showing reliable performance. The authors have reported that GBC shows better classification performance over ABC. The major drawback of this work is the lack of biological significance of the predicted genes.

Tabakhi et al. [57] suggested an unsupervised gene selection method incorporating ant colony optimization (ACO) into the filter method. They have tested the performance upon five standard datasets. This work too lacks in biological interpretation of selected genes. Apolloni et al. [5] proposed an algorithm that combines a wrapper FS technique based on binary differential evolution (BDE). The robustness of the method is tested using four machine learning techniques over six datasets. The advantage of this technique is that it reduces the original gene size by more than 99%. No biological analysis of the resulted genes has been provided by the authors, which can be thought of as a limitation of the work. Chinnaswamy and Srinivasan [10] proposed a hybrid FS method using correlation coefficient filter and PSO-based extreme learning machine (ELM). This method is capable of selecting 2–8% of the informative genes from the original dataset. The authors have demonstrated that ELM classifier produces better results compared to tree-based classifiers.

Elyasigomari et al. [18] presented a two-stage FS approach using mRMR filter followed by a combination of cuckoo optimization algorithm (COA) and harmony search (HS)-based SVM. The authors claim that their approach significantly outperforms other existing methods in selecting less number of genes with higher accuracy. The biological relevance of the selected genes is confirmed in each of the cancer types. However, the computational complexity of this technique may be high due to the combination of multiple evolutionary algorithms. Motieghader et al. [42] proposed a hybrid meta-heuristic method integrating GA and learning automata (LA). The experimental results demonstrate remarkable performance compared to other existing algorithms. The major drawback of this technique is its high computational complexity.

Nieto and Alba [23] introduced a parallel PSO for gene selection where a set of independent PSOs run simultaneously. This parallel algorithm outperforms its sequential version in terms of computational time and classification accuracy. However, the authors have not verified their approach on multi-class datasets, which is a limitation of the work. Wang et al. [60] introduced Markov Blanket into incremental wrapper-based FS, in which the relevant genes are selected eliminating the redundant ones. The authors have proven the effectiveness of the technique by

experimenting over six popular microarray datasets. The major disadvantage of this method is that it is more time-consuming because of interdependence and redundancy between the genes.

Kar et al. [33] suggested a wrapper FS method based on PSO and adaptive KNN classifier. The experimental results show the efficiency of the method in terms of CA, number of informative genes selected, and computing time. Mohapatra et al. [41] proposed a wrapper gene selection approach based on the principle of modified cat swarm optimization (MCSO) and kernel ridge regression (KRR). The efficacy of the model is examined using CA, sensitivity, specificity, G-mean, *F*-score, and area under curve (AUC). The authors have implemented the technique on both binary-class and multi-class datasets and observed that the model works better on binary-class datasets compared to multi-class datasets. They have not tested the biological significance of the selected genes which is a major drawback of the suggested work.

Wang et al. [62] developed a method combining adaptive elastic net (AEN) with conditional mutual information (CMI) for microarray gene selection. The major advantage of this method is that it reduces the influence of the wrong initial estimation to gene selection and classification. This approach has been applied only on two binary-class datasets, which can be considered as a limitation. Algamal and Lee [1] proposed a two-stage sparse logistic regression for an efficient gene selection and cancer classification. Experimental results show that the suggested method significantly outperforms other existing techniques in terms of CA, AUC, and G-mean.

Liu et al. [38] proposed a hierarchical ensemble approach using error correcting output codes (ECOC) for multi-class microarray classification. The limitation of this work is the lack of description regarding the conservation of most important oncogenes. Canedo et al. [9] presented a distributed FS approach which divided the features vertically and applied filters on feature subsets independently then merged the feature subsets based on their CA. The major advantage of this method is the shorter execution time compared to other standard algorithms applied on non-partitioned datasets. Again this work lacks in biological analysis of genes selected. Sharma et al. [53] proposed a FS method exploring improved regularized linear discriminant analysis (IRLDA) for microarray gene selection and cancer classification. The effectiveness of the method is tested on three standard datasets which shows promising results.

Although several researchers have used hybrid FS techniques on the basis of evolutionary search techniques, to the best of the authors' knowledge, this is the first attempt at exploring EJFOA gene selection algorithm for cancer classification using gene expression microarray profiles.

3 Proposed methodology

The primary goal of the proposed work (see Fig. 1) is to develop a hybrid algorithm which can determine the optimal set of relevant genes with an improved CA.

Each of the stages of the proposed method is briefly discussed as follows.

i. *Data collection*
   In the first step, the gene expression microarray data is collected from Kent Ridge Biomedical Dataset Repository [26, 69] and ELVIRA Biomedical Data Set Repository (http://leo.ugr.es/elvira/DBCRepository/).
ii. *Imputation of missing value and data normalization*
   In the second step, a missing data imputation technique is used to fill the missing values in the dataset, followed by normalization. Missing data are imputed by computing the mean value of the respective feature vector. The values of genes are further normalized in the range of [0, 1] employing a min—max normalization method.
iii. *First stage gene selection using filter*
   In the third step, a filter technique, namely ANOVA, is employed on each of the datasets to select the most relevant genes. This reduced gene subset is passed to the next stage for further optimization.
iv. *Second stage gene selection using wrapper*
   In the fourth step, to further optimize the selected genes obtained from the filter as well as to improve the overall performance, a recently developed evolutionary technique, namely FOA, is used. However, the performance of FOA depends on two important parameters, namely LSC and GSC. Hence, to optimize the values of these two parameters, an EJaya algorithm is proposed.

3.1 Basic concepts and preliminaries

3.1.1 ANOVA

ANOVA is a parametric method [55] which can be used to compare the 'class means' for a particular feature of the dataset. According to the null hypothesis, there is no significant difference between the class means of a feature; therefore, that particular feature can be discarded from the feature set. As per the alternate hypothesis, there is a significant difference between the class means, and hence, the

