Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 3165 – 3179
© TÜBİTAK
doi:10.3906/elk-2102-110

Gene expression data classification using genetic algorithm-based feature selection

Öznur Sinem SÖNMEZ[1,*], **Mustafa DAĞTEKİN**[1], **Tolga ENSARİ**[2]
[1]Department of Computer Engineering, Faculty of Engineering, Istanbul University,Cerrahpasa, Istanbul, Turkey
[2]Department of Computer and Information Science, College of Engineering and Applied Sciences,
Arkansas Tech University, Russellville, Arkansas, USA

Received: 22.02.2021 • **Accepted/Published Online:** 09.09.2021 • **Final Version:** 30.11.2021

Abstract: In this study, hybrid methods are proposed for feature selection and classification of gene expression datasets. In the proposed genetic algorithm/support vector machine (GA-SVM) and genetic algorithm/k nearest neighbor (GA-KNN) hybrid methods, genetic algorithm is improved using Pearson's correlation coefficient, Relief-F, or mutual information. Crossover and selection operations of the genetic algorithm are specialized. Eight different gene expression datasets are used for classification process. The classification performances of the proposed methods are compared with the traditional GA-KNN and GA-SVM wrapper methods and other studies in the literature. Classification results demonstrate that higher accuracy rates are obtained with the proposed methods compared to the other methods for all datasets.

Key words: Feature selection, gene expression datasets, hybrid method, genetic algorithm, support vector machine, cancer classification

1. Introduction

Microarray and RNA sequencing are the technologies that allow analysis and quantification of gene expression levels. With these technologies, studies are carried out on the identification of cancer-related genes and gene profiling of diseases. Cancer-healthy tissue or cancer subtypes can be differentiated by evaluation of the gene expressions. Thousands of gene expressions need to be evaluated together for cancer diagnosis and classification. Due to a large amount of interrelated data in this process, computer algorithms and artificial intelligence methods are needed to make the process more efficient and the results more accurate.

Gene expression datasets include 2000–60000 genes and their feature sizes are greater than the number of samples. Therefore, dimensionality reduction methods have a significant role in classification of these datasets. These methods are divided into feature selection and feature extraction methods. Feature extraction methods reduce dimension by transforming the available data into new features. Feature selection methods determine group of features that best reflect data. The features in the ideal subset to be selected should have a high relevance to the class but a low relevance to each other. These methods help to improve performance metrics and decrease computation time. Filter, wrapper, embedded, and hybrid methods are subcategories of feature selection methods.

Filter methods are generally statistical-based methods that do not involve learning. Subset selection is usually performed by measuring the relationship of features to the classes. As a result, the redundant features may not be eliminated. On the other hand, these methods are practical to implement and the results can be

[*]Correspondence: osonmez@ogr.iu.edu.tr

achieved quickly. They are usually preferred as the preprocess methods in the analysis and classification of gene expressions in the literature. Thus, the number of genes is reduced to 500–1000 level. Pearson's correlation coefficient, mutual information, and Relief-F [1] are some of the most preferred examples of filter methods.

Wrapper methods consist of a classifier and the search algorithm. The most successful subset is selected with evaluation of the classification accuracy. The goal is optimizing classification accuracy or an equation related to accuracy for selected subset. For a high-dimensional dataset with N number of features, the evaluation of $(2^N - 1)$ subsets will result in considerable computational load. Thus, heuristic or evolutionary algorithms are preferred for subset selection. Wrapper methods consider the dependencies and relationship between genes. Although their computational complexity is higher than filter methods, the accuracy of wrapper methods is much better.

The most commonly used algorithms for subset selection are ant colony optimization (ACO), genetic algorithm, and particle swarm optimization (PSO) algorithms. In much of the current research, SVM and KNN methods are used as a classifier in combination with these algorithms to form GA-SVM [2, 3], GA-KNN [4], PSO-SVM [3, 5], PSO-KNN [5], ACO-SVM [6] methods. For instance, Alba et al. [3] compared the classification accuracy of GA-SVM and PSO-SVM and developed a geometric particle swarm optimization–support vector machine (GPSO-SVM) method. Kar et al. [5] proposed the PSO-AKNN method by combining particle swarm optimization with adaptive k nearest neighbor (AKNN). Adaptive genetic algorithm (AGA) is used for subset selection in the AGA-KNN [7] wrapper method. Arunkumar et al. [8] proposed the GA-ELM method by using an extreme learning machine (ELM) as the classifier.