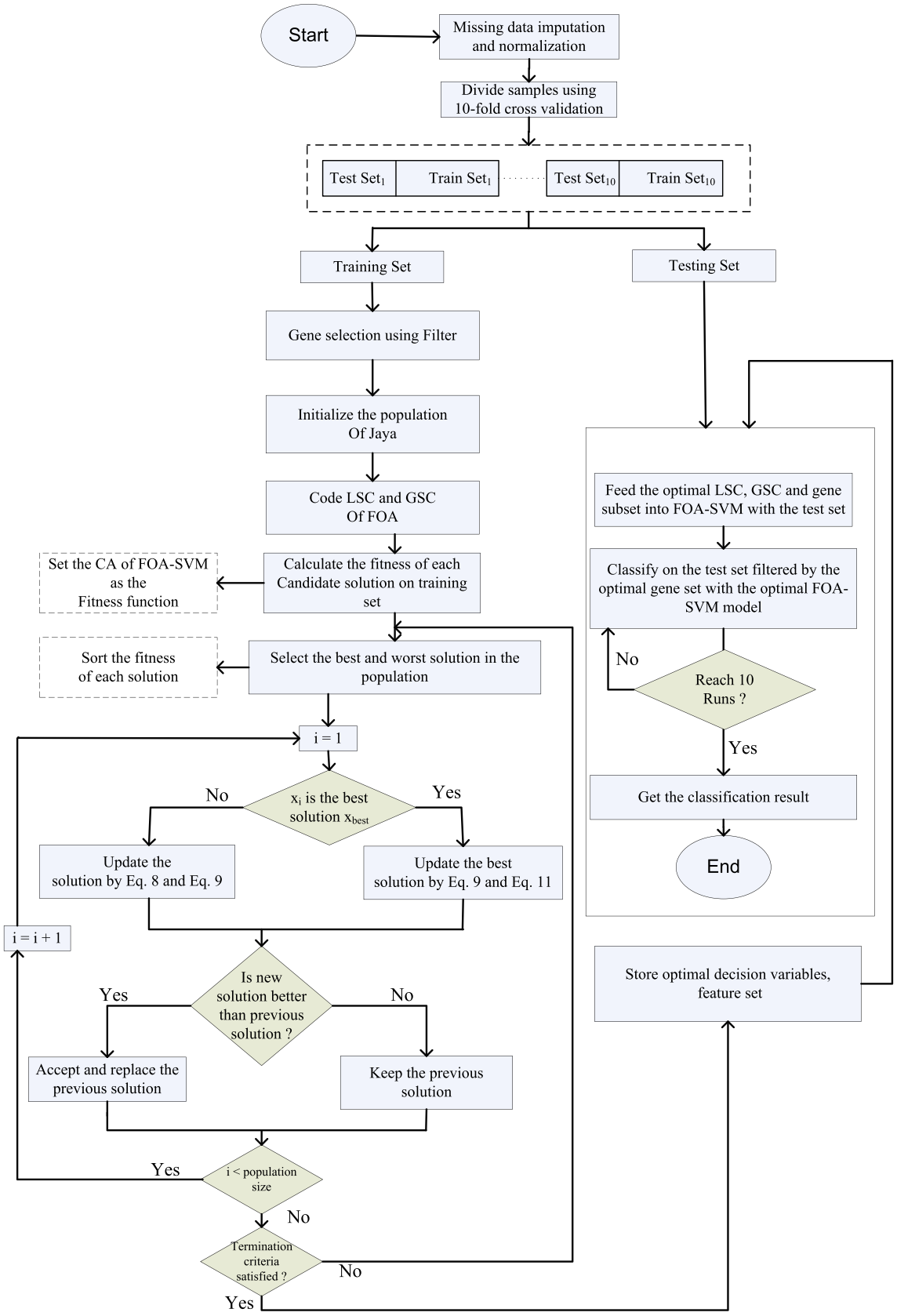


Fig. 1 Block diagram of the proposed framework

feature will be retained. A feature will be considered in the final feature set depending on the p-value of the feature. The p-value is computed from the F-statistic of ANOVA. The F-statistic is calculated as follows:

1. Inter-class variation is computed as:

$$\text{Sum of square errors between the classes (SSB)} = n_1(\mu_1 - \mu)^2 + n_2(\mu_2 - \mu)^2 + \cdots \quad (1)$$

$$\text{Degree of freedom between the classes (DFB)} = \text{Total number of classes} - 1 \quad (2)$$

$$\text{Mean squared error between the classes (MSB)} = \text{SSB}/\text{DFB} \quad (3)$$

2. Intra-class variation is computed as:

$$\text{Sum of square errors within the classes (SSW)} = (n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2 + \cdots \quad (4)$$

$$\text{Degree of freedom within the classes (DFW)} = \text{Total number of samples} - \text{Total number of classes} \quad (5)$$

$$\text{Mean squared error within the classes (MSW)} = \text{SSW}/\text{DFW} \quad (6)$$

where 'n' represents sample size, 'n_c' represents the number of samples in class 'c', 'c' represents the number of classes, 'μ' represents mean of all classes, 'μ_c' represents mean of class 'c', 'σ_c' represents the standard deviation of class 'c'.

3. F-statistic is computed as:

$$F = \text{MSB}/\text{MSW} \quad (7)$$

The implementation of ANOVA for FS is shown in Algorithm 1

Algorithm 1 ANOVA for Feature Selection

Input: M : Feature matrix of size $S \times G$; where S represents sample size and G represents feature size
Output: Select top N features

for each feature f_i **do**
  ▷ i=1,2,...,G.
  Evaluate the value of MSB using Eq. (3)
  Compute the value of MSW using Eq. (6)
  Compute the F-Statistics value (F_i) using Eq. (7)
  Compute the p-value (p_i) for each F-Statistics using the F-distribution table
  if $p_i < 0.001$ **then**
    Select the feature f_i
    Append f_i to a feature matrix G_M
  else
    feature f_i is discarded
  end if
  Sort the features in ascending order of their p value
  if $size(G_M) > 500$ **then**
    Select only top-500 features
  else
    Keep the feature matrix G_M as it is
  end if
end for
Return the features from feature matrix G_M

3.1.2 Forest optimization algorithm

FOA is a relatively new evolutionary technique inspired by the concept of trees in a forest [24]. This algorithm has three major steps, namely population initialization, local seeding changes in the tree, and global seeding changes in the tree. Similar to the other evolutionary algorithms, this algorithm also starts with population initialization of the trees that form the forest. Each tree in the forest is an array of 0's and 1's having size '$G + 1$', where 'G' represents the number of features in the dataset. One part of every tree is reserved to store its *'Age'* which is initialized to zero for a newly generated tree. Each tree in the forest is a potential solution for the optimization problem. A '1' in the tree signifies the corresponding feature is selected, and a '0' indicates exclusion of that feature. This is shown in Fig. 2.

Age	0	1	1	0	1	0	0	1

Fig. 2 Initialization of a tree in FOA

At the time of seeding process of the trees, new seeds are fallen below their parent trees and, in turn, evolve as independent young trees. This process is mimicked by a parameter called *'Local Seeding Changes'* (LSC) in FOA. This seeding process is applied on every zero-aged tree. At this stage, the *'Age'* of the parent trees excluding the newly generated trees is incremented by 1. For example, let the total number of features in the dataset be 8, and the value of LSC is 25% of the total number of features. Hence, 2 bits are changed in the local seeding process, which is shown in Fig. 3.

Algorithm 2 Gene selection using FOA

Input: M : Feature matrix of size $S \times G$; where S represents samples and G represents feature set
LT=Life time of tree
GSC=Global seeding changes
LSC=Local seeding changes
TR=Transfer rate
AL=Area limit
Output: M : Feature matrix of size $S \times f$, where S is the sample size and f is the features set ($f < G$) with optimum CA
Initialize the population with random binary trees(0's or 1's)
  Each tree is a (G+1)- dimensional array x, x = ($x_1, x_2, ..., x_G$, age)
  The 'Age' of each tree is initialized to '0' which is represented by the last bit of the tree
while maximum iteration is not reached **do**
  Apply local seeding operation on each trees having Age 0
  for $i = 1$ to LSC **do**
    Flip any one bit of each of the trees from 1 to 0 or 0 to 1 at a random
  end for
  Increment the Age of each tree by one except for new generated trees
  if Age of the trees > LT **then**
    Remove those trees from forest
    Add these trees to candidate population
  end if
  Compute the CA of each tree
  Sort the trees by their CA
  if number of trees > AL **then**
    Remove those trees from forest
    Add these trees to candidate population
  end if
  Select TR percentage of trees from the candidate population
  for every selected tree **do**
    Select GSC parameter
    Randomly flip GSC percentage of bits of each tree from 1 to 0 or 0 to 1
  end for
  Sort the trees by their CA
  Select the best trees based on CA
  if Age of the best tree > 0 **then**
    Update its 'Age' to 0
    Add this best tree to forest
  end if
end while
Return the best trees with CA

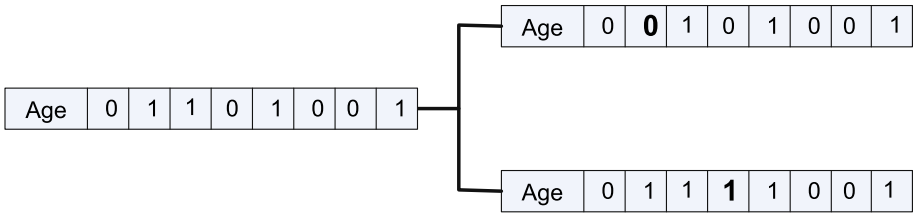


Fig. 3 Local seeding operation on a tree with LSC = 25%

Next step is limiting the population of the trees which is carried out by two parameters such as life time (LT) and area limit (AL). Initially, the trees with *Age* greater than the pre-defined life time are discarded from the forest to form the candidate population (CP). Thereafter, the rest of the trees are arranged in terms of their fitness values (CA). Moreover, when the number of trees in the forest goes beyond the pre-defined parameter AL, then the additional trees are eliminated from the forest and are included in CP. Further, based on the transfer rate (TR) parameter, a certain percentage of CP trees will participate in global seeding (GS) process. In GS process, GSC number of bits are selected randomly from each of the trees in CP, and the values of these bits are flipped. For example, consider the total number of features are 8, and GSC value is 50% of the total number of features, therefore, 4 bits are randomly flipped which is shown in Fig. 4. CA is then calculated for each of the modified trees using SVM classifier and sorted according to the obtained CA. The tree having highest CA is selected as the best tree, and the age of that tree is changed to 0. This selected tree will become the parent tree for the next iteration.