Embedded methods use the classifier to establish a criterion for ranking the features. The results depend on the classifier as the feature selection is performed in the classifier training. Since the selected features are dependent on the classifier, their performance may not be the same in different classifiers [9]. One of the most well-known embedded methods is SVM-RFE [10] that includes recursive feature elimination (RFE). In [10], SVM classifier weight of feature is used to rank genes. Mundra and Rajapakse [11] introduced a method that combines minimum redundancy maximum relevancy (MRMR) [12] and SVM-RFE to improve its performance. Turgut et al. [13] used RFE and randomized logistic regression for feature elimination and compared the classification performances of different classifiers. Luo et al. [14] proposed a method to consider correlation between features and intrinsic properties by improving SVM-RFE method with F-statistic, distance correlation coefficient, and Pearson's correlation coefficient-based correlation metric.

Hybrid methods consist of wrapper and filter methods. The purpose is to optimize the performance by taking advantage of each method. In hybrid methods, generally, the number of genes is reduced to around 500–1000 using filter methods, and then gene subsets are selected with the wrapper method. In the literature, there are various hybrid method studies such as those combining Relief-F, MRMR, and GA-wrapper methods [15], using multiple filter and multiple wrapper methods [16], combining dynamic parameter genetic algorithm (GADP) and chi-square test [17], using information gain in combination with micro-GA-SVM [18] or with binary krill herd algorithm [19], and performing feature selection with chaotic harmony search (CHS) after MRMR filter method [20]. In addition, there are also hybrid methods that use various optimization algorithms and evolutionary operators without including filter methods. For example, Othman et al. [21] proposed a hybrid method by combining multiobjective cuckoo search with mutation and crossover evolutionary operators (MOCS-EO) to improve the exploration ability of the algorithm. Meenachi et al. [22] hybridized GA and ACO with tabu search. Qaraad et al. [23] used different optimization algorithms to determine elastic net (EN) with

SVM.

In this study, GA-KNN and GA-SVM hybrid methods are proposed in which the genetic algorithm is improved with the Pearson correlation coefficient, Relief-F, or mutual information. In the proposed methods, the crossover and selection operations of the genetic algorithm are specialized. The goal of the proposed approach is to determine best group of genes for maximizing the accuracy rate of the classification and minimizing the number of selected genes. In Section 2, filter methods that are used to improve genetic algorithm are described briefly and proposed methods are accounted. In Section 3, datasets and performance metrics are shown and the classification results of proposed methods are compared with the wrapper methods and other studies in the literature by using eight different gene expression datasets. Experimental results demonstrate that accuracy rates of the proposed methods are higher than the wrapper methods and other studies in the literature for all datasets. Section 4 contains the summary of this study and future research directions.

2. Methods

2.1. The Pearson correlation coefficient

The Pearson correlation rates the linear dependence between the feature and the class. It is defined as in Equation 1, where N is the sample size, X^i is the i^{th} feature, and Y is the target value.

$$r = \frac{cov(X^i, Y)}{\alpha_X \alpha_Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}} = \frac{\sum_{j=1}^{N}(x_j-\bar{x})(y_j-\bar{y})}{\sqrt{\sum_{j=1}^{N}(x_j-\bar{x})^2(y_j-\bar{y})^2}} \tag{1}$$

This method allows ranking the relationship between the genes and the class. Thus, the number of genes can be reduced by eliminating the low-ranked genes. Moreover, similar genes can be determined by calculating the degree of relationship between genes.

2.2. Relief-F

Relief-F [1] is a multivariate method that aims to determine features which best distinguish each class. It is a version of the Relief algorithm and developed for multiclass datasets. It gives better results in the noisy and missing data [24]. The main objective of this method is to rate each feature according to its ability to distinguish nearest samples. The algorithm selects an arbitrary x instance, determines the nearest k instances from same (a) and different (b) classes. The distances between random x and nearest k samples from same class and different classes are calculated. Accordingly, the degree of the feature decreases as the distance of the x instance to the instances of the same class increases. It increases as the distance between x and instances of different classes increases.

The degree of f feature $R(f)$ can be expressed as in Equation 2, where $R(f)_a$ represents the distance between random x and its nearest neighbors with same class, n is the number of random selection.

$$R(f) = R(f)_a - R(f)_b \tag{2}$$

$$R(f)_a = \frac{1}{n \cdot k}\sum_{i=1}^{k} diff(f, x, a_i) \tag{3}$$

$$R(f)_a = \sum_{s \neq s_x} \frac{P(s)}{[1-P(s_x)]n \cdot k}\sum_{i=1}^{k} diff(f, x, b_i(s)) \tag{4}$$

$R(f)_b$ represents the distance between x and its nearest neighbors with different classes and calculated together for all other classes (s) except the class of x. $P(s_x)$ indicates the probability of the class of x, $P(s)$ indicates the probability of the class of s.

2.3. Mutual information

Mutual information (MI) indicates the measure of the dependence between two random variables. It is calculated as in Equation 5.

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{5}$$

In Equation 5, X and Y are random variables, $p(x)$ indicate the probability distribution of X, $p(y)$ indicate the probability distribution of Y, and $p(x,y)$ represents the joint probability distribution.

2.4. Proposed methods

In the proposed methods, the genetic algorithm in GA-KNN and GA-SVM methods is improved by the filter feature selection method. The Pearson correlation, Relief-F, or mutual information methods are used as filter methods. The flowchart that shows the main structure of the proposed approach is shown in Figure 1.

In the genetic algorithm, chromosomes are encoded by using binary coding. The size of each chromosome is the same with feature size of the dataset and individuals are randomly generated. In parent selection, 20% of the current population with the best fitness scores is determined as the elite parent population. They are guaranteed to survive and to be selected as parents. Other parents are selected by stochastic universal selection (SUS) from the remaining population based on scaled fitness scores. The scaling is performed to convert the range of the fitness scores into more appropriate range for SUS method. The calculated scaled value is proportional to $1/\sqrt{n}$ where n is the rank of the individual that corresponds to its position after sorting fitness scores in descending order.

Crossover is performed in two different ways by using logical AND and logical OR operations. AND operation is intended to transfer the genes in both parents to the next generation. The OR operation is used to provide diversity. Before crossover by OR, a certain percentage of the genes with the lowest score obtained from the filter method is converted to zero in both parents. Thus, the diversity can be maintained at a certain level and the results can be achieved faster. Two offsprings are generated from each pair of parents in crossover operation, one from logical AND and the other from logical OR operation. After crossover of the parents, the child chromosomes and the elite parents are transferred to the next generation. The parents other than elite parents are deleted from the population to maintain the population size.

The AccDimensionScore function shown in Equation 6 is determined as a function of the classification accuracy obtained from the classifier and the number of genes on the chromosomes. It is aimed to maximize the accuracy of classification and minimize the number of selected genes. KNN and SVM are used as a classifier in the proposed methods.

$$AccDimensionScore(x) = \alpha\Big(1 - Acc(x)\Big) + (1-\alpha)\frac{s(x)}{t} \tag{6}$$

In Equation 6, $Acc(x)$ represents the accuracy of individual x, $s(x)$ represents the number of genes that are 1 in x, t corresponds to the total number of genes, and α indicates the coefficient whose value varies in the