The above steps are repeated until the stopping criterion is met. The stopping criterion can be the number of iterations or no improvement in the CA for successive iterations. In the proposed approach, number of iterations is considered as the stopping criterion which is taken as 10. Further, the working principle of FOA for gene selection is presented in Algorithm 2.

3.1.3 Jaya algorithm

Jaya is a simple and powerful global optimization algorithm proposed by Rao et al. [52] for solving constrained and unconstrained optimization problems. From the application point of view, this is a simple and novel optimization technique. It is a parameter-less algorithm that converges to the optimal solution in comparatively less number of function evaluations. Due to the above advantages, this algorithm can be used for solving different engineering optimization problems. This algorithm is based on the concept that the solution obtained for a given problem should move toward the best solution and should avoid the worst solution.

Let there be R number of candidate solutions (i.e., population size, $i = 1, 2, \ldots, R$) and C number of decision variables (i.e., $j = 1, 2, \ldots, C$) for each candidate solution. At any zth iteration, the best solution obtained out of all candidate solutions is denoted as x^z_{best} and the worst solution is denoted as x^z_{worst}. If $x^z_{i,j}$ is the value of the jth decision variable for the ith candidate solution during zth iteration, then $x^z_{i,j}$ is updated according to following equation as

$$X^z_{i,j} = x^z_{i,j} + \mu^z_{1,j}(x^z_{j,\text{best}} - |x^z_{i,j}|) - \mu^z_{2,j}(x^z_{j,\text{worst}} - |x^z_{i,j}|) \quad (8)$$

where $x^z_{j,\text{best}}$ and $x^z_{j,\text{worst}}$ are the best and worst values of the jth decision variable, respectively, and $X^z_{i,j}$ is the updated value of $x^z_{i,j}$. $\mu^z_{1,j}$ and $\mu^z_{2,j}$ are the two random numbers during zth iteration in the range [0,1]. $X^z_{i,j}$ is accepted in $x^z_{i,j}$ if it gives a better objective function value. Thus, modified $x^z_{i,j}$, at the end of each iteration, is then allowed to take part in the next iteration. This process continues till the number of iterations is completed.

3.2 Proposed EJFOA algorithm

From the literature, it has been noticed that selection of optimal features plays an important role in the performance of the classifier in terms of complexity and CA [27]. In this paper, FOA is employed as a wrapper method in combination with SVM, to select the optimal subset of features. However, the performance of FOA mainly depends on two important parameters, namely LSC and GSC. To choose the optimal values of these two parameters, an enhanced Jaya algorithm is proposed. The reason for choosing Jaya over other algorithms is that it is a parameter-less algorithm.

The values of the two dependent parameters of FOA, namely, LSC and GSC, can vary from 1 to 50% of the total number of features, which means the values can lie in the range of [0.01, 0.5]. This is due to the fact that the number of child tree generation of FOA depends on the upper and lower bound values of the LSC parameter. How many bits of a tree will be flipped is decided by the GSC parameter. Hence, if the upper bound value is set to be more than 50%, it will increase the computational overhead in terms of elapsed time of the algorithm. For various combinations of LSC and GSC within this range, the combinatorial search

Fig. 4 Global seeding operation on a tree with GSC = 50%

space becomes exponentially large. Hence, to efficiently explore the huge search space, Jaya algorithm is employed. However, it may so happen that, the values of LSC and GSC parameters might go beyond their lower and upper bounds. Hence, to deal with the out-of-bound problem, Eq. 9 is utilized.

$$X_{i,j}^{z+1} = \begin{cases} 0.01 & \text{if } X_{i,j}^{z+1} < 0.01 \\ 0.5 & \text{if } X_{i,j}^{z+1} > 0.5 \end{cases} \quad (9)$$

The best solution of Jaya algorithm plays a crucial role throughout the search process as it leads and drags other solutions toward its own area. However, there is a possibility that the best solution may be present in a local optimum. In this type of situation, other solutions are comfortably drawn toward the region of the best solution which causes premature convergence. To overcome this problem, chaotic map is introduced in this study to improve the convergence rate of the Jaya algorithm. Due to the ergodicity and randomicity properties of chaos, it can perform overall searches at higher speed compared to the stochastic searches [45, 46, 66]. In the present work, a well-known logistic map is employed as a chaotic sequence which is defined by Eq. 10. Using this chaotic map, an enhanced Jaya algorithm is proposed where the best solution of Jaya algorithm is updated as per Eq. 11.

$$R_{z+1} = 4 \cdot R_z \cdot (1 - R_z) \quad (10)$$

$$X_{j,\text{best}} = x_{j,\text{best}} + 4 \cdot R_z \cdot (1 - R_z) \quad (11)$$

where z denotes the iteration number, R_z denotes the value of zth chaotic iteration. The value of R_0 is initialized in the range of [0, 1].

Algorithm 3 Proposed EJFOA algorithm

Input: M : Feature matrix of size $S \times G$; where S represents samples and G represents feature set
PS=Population size
MI=Maximum iteration
NDV=Number of decision variables (LSC, GSC)
Output: Optimal LSC, GSC, M : Feature matrix of size $S \times f$, where S is the sample size and f is the features set $(f < G)$ with optimum CA

```
Randomly initialize the population for the decision variables(LSC, GSC) within the range of [0.01, 0.5]
while maximum iteration is not reached do
    for each population do
        Obtain the fitness value (CA) using Algorithm 2
    end for
    Sort the candidate population as per the obtained fitness value and find the best and worst solution
    if x_i is the best population x_best then
        Update the best population using Equation 9 and 11
    else
        Update the candidate population using Equation 8 and 9
    end if
end while
Return the optimized LSC, GSC along with reduced feature set and CA
```

The two parameters LSC and GSC are the two decision variables for EJaya algorithm. In order to find their optimum value, the proposed EJFOA algorithm is outlined in Algorithm 3.

4 Experimental design and evaluation procedure

This section presents the detailed experimental setup to evaluate the performance of the proposed work. The proposed hybrid method is validated on both binary-class and multi-class microarray datasets. First, the dataset is divided into train and test set using tenfold cross-validation. Then, the train set is passed through a filter, namely ANOVA, to obtain the relevant feature vectors from the original set of feature vectors. The feature vectors thus obtained are introduced to an EJFOA-based wrapper module. This module utilizes an EJaya algorithm to optimize the parameters of FOA followed by an SVM classifier to

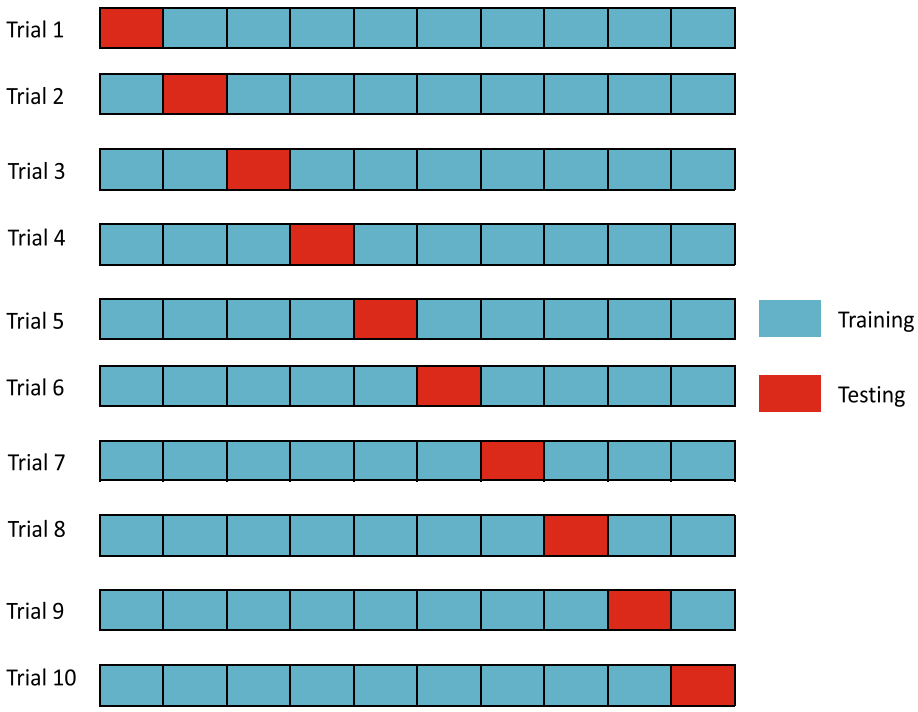


Fig. 5 Illustration of tenfold cross-validation setting for a single run

compute the overall accuracy (fitness value) of the proposed method.