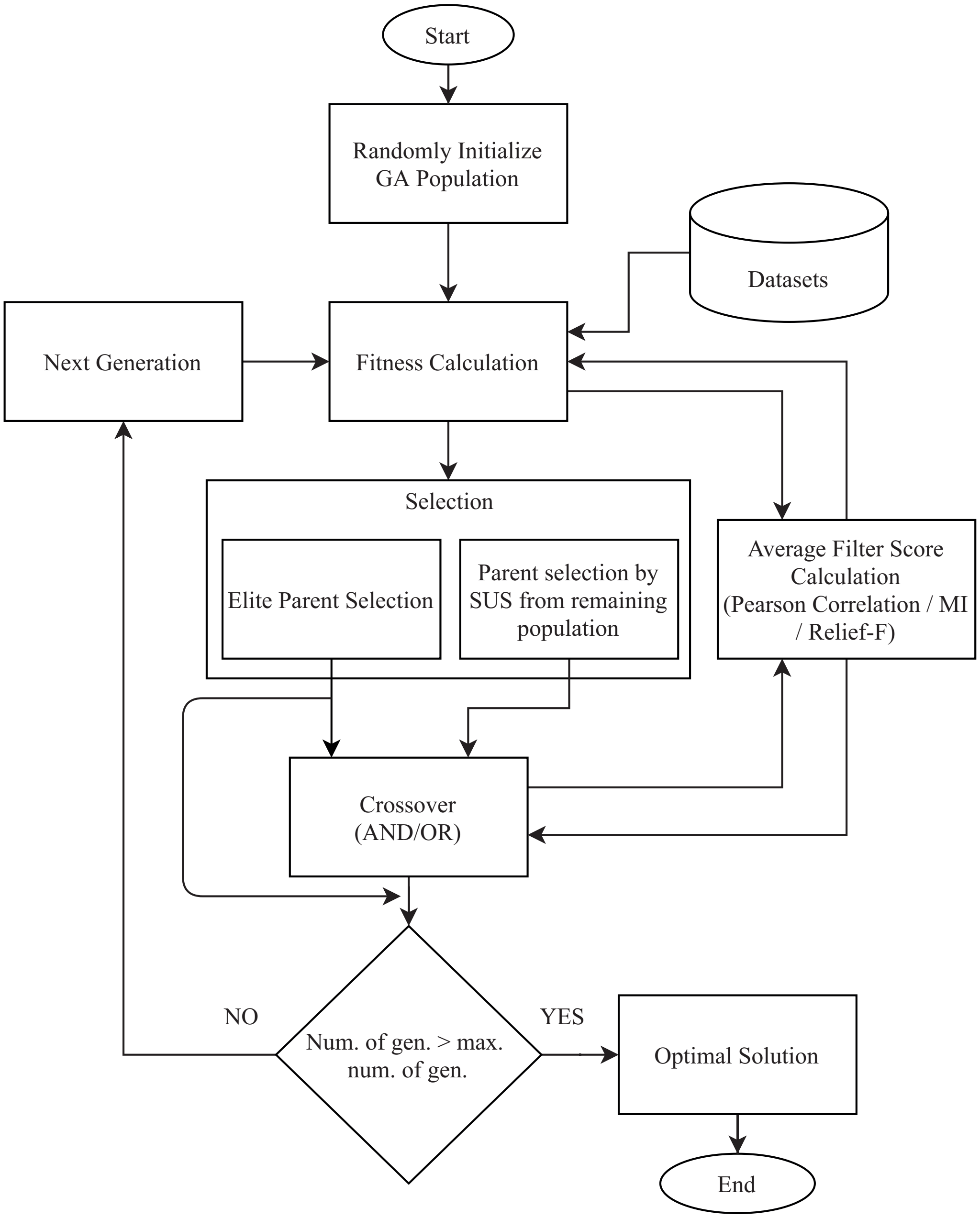

Start
Randomly Initialize GA Population
Datasets
Next Generation
Fitness Calculation
Selection
Elite Parent Selection
Parent selection by SUS from remaining population
Average Filter Score Calculation (Pearson Correlation / MI / Relief-F)
Crossover (AND/OR)
NO
YES
Num. of gen. > max. num. of gen.
Optimal Solution
End

Figure 1. The flowchart of the proposed method.

range [0,1].

In the proposed methods, average score of each chromosome where each gene is associated with the class is calculated by using the Pearson correlation, Relief-F, or mutual information scores. Thus, the filter method score of chromosomes is used together with the AccDimensionScore function values in the selection process.

Average score of each individual is calculated as:

$$AvgScore(x) = \frac{\sum_{i=1}^{t} g(i)Score(i)}{s(x)} \tag{7}$$

In Equation 7, $g(i)$ corresponds to 0 or 1 that represents if the i^{th} gene of individual x is selected or not, and $Score(i)$ is the filter method score of i^{th} gene in individual x.

The fitness function that corresponds to the total score of each chromosome can be expressed as:

$$Fitness(x) = \beta(AccDimensionScore(x)) + (1 - \beta)(1 - AvgScore(x)) \tag{8}$$

where $AvgScore(x)$ represents the average score value of the genes that are 1 in individual x, β represents the coefficient that takes value in the range [0,1].