To avoid the over-fitting problem and to achieve an unbiased classification result, tenfold CV is used to test the performance of the proposed model. Figure 5 demonstrates the setting of a tenfold CV for a single run. In each trial, onefold is used for testing, and the rests for training. Further to overcome the randomness behavior of the meta-heuristic model, ten independent iterations are performed, and the average of ten runs is reported as the final result.

The simulations are carried out using MATLAB 2015a on a PC with Intel Core i5 CPU(2.70 GHz) and 8GB of RAM. A brief description of the datasets used in the experiment and the parameters is discussed in the following subsections.

4.1 Datasets

To carry out the experiments, seven different standard datasets are considered from the Kent Ridge Biomedical Dataset Repository [26, 69]. Table 1 briefly lists the important attributes of each of the datasets. Before the data analysis by the proposed method, the datasets are normalized in the range of [0, 1].

4.2 Parameter settings

As discussed in Sect. 3.1.2, it requires five distinct parameters to be initialized, namely *life time*, *area limit*, *transfer rate*, LSC, and GSC. To perform the experiments, the values initialized to the respective parameters are depicted in Table 2. Moreover, the values of LSC and GSC parameters are randomly initialized within the specified range.

Table 1 Description of the datasets

Dataset	#Features	#Samples	#Classes
Leukemia-2 [26]	7129	72	2
Colon tumor [2]	2000	62	2
Ovarian cancer [50]	15,154	253	2
Leukemia-3 [6]	7129	72	3
Leukemia-4 [37]	7129	72	4
Lymphoma-3 [69]	4026	62	3
SRBCT [69]	2308	83	4
Lung cancer-5 [8]	12,600	203	5

5 Results analysis

To evaluate the efficiency of the proposed method, six different performance metrics, namely CA, sensitivity, specificity, Matthews correlation coefficient (MCC), and F-measure, are considered.

5.1 Performance metrics

- *Confusion matrix* This is used to describe the performance of a classification model on a set of test data for

Table 2 Parameter initialization of FOA

Parameter	Value
Life time	15
Area limit	50
Transfer rate	0.05
LSC	[0.01, 0.5]
GSC	[0.01, 0.5]

which the target classes are known [56]. Table 3 shows the confusion matrix and performance measures for a binary classifier.

- *Matthews correlation coefficient (MCC)* The MCC, introduced by biochemist Matthews [39] is employed to measure the quality of the classification algorithm. The MCC is primarily a correlation coefficient between the observed class and the predicted class. The MCC value can lie in the range of $[-1, 1]$. An MCC 1 represents a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation. The MCC can be computed from the confusion matrix using Eq.12.

$$\mathrm{MCC} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{12}$$

- *F-measure* F-measure is a measure of test accuracy. It takes into account both the precision and the recall to compute the score. The best value of F-measure can be 1 indicating perfect precision and recall and worst value can be 0. The F-measure can be computed from the confusion matrix using Eq. 13.

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

5.2 Experimental results of feature pre-selection using ANOVA

The number of features in the original microarray dataset is very large containing irrelevant genes which cause the curse of dimensionality problem. To overcome this, a filter technique, namely ANOVA, is used. This filter is applied to each of the features of the dataset. ANOVA is based on a hypothesis testing, wherein the null hypothesis states that there is no significant difference between the classes in terms of their mean, median, and variance. The alternate hypothesis states that there is a significant difference. So, it is implied that the features which satisfy the null hypothesis have no effect on the classification result, and therefore, those features can be removed. On the other hand, the features that satisfy the alternate hypothesis influence the classification accuracy, and therefore, they are accepted. To accept or discard a feature is decided by the corresponding p-value of the feature, which tells how effective a feature is at separating the classes.

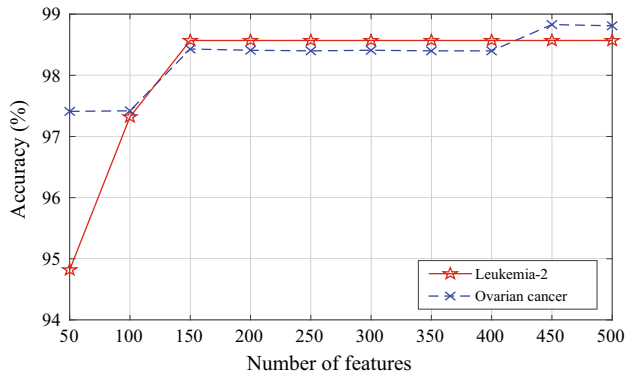
By considering the 99.9% of the confidence interval, if the p-value is less than 0.001, the null hypothesis is rejected, and alternate hypothesis is accepted. Sorting these features in ascending order of their p-value helps identifying the features with strong representations. Based on their p-value, a threshold of top 500 genes (as suggested by [16]) is selected for the next stage for all datasets except Colon tumor dataset where only top-53 genes are qualified to be selected whose p-value is less than 0.001.

After FS, the proposed EJFOA-SVM method is applied on each of the datasets. However, without the previous knowledge of the dataset, it is hard to decide the optimal number of genes necessary for classification. To resolve this issue, forward FS approach is used, wherein the top ranked genes corresponding to ascending p-value are considered. Various subsets of the top ranked genes are employed for classification, and their corresponding CA is reported in Fig. 6a–c.

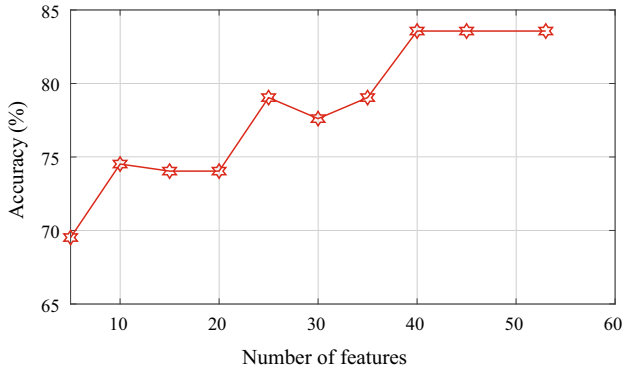
Figure 6a shows the change in CA with increase in the number of genes on Leukemia-2 and Ovarian cancer dataset. From the figure, it is observed that in the case of Leukemia-2 dataset, a maximum accuracy of 98.57% is achieved with top 150 genes. For Ovarian cancer, an accuracy of 98.83% is obtained with top 450 genes. Figure 6b shows the change in CA of Colon cancer with change in the number of selected genes. In Colon cancer

Table 3 Confusion matrix

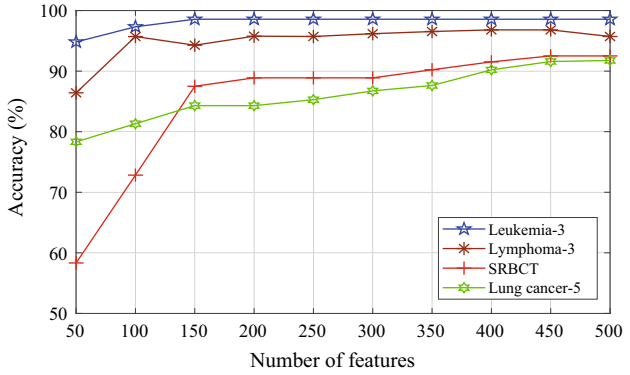
	Target class		
	neg	pos	
Output class			
Classified as neg	tn	fn	
Classified as pos	fp	tp	$\text{Precision} = \frac{tp}{tp + fp}$
	$\text{Specificity} = \frac{tn}{tn + fp}$	$\text{Recall} = \frac{tp}{tp + fn}$	$\text{Accuracy} = \frac{tp + tn}{tp + fp + fn + tn}$



(a) Binary-class dataset results



(b) Colon tumor results



(c) Multi-class dataset results

Fig. 6 Number of features versus accuracy

dataset, a maximum accuracy of 83.57% is achieved with top 400 genes. Similar results have been obtained for multi-class datasets which is shown in Fig. 6c. From the figure, it is observed that a maximum accuracy of 98.57%, 96.81%, 92.51%, and 91.76% is achieved with 150, 400, 450, and 500 genes for Leukemia-3, Lymphoma-3, SRBCT, and Lung cancer-5 datasets, respectively.

5.3 Experimental results using EJFOA-SVM

Further, to get the optimized set of features from the reduced set of features obtained from ANOVA, a wrapper-based approach, namely EJFOA-SVM, is utilized. Here, EJaya algorithm is used to optimize the two parameters of FOA.

5.3.1 Results of binary-class dataset

There are three binary-class datasets used in this experiment, namely Leukemia-2, Colon tumor, and Ovarian cancer. After applying ANOVA as a filter technique, 500 top ranked genes have been selected from Leukemia-2 and Ovarian cancer dataset. However, only 53 genes are selected from Colon cancer dataset.

The average classification results obtained with these top ranked genes by the proposed method are reported in Table 4. From the table, it is observed that for Ovarian cancer our method achieves 100% accuracy with only four genes. For Leukemia-2 dataset, selecting four genes by the proposed approach leads to 98.57% accuracy. Similarly, for Colon cancer dataset, our method provides 96.90% accuracy with only three genes. The optimal values of LSC, and GSC parameters obtained by the best results across ten runs are reported in Table 5.

5.3.2 Results of multi-class datasets

This section describes the results obtained by the multi-class datasets used in this experiment, namely Leukemia-3, Lymphoma-3, SRBCT, and Lung cancer-5. The average classification performance resulted by the proposed method with the top ranked genes is reported in Table 6. From the table, it is noticed that for Leukemia-3 our method achieves 98.55% accuracy with only four genes. For Lymphoma-3 dataset, selecting three genes by the proposed approach leads to 99.87% accuracy. In the case of SRBCT, the obtained accuracy is 97.77% with five genes. Similarly for Lung cancer-5 dataset, our method provides 94.56% accuracy with only four genes.

5.3.3 Validation on independent datasets

We have conducted experiments to see the goodness of a feature set obtained from one dataset on different independent datasets. In our case, the predictive genes selected from the Leukemia-2 dataset are validated on Leukemia-3 and Leukemia-4 datasets independently which is reported in Table 7. From the table, it is observed that the four biomarkers selected by the proposed method for Leukemia-2 dataset are Myeloperoxidase, Oncoprotein 18, Homeo box A9, and Proteasome iota chain leading to 98.57% accuracy. Whenever these four genes are tested on Leukemia-3 and Leukemia-4 datasets, the accuracies are found to be 98.55% and 98.51%, respectively. The maximum difference in accuracy among the three datasets is 0.06%, which is very small. Therefore, it can be inferred that the

Table 4 Average classification results of the proposed model on binary-class datasets

Dataset	Accuracy (%)	Sensitivity	Specificity	MCC	*F*-measure	Kappa
Leukemia-2	98.57(4)	0.95	1	0.964	0.966	0.958
Colon tumor	96.90(3)	0.925	0.916	0.888	0.941	0.869
Ovarian cancer	100(4)	1	1	1	1	1

The number of selected genes is shown in parenthesis

Table 5 The optimal values of LSC and GSC obtained by the best solutions across ten runs

Dataset	LSC	GSC
Leukemia-2	0.026	0.191
Colon tumor	0.041	0.372
Ovarian cancer	0.036	0.241
Leukemia-3	0.026	0.191
Lymphoma-3	0.040	0.168
SRBCT	0.018	0.297
Lung cancer-5	0.050	0.325

genes selected from our approach are good enough for prediction.

5.4 Execution time of the proposed method

The total elapsed time taken by the filter stage and wrapper stage with the proposed method is shown in Table 8. From the table, it can be observed that the overall time consumed by the proposed method is 22.523, 12.355, 36.516, 23.964, 17.361, 17.924, and 29.563 seconds for Leukemia-2, Colon tumor, Ovarian cancer, Leukemia-3, Lymphoma-3, SRBCT, and Lung cancer-5 datasets, respectively.

5.5 Biological functions of selected informative genes

The biological significance of the selected genes to each cancer type is investigated and reported in Table 9. Furthermore, their functions are studied and described as follows.

Myeloperoxidase (MPO) It is the hallmark enzyme of the myeloid lineage. The diagnosis of acute myeloid leukemia (AML) is easy if more than 3% of blast cells are confirmed to be cytochemically MPO positive. The expression of Myeloperoxidase is widely accepted as an important marker for the diagnosis of AML in WHO classifications [35].

Oncoprotein 18 (Op18) This gene belongs to the stathmin family of genes which encodes a proliferation-related cytosolic phosphoprotein and induced in normal lymphocytes following mitogenic stimulation. Op18 gene expression is greatly increased in acute leukemia cells [40].

Homeo box A9 This gene is commonly known as HOXA9. It is a homeodomain-containing transcription factor which has a crucial role in hematopoietic stem cell expansion and is commonly deregulated in acute leukemia. The upstream genetic alterations in acute myeloid leukemia lead to overexpression of HOXA9 [15].

Proteasome iota chain This gene is also known as PSMA6 whose main function is to degrade unneeded or damaged proteins. The proteasome is overexpressed in acute leukemia, and its inhibition is used for the treatment of acute leukemia [61].

Homo sapiens RON mRNA for tyrosine kinase (MST1R) MST1R is a protein coding gene that belongs to the MET proto-oncogene family. This gene encodes a cell surface

Table 6 Average classification results of the proposed model on multi-class datasets

Dataset	Accuracy (%)	Sensitivity	Specificity	MCC	*F*-measure	Kappa
Leukemia-3	98.55(4)	0.991	0.993	0.982	0.988	0.967
Lymphoma-3	99.87(3)	0.980	0.980	0.980	0.980	0.980
SRBCT	97.77(5)	0.983	0.991	0.977	0.982	0.940
Lung cancer-5	94.56(4)	0.917	0.971	0.922	0.934	0.868

The number of selected genes is shown in parenthesis

Table 7 Validation of feature set on independent datasets

Gene Index	Accession	Gene name	Accuracy (%)		
			Leukemia-2	Leukemia-3	Leukemia-4
1780, 1928, 3848, 4328	M19507, M31303_rna1_at, U82759, X59417_at	MPO, STMN1, HOXA9, PSMA6	98.57	98.55	98.51

Table 8 The running time (in s) of the proposed method on seven datasets

Class	Dataset	Filter time (ANOVA)	Wrapper time (EJFOA-SVM)	Total time
Binary-class	Leukemia-2	14.329	8.194	22.523
	Colon tumor	4.141	8.214	12.355
	Ovarian cancer	29.423	7.093	36.516
Multi-class	Leukemia-3	14.116	9.848	23.964
	Lymphoma-3	7.966	9.395	17.361
	SRBCT	4.697	13.227	17.924
	Lung cancer-5	25.000	4.563	29.563

Table 9 Most informative genes obtained by the best solution across ten runs using the proposed method