As seen in Figure 1, there is no mutation process in the proposed methods owing to the diversity provided by the OR operation.

3. Experimental analyses

In this section, performances of the proposed methods are evaluated with a 10-fold cross-validation (CV) protocol by 8 datasets containing gene expressions. Moreover, the details of datasets and performance metrics are presented.

In Table 1, genetic algorithm parameters are shown. The parameters of genetic algorithm are determined according to the optimization problem to avoid convergence to the local optimum. The population size should be chosen by considering the balance of genetic diversity and computational load. If the population size is too small, it may cause premature convergence due to insufficient genetic diversity. If it is too large, it leads to high computational load. Similarly, the percentage of elitism should be at a level that allows to transfer the individuals with the best scores of the population to the next generation and at the same time does not reduce the diversity. The maximum number of generations should be large enough to achieve the optimal solution. By considering these trade-offs, the parameters are determined by evaluating various test results for different datasets and methods to achieve the best results in terms of fitness score and accuracy. The best results are obtained with the values shown in Table 1. Accordingly, the α and β coefficients of fitness function are set to 0.9 and 0.8, respectively.

Table 1. Parameters of genetic algorithm in the proposed methods.

Parameters	Value
Population size	100
Crossover	AND(50%), OR (50%)
Elite rate	20%
Selection method	Stochastic universal selection
Maximum number of generations	200

GA is a stochastic evolutionary algorithm. Therefore, the experimental results are calculated as the mean values of the results obtained from 10 runs of each method for each dataset. The number of selected gene results are average results of 10 runs rounded to the nearest integer.

3.1. Datasets

Datasets that are used in the comparison of the methods are shown in Table 2. These datasets are among the main datasets frequently used in computer-assisted cancer diagnosis and classification studies in the literature. BRCA and COAD datasets include gene expression levels determined by the RNA sequencing method and others include gene expression levels obtained by the microarray method.

The datasets shown in Table 2 consist of two classes and are mostly composed of the samples from healthy and cancer patients. Most of these datasets are used for cancer-healthy tissue or cancer subtype classification. Moreover, some of them are used for the estimation of survival and determining whether the cancer recurred.

Table 2. Datasets and their properties.

Datasets	Classification task	Number of genes	Number of samples	Number of classes	Class distribution
Colon cancer [25]	Normal samples Cancer samples	2000	62	2	Normal: 22 Cancer: 40
Prostate cancer [26]	Normal samples Cancer samples	12,600	136	2	Normal: 59 Cancer: 77
DLBCL [27]	Diffuse large B-cell lymphoma (DLBCL) Follicular lymphoma (FL)	7129	77	2	DLBCL: 58 FL: 19
Breast cancer [28]	Luminal type (L) Nonluminal type (NL)	47,293	128	2	L: 84 NL: 44
Breast cancer 2 [29]	Relapse (distance metastases within 5 years) Nonrelapse (NR)	24,481	97	2	Relapse: 46 NR: 51
CNS [30]	Survivor (Class 1) Failure (Class 0)	7129	60	2	Class 0: 39 Class 1: 21
BRCA[1]	Normal samples Cancer samples	60,483	1211	2	Normal: 113 Cancer: 1098
COAD[1]	Normal samples Cancer samples	60,483	430	2	Normal: 40 Cancer: 390

3.2. Performance metrics

The number of selected genes, accuracy (Acc.), precision (Pre.), negative predictive ratio (NPR), specificity (Spec.), F1 score and sensitivity (Sens.) are used as performance metrics. These performance metrics can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$NPR = \frac{TN}{TN + FN} \tag{11}$$

[1] NCI (2021). Genomic Data Commons Data Portal [online]. Website https://portal.gdc.cancer.gov [accessed: 2 February 2020]

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{13}$$

$$F1score = \frac{2 \times (Precision \times Sensitivity)}{Precision + Sensitivity} \tag{14}$$

True positive (TP) indicates positive samples that are assigned to positive. False positive (FP) represents negative samples that are assigned to positive. Similarly, true negative (TN) and false negative (FN) correspond to positive and negative samples, respectively, for samples assigned to negative.