Dataset	Predictive genes
Leukemia-2	Myeloperoxidase (MPO)
	Oncoprotein 18 (STMN1)
	Homeo box A9 (HOXA9)
	Proteasome iota chain (PSMA6)
Colon tumor	Homo sapiens RON mRNA for tyrosine kinase (MST1R)
	Homo sapiens pterin 4a carbinolamine dehydratase (PCBD) mRNA complete cds (PCBD1)
	Cell division cycle 42 (CDC42), GTP binding protein, 25 KDA
Ovarian cancer	Decorin (DCN)
	Microfibrillar-associated protein 4 (MFAP4)
	PDZ domain containing ring finger 3 (PDZRN3)
	Semaphorin 3C (SEMA3C)
Lymphoma-3	UG Ha. 1 69081 ets variant gene 6 (ETV6)
	UG Hs. 120716 ESTs
	MCL1 myeloid cell differentiation protein
Lung cancer-5	Microtubule associated protein RP/EB family member 3 (MAPRE3)
	ATP-binding cassette subfamily C member 3 (ABCC3)
	Claudin 4 (CLDN4)
	Erb-b2 receptor tyrosine kinase 2 (ERBB2)
SRBCT	Fc fragment of IgG receptor and transporter (FCGRT)
	Caveolin 1 (CAV1)
	CD99 molecule (Xg blood group)
	Protein tyrosine phosphatase non-receptor type 13 (PTPN13)
	Major histocompatibility complex, class II, DM alpha (HLA-DMA)

receptor for macrophage-stimulating protein (MSP) with tyrosine kinase activity. Research suggests that RON expression is altered in colon cancer, and abnormal expression of RON variants could lead to the progression of colon cancer [36].

Homo sapiens pterin 4a carbinolamine dehydratase(PCBD) mRNA complete cds This gene encodes pterin-4 alphacarbinolamine dehydratase, which helps recycle a molecule known as tetrahydrobiopterin (BH4). Research suggests that the PCB gene could potentially serve as a new marker of malignant colon cells [19].

Cell division cycle 42 (CDC42), GTP binding protein, 25 KDA CDC42 is an oncogenic Rho GTPase overexpressed in colorectal cancer. Research suggests that CDC42 regulates gene transcription and several cancer-related signaling pathways, including cell migration and cell proliferation [59].

Decorin (DCN) This gene encodes a member of the small leucine-rich proteoglycan family of proteins. This gene and the related gene biglycan are thought to be the result of gene duplication. Research suggests that expression of decorin leads to ovarian cancer cell growth [44].

Microfibrillar-associated protein 4 (MFAP4) This gene encodes a protein which has binding specificities for both collagen and carbohydrate. It is thought to be an extracellular matrix protein which is involved in cell adhesion or intercellular interactions. The gene is located within the Smith–Magenis syndrome region. MFAP4 is found highly expressed in patients of serous ovarian cancer [67].

PDZ domain containing ring finger 3 (PDZRN3) This gene encodes a member of the LNX (Ligand of Numb Protein-X) family of RING-type ubiquitin E3 ligases. This protein may be targeted for degradation by the human papilloma virus E6 protein. Research suggests that PDZRN3 gene is responsible for ovarian cancer [63].

Semaphorin 3C (SEMA3C) This gene encodes a secreted glycoprotein that belongs to the semaphorin class 3 family of neuronal guidance cues. The encoded protein contains an N-terminal sema domain, integrin and immunoglobulin-like domains, and a C-terminal basic domain. An increase in the expression of this gene correlates with an increase in cancer cell invasion and adhesion. In ovarian cancer, high levels of SEMA3C are associated with shorter patient survival [21].

UG Ha. 1 69081 ets variant gene 6 It is known as ETV6 and facilitates a protein that functions as a transcription factor. The ETV6 gene plays a key role in regulating blood cell formation. Fusions of the ETV6 gene to fibroblast growth factor receptor 3 in peripheral T cell lymphoma have been reported in [64].

UG Hs. 120716 ESTs This gene is linked to serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 9 (SERPINA9), which is strongly expressed in B cell lymphomas [28].

MCL1 myeloid cell differentiation protein It is a protein that, in humans, is encoded by the MCL1 gene. This gene encodes an antiapoptotic protein that is a member of the Bcl-2 family. Bcl-2 plays an important role in some cancers such as leukemia-2 and lymphoma [11].

Microtubule associated protein RP/EB family member 3 (MAPRE3) The protein encoded by this gene is a member of the RP/EB family of genes. The protein localizes to the cytoplasmic microtubule network and binds APCL, a homolog of the adenomatous polyposis coli tumor suppressor gene. This gene has been identified as oncogenic drivers in lung adenocarcinomas in partnership with other genes [65].

ATP-binding cassette subfamily C member 3 (ABCC3) The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. Research suggests that ABCC3 expression may serve as a marker for multidrug resistance (MDR). MDR contributes to the failure of chemotherapy and high mortality in non-small cell lung cancer [68].

Claudin 4 (CLDN4) The protein encoded by this intronless gene belongs to the claudin family. Claudins are integral membrane proteins that are components of the epithelial cell tight junctions, which regulate movement of solutes and ions through the paracellular space. Study suggests that expression of Claudin 4 gene is frequently altered in human lung cancers [32].

Erb-b2 receptor tyrosine kinase 2 (ERBB2) This gene encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases. Amplification or over expression of this gene has been reported in numerous cancers, including lung tumors [14].

Fc fragment of IgG receptor and transporter (FCGRT) This gene encodes a receptor that binds the Fc region of monomeric immunoglobulin G. The encoded protein transfers immunoglobulin G antibodies from mother to fetus across the placenta. FCGRT gene is found as a biomarker in small-blue-round-cell tumor [48].

Caveolin 1 (CAV1) Caveolin 1 and caveolin 2 are located next to each other on chromosome 7 and express colocalizing proteins that form a stable hetero-oligomeric complex. Mutations in this gene have been associated with Berardinelli–Seip congenital lipodystrophy. In some studies, Caveolin 1 is found to be an informative gene for SRBCT [43].

CD99 molecule (Xg blood group) The protein encoded by this gene is a cell surface glycoprotein involved in leukocyte migration and T cell adhesion. Cases of small-blue-round-cell tumor are confirmed using CD99 gene [20].

Protein tyrosine phosphatase non-receptor type 13 (PTPN13) The protein encoded by PTPN13 is a member of the protein tyrosine phosphatase (PTP) family. PTPs are signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. Research shows PTPN13 to be an informative gene for SRBCT [43].

Major histocompatibility complex, class II, DM alpha (HLA-DMA) HLA-DMA belongs to the HLA class II alpha chain paralogues. It plays a central role in the peptide loading of MHC class II molecules by helping to release the CLIP molecule from the peptide binding site. This biomarker is discovered in the case of SRBCT in [48].

5.6 Comparison with benchmark algorithms

This section deals with the comparative analysis of the results obtained with the proposed hybrid gene selection method (ANOVA-EJFOA-SVM) along with other benchmark schemes. Some of the benchmark techniques justify their results using a testing dataset, or cross-validation or using both. Moreover, different algorithms use different sample sizes for validation. Various researchers have used

different approaches to validate their results, which implies that the comparisons cannot be definitive or precise. Nevertheless, a comparison gives an approximative measurement of the performance of the proposed scheme, and any technically sound work should include, as done in this article, a comparison to the best techniques available in the domain.

Tables 10 and 11 depict the comparative analysis of the results obtained with ANOVA-FOA-SVM, ANOVA-EJFOA-SVM, and other benchmark schemes, for all the well-known datasets. The values represent the CA along with the number of genes selected. The number of genes selected is shown in parenthesis. The number of genes selected with the proposed method is from the best result across ten runs. From Table 10, it is observed that for Colon cancer, not only does the proposed method outperform other algorithms achieving 96.90% accuracy, but only three genes are required to achieve this. For Ovarian cancer, our method is at par with DRFO-CFS technique resulting in 100% accuracy. However, DRFO-CFS achieves 100% result with 16 genes in contrast to only four genes by our approach. In case of Leukemia-2, the proposed approach

achieves 98.57% accuracy with four genes and comes second to an existing technique DFS which gives 98.61% accuracy.

Table 11 shows the performance of existing algorithms in terms of accuracy and the number of selected genes for multi-class datasets. Out of four multi-class datasets, the proposed method outperforms in three datasets, namely Leukemia-3, Lymphoma-3, and Lung cancer-5. For Leukemia-3, our result is 98.55% with only four genes which is at least 1% higher than the existing methods. In Lymphoma-3 dataset, selecting three genes by the proposed

Table 10 Comparative analysis of the proposed method with benchmark algorithms for binary-class datasets (columns 2 to 8); the symbol '–' denotes no data available

Methods	Datasets		
	Leukemia-2	Colon tumor	Ovarian cancer
BCGS [49]	94.1(35)	83.8(23)	98.8(26)
BDE-XRankf [5]	82.4(6)	75(4)	95(3)
DRFO-CFS [9]	91.18(13)	90(10)	**100(16)**
GEM [29]	91.5(3)	91.2(8)	–
IRLDA [53]	97(72)	–	–
IWSS [60]	94.4(7.9)	–	–
IWSS-MB-NB [60]	97.1(6.4)	86(5.2)	–
8-S PMSO [23]	98.1(20)	94.2(20)	–
AEN-CMI [62]	91.05(26.85)	89.30(25.20)	–
SLR [1]	95.51(7)	94.61(5)	–
DFS [51]	**98.61**	87.09	–
ANOVA-FOA-SVM	96(8)	94.44(6)	98.33(12)
ANOVA-EJFOA-SVM	98.57(4)	**96.90(3)**	**100(4)**

The best results are shown in bold font. Also the name of the proposed method is shown in bold font

Table 11 Comparative analysis of the proposed method with benchmark algorithms for multi-class datasets (columns 2 to 8) the symbol '–' denotes no data available

Methods	Datasets			
	Leukemia-3	Lymphoma-3	SRBCT	Lung cancer-5
mRMR-ABC [3]	96.12(20)	96.96(5)	96.30(10)	–
GBC [4]	95.83(8)	98.48(5)	96.38(6)	–
CC-PSO [10]	–	96.8(306)	93.7(63)	–
PSO-AKNN [33]	90.66(3.3)	–	94(8.5)	–
GALA [42]	93.96(3)	–	99.34(6)	–
MCSO [41]	–	–	71.04(100)	–
D-ECOC [38]	79.79	–	98.70	–
MGSACO [57]	–	–	74.49(20)	85.72(20)
DFS [51]	97.22	98.48	**100**	–
ANOVA-FOA-SVM	96(9)	98.04(10)	95.28(10)	92.55(8)
ANOVA-EJFOA-SVM	**98.55(4)**	**99.87(3)**	97.77(5)	**94.56(4)**

The best results are shown in bold font. Also the name of the proposed method is shown in bold font

approach leads to 99.87% accuracy. For Lung cancer-5, our method achieves 94.56% accuracy with only four genes. Only, in the case of SRBCT dataset, the results obtained with the DFS algorithm are better than the proposed method. From the above results, it is clear that the proposed learning algorithm is found to be the most suitable algorithm among all other algorithms in the context of CA and number of features selected.

6 Conclusion

In this article, a simple, yet efficient gene selection method using the principles of enhanced Jaya and FOA is proposed. Initially, a statistical filter is used which sorts the features according to their *p*-values; then, an EJFOA-based wrapper technique searches the optimal feature subset. An important aspect of the proposed method is that the initial population is generated with small feature size.

To evaluate the performance of the proposed technique, seven different binary-class and multi-class gene expression microarray datasets are used along with a well-known classifier, namely SVM. In addition, a comparative analysis is made with state-of-the-art techniques. From the extensive simulation and result analysis, it is observed that the proposed method is competitive and selects a minimum subset of genes with maximum accuracy for all the benchmark datasets. We must remark that our proposed approach is able to reduce the original size of feature set by more than 99% and simultaneously achieve highly robust results.

In the future, the authors will apply some advanced machine learning techniques, namely deep learning and reinforcement learning, for gene selection and classification.

Acknowledgements This research is partially supported by the following Grant: Grant No. SR/FST/ETI-335/2013 by Fund for Improvement of S&T Infrastructure in Higher Educational Institutions (FIST) Program of Department of Science and Technology, Government of India to International Institute of Information Technology, Bhubaneswar, Odisha, India.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Algamal ZY, Lee MH (2018) A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. In: Advances in data analysis and classification. Springer, Berlin, pp 1–19
2. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 96(12):6745–6750
3. Alshamlan H, Badr G, Alohali Y (2015a) mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. BioMed Res Int 2015:604910–604910
4. Alshamlan HM, Badr GH, Alohali YA (2015b) Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification. Comput Biol Chem 56:49–60
5. Apolloni J, Leguizamón G, Alba E (2016) Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 38:922–932
6. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ (2001) Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 30(1):41
7. Baliarsingh SK, Vipsita S, Muhammad K, Dash B, Bakshi S (2019) Analysis of high-dimensional genomic data employing a novel bio-inspired algorithm. Appl Soft Comput 77:520–532
8. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M et al (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci 98(24):13790–13795
9. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2015) Distributed feature selection: an application to microarray data classification. Appl Soft Comput 30:136–150
10. Chinnaswamy A, Srinivasan R (2016) Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data. In: Innovations in bio-inspired computing and applications. Springer, Cham, pp 229–239
11. Cho-Vega JH, Rassidakis GZ, Admirand JH, Oyarzo M, Ramalingam P, Paraguya A, McDonnell TJ, Amin HM, Medeiros LJ (2004) Mcl-1 expression in b-cell non-hodgkin's lymphomas. Hum Pathol 35(9):1095–1100
12. Chouhan SS, Kaul A, Singh UP (2018a) Soft computing approaches for image segmentation: a survey. Multimed Tools Appl 77(21):28483–28537
13. Chouhan SS, Kaul A, Singh UP, Jain S (2018b) Bacterial foraging optimization based radial basis function neural network (BRBFNN) for identification and classification of plant leaf diseases: an automatic approach towards plant pathology. IEEE Access 6:8852–8863
14. Chuang JC, Stehr H, Liang Y, Das M, Huang J, Diehn M, Wakelee HA, Neal JW (2017) Erbb2-mutated metastatic non-small cell lung cancer: response and resistance to targeted therapies. J Thorac Oncol 12(5):833–842
15. Collins CT, Hess JL (2016) Role of hoxa9 in leukemia: dysregulation, cofactors and essential targets. Oncogene 35(9):1090
16. Dashtban M, Balafar M (2017) Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics 109(2):91–107
17. Dwivedi AK (2018) Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Comput Appl 29(12):1545–1554
18. Elyasigomari V, Lee D, Screen H, Shaheed M (2017) Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. J Biomed Inform 67:11–20
19. Eskinazi R, Thöny B, Svoboda M, Robberecht P, Dassesse D, Heizmann CW, Van Laethem JL, Resibois A (1999) Overexpression of pterin-4a-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 in human colon cancer. Am J Pathol 155(4):1105–1113

20. Ezejiofor IF, Adelusola K, Durosinmi MA, Leoncini L, Odesanmi WO, Ambrosio MR, Lazzi S, Olaofe RO, Gbutorano G et al (2018) Immunohistochemical characterization of small round blue cell tumors of childhood at ile-ife, Nigeria: a 10-year retrospective study. Arch Med Health Sci 6(1):64
21. Galani E, Sgouros J, Petropoulou C, Janinis J, Aravantinos G, Dionysiou-Asteriou D, Skarlos D, Gonos E (2002) Correlation of mdr-1, nm23-h1 and h sema e gene expression with histopathological findings and clinical outcome in ovarian and breast cancer patients. Anticancer Res 22(4):2275–2280
22. García-Nieto J, Alba E (2012a) Parallel multi-swarm optimizer for gene selection in DNA microarrays. Appl Intell 37(2):255–266
23. García-Nieto J, Alba E (2012b) Parallel multi-swarm optimizer for gene selection in DNA microarrays. Appl Intell 37(2):255–266
24. Ghaemi M, Feizi-Derakhshi MR (2014) Forest optimization algorithm. Exp Syst Appl 41(15):6676–6687
25. Ghosh M, Guha R, Sarkar R, Abraham A (2019) A wrapper-filter feature selection technique based on ant colony optimization. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04171-3
26. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
27. Hall MA (1999) Correlation-based feature selection for machine learning. Doctoral dissertation, The University of Waikato
28. Heit C, Jackson BC, McAndrews M, Wright MW, Thompson DC, Silverman GA, Nebert DW, Vasiliou V (2013) Update of the human and mouse serpin gene superfamily. Hum Genom 7(1):22
29. Hernandez JCH, Duval B, Hao JK (2007) A genetic embedded approach for gene selection and classification of microarray data. In: European conference on evolutionary computation, machine learning and data mining in bioinformatics, Springer, pp 90–101
30. Ibrahim AO, Shamsuddin SM, Abraham A, Qasem SN (2019) Adaptive memetic method of multi-objective genetic evolutionary algorithm for backpropagation neural network. Neural Comput Appl. https://doi.org/10.1007/s00521-018-03990-0
31. Jothi G, Inbarani HH, Azar AT, Devi KR (2018) Rough set theory with jaya optimization for acute lymphoblastic leukemia classification. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3359-7
32. Jung JH, Jung CK, Choi HJ, Jun KH, Yoo J, Kang SJ, Lee KY (2009) Diagnostic utility of expression of claudins in non-small cell lung cancer: different expression profiles in squamous cell carcinomas and adenocarcinomas. Pathol Res Pract 205(6):409–416
33. Kar S, Sharma KD, Maitra M (2015) Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Exp Syst Appl 42(1):612–627
34. Kečo D, Subasi A, Kevric J (2018) Cloud computing-based parallel genetic algorithm for gene selection in cancer classification. Neural Comput Appl 30(5):1601–1610
35. Kim Y, Yoon S, Kim SJ, Kim JS, Cheong JW, Min YH (2012) Myeloperoxidase expression in acute myeloid leukemia helps identifying patients to benefit from transplant. Yonsei Med J 53(3):530–536
36. Lee CT, Chow NH, Su PF, Lin SC, Lin PC, Lee JC (2008) The prognostic significance of ron and met receptor coexpression in patients with colorectal cancer. Dis Colon Rectum 51(8):1268–1274
37. Li T, Zhang C, Ogihara M (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20(15):2429–2437
38. Liu KH, Zeng ZH, Ng VTY (2016) A hierarchical ensemble of ECOC for cancer classification based on multi-class microarray data. Inf Sci 349:102–118
39. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA) Protein Struct 405(2):442–451
40. Melhem R, Xx Zhu, Hailat N, Strahler JR, Hanash SM (1991) Characterization of the gene for a proliferation-related phosphoprotein (oncoprotein 18) expressed in high amounts in acute leukemia. J Biol Chem 266(27):17747–17753
41. Mohapatra P, Chakravarty S, Dash P (2016) Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. Swarm Evolut Comput 28:144–160
42. Motieghader H, Najafi A, Sadeghi B, Masoudi-Nejad A (2017) A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Inform Med Unlocked 9:246–254
43. Mukhopadhyay A, Bandyopadhyay S, Maulik U (2010) Multi-class clustering of cancer subtypes through svm based ensemble of pareto-optimal solutions for gene marker identification. PloS One 5(11):e13803
44. Nash MA, Deavers MT, Freedman RS (2002) The expression of decorin in human ovarian tumors. Clin Cancer Res 8(6):1754–1760
45. Niu Q, Zhang H, Li K (2014a) An improved TLBO with elite strategy for parameters identification of PEM fuel cell and solar cell models. Int J Hydrog Energy 39(8):3837–3854
46. Niu Q, Zhang L, Li K (2014b) A biogeography-based optimization algorithm with mutation strategies for model parameter estimation of solar and fuel cells. Energy Convers Manag 86:1173–1185
47. Orujpour M, Feizi-Derakhshi MR, Rahkar-Farshi T (2019) Multimodal forest optimization algorithm. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04113-z
48. Pal NR, Aguan K, Sharma A, Amari Si (2007) Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. BMC Bioinform 8(1):5
49. Pang S, Havukkala I, Hu Y, Kasabov N (2007) Classification consistency analysis for bootstrapping gene selection. Neural Comput Appl 16(6):527–539
50. Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC et al (2002) Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359(9306):572–577
51. Potharaju SP, Sreedevi M (2019) Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. Clin Epidemiol Glob Health 7(2):171–176
52. Rao R (2016) Jaya: a simple and new optimization algorithm for solving constrained and unconstrained optimization problems. Int J Ind Eng Comput 7(1):19–34
53. Sharma A, Paliwal KK, Imoto S, Miyano S (2014) A feature selection method using improved regularized linear discriminant analysis. Mach Vis Appl 25(3):775–786
54. Sharma S, Kaul A (2018) Hybrid fuzzy multi-criteria decision making based multi cluster head dolphin swarm optimized IDS for VANET. Veh Commun 12:23–38
55. Sheskin DJ (2003) Handbook of parametric and nonparametric statistical procedures. CRC Press, Boca Raton
56. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437
57. Tabakhi S, Najafi A, Ranjbar R, Moradi P (2015) Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing 168:1024–1036

58. Tang B, Xiang K, Pang M (2018) An integrated particle swarm optimization approach hybridizing a new self-adaptive particle swarm optimization with a modified differential evolution. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3878-2
59. Valdés-Mora F, Locke WJ, Bandrés E, Gallego-Ortega D, Cejas P, García-Cabezas MA, Colino-Sanguino Y, Feliú J, del Pulgar TG, Lacal JC (2017) Clinical relevance of the transcriptional signature regulated by cdc42 in colorectal cancer. Oncotarget 8(16):26755
60. Wang A, An N, Chen G, Yang J, Li L, Alterovitz G (2014a) Incremental wrapper based gene selection with Markov blanket. In: 2014 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 74–79
61. Wang X, Gotoh O (2009) Accurate molecular classification of cancer using simple rules. BMC Med Genom 2(1):64
62. Wang Y, Yang XG, Lu Y (2019) Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. Appl Math Model 71:286–297
63. Wang ZQ, Bachvarova M, Morin C, Plante M, Gregoire J, Renaud MC, Sebastianelli A, Bachvarov D (2014b) Role of the polypeptide n-acetylgalactosaminyltransferase 3 in ovarian cancer progression: possible implications in abnormal mucin o-glycosylation. Oncotarget 5(2):544
64. Yagasaki F, Wakao D, Yokoyama Y, Uchida Y, Murohashi I, Kayano H, Taniwaki M, Matsuda A, Bessho M (2001) Fusion of etv6 to fibroblast growth factor receptor 3 in peripheral t-cell lymphoma with at (4; 12)(p16; p13) chromosomal translocation. Cancer Res 61(23):8371–8374
65. Yakirevich E, Resnick MB, Mangray S, Wheeler M, Jackson CL, Lombardo KA, Lee J, Kim KM, Gill AJ, Wang K et al (2016) Oncogenic alk fusion in rare and aggressive subtype of colorectal adenocarcinoma as a potential therapeutic target. Clin Cancer Res 22(15):3831–3840
66. Yu K, Wang X, Wang Z (2016) An improved teaching-learning-based optimization algorithm for numerical and engineering optimization problems. J Intell Manuf 27(4):831–843
67. Zhao H, Sun Q, Li L, Zhou J, Zhang C, Hu T, Zhou X, Zhang L, Wang B, Li B et al (2019) High expression levels of aggf1 and mfap4 predict primary platinum-based chemoresistance and are associated with adverse prognosis in patients with serous ovarian cancer. J Cancer 10(2):397
68. Zhao Y, Lu H, Yan A, Yang Y, Meng Q, Sun L, Pang H, Li C, Dong X, Cai L (2013) Abcc3 as a marker for multidrug resistance in non-small cell lung cancer. Sci Rep 3:3120
69. Zhu Z, Ong YS, Dash M (2007) Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognit 40(11):3236–3248