3.3. Experimental results

The classification accuracy rates of the proposed and wrapper methods and the number of genes selected with these methods are shown in Table 3. According to these results, it is observed that the highest accuracy rates and the minimum number of selected gene combinations for all datasets are obtained by the proposed methods.

Table 3. Accuracy of classification (%) and the number of selected genes for all datasets.

Methods	Colon cancer	Prostate cancer	DLBCL	Breast cancer	Breast cancer (2)	CNS	BRCA	COAD
GA-KNN	83.87(171)	80.00(4064)	86.25(1853)	70.00(1977)	81.54(21338)	65.00(1945)	90.91(1892)	96.51(28,901)
GA(Pearson)-KNN	88.33 (3)	94.00(8)	96.25(2)	77.00(22)	90.00(304)	76.67(3)	97.44(8)	98.37(19)
GA(Relief-F)-KNN	86.67(1)	86.00(1)	100.00(82)	77.00(20)	85.38(24)	81.67(196)	98.18(30)	97.91(2)
GA(MI)-KNN	86.67(1)	85.00(1)	97.50(2)	80.00(9)	83.85(274)	73.33(1)	98.68(433)	98.14(1)
GA-SVM	91.67(208)	89.00(3931)	98.75(1814)	58.00(9080)	83.08(21103)	73.33(1807)	99.09(28,773)	99.07(28,877)
GA(Pearson)-SVM	95.00(20)	97.00(275)	100.00(101)	71.00(21)	89.23(216)	93.33(28)	99.01(529)	99.30(219)
GA(Relief-F)-SVM	98.33(33)	99.00(210)	100.00(214)	67.00(376)	90.77(227)	95.00(100)	99.34(625)	99.07(124)
GA(MI)-SVM	98.33(49)	97.00(437)	100.00(38)	82.00(352)	90.00(396)	90.00(119)	99.01(629)	97.91(1)

The best results are obtained with the GA(Relief-F)-SVM method for five datasets except DLBCL, breast cancer and COAD datasets. The traditional GA-KNN wrapper method outperforms GA-SVM for breast cancer dataset. In contrast, the best accuracy rate for breast cancer dataset is obtained by the proposed GA(MI)-SVM method. This indicates that the proposed methods are more stable and robust against changes in data. For COAD dataset, GA(Pearson)-SVM method achieves the best accuracy rate. Relief-F method outperforms the other filter methods for five datasets.

The proposed methods that include KNN outperform the traditional GA-KNN wrapper method for all datasets and GA-SVM method for five datasets. They also significantly reduce the number of selected genes compared to the traditional wrapper methods.

In Table 4, the performances of the proposed methods and other studies in the literature are compared. The classification accuracy rates of the proposed methods are higher than those of the other methods. For colon cancer, breast cancer, CNS, and prostate cancer datasets, the GA(Relief-F)-SVM method achieves the best accuracy rates. For breast cancer dataset, this method achieves 90.77% accuracy rate with 227 genes. For prostate cancer dataset, its accuracy rate is 99% with 210 genes. For colon cancer and CNS datasets, its accuracy rates are 98.33% and 95%, respectively. For DLBCL dataset, 100% accuracy rate is obtained by most of the methods, including the proposed methods. However, the GA (MI)-SVM method achieves 100% with only 38 genes, the number of selected genes was not specified in the other studies.

Table 4. Comparison of methods (accuracy % and the number of selected genes).

Methods	Colon cancer	Prostate cancer	DLBCL	Breast cancer (2)	CNS
Bootstrapped margin [2]	92.40	-	-	-	-
3-fold CV					
GA-ELM [8]	-	-	-	84.00 (9)	-
10-fold CV					
ACO-S-SVM [6]	81.42 (69)	-	-	-	-
10-fold CV					
Max-Min ACO-SVM [31]	95.00 (10.8)	-	100	-	-
Leave one out CV					
PSO-SVM [32]	94.0	-	-	-	-
Leave half out CV					
F-test + GA-SVM [4]	85	92.68	84	-	81.25
5-fold CV					
SNR + GA-SVM [4]	95	65.85	100	-	81.25
5-fold CV					
MIM + AGA-ELM [33]	89.09	96.54	-	-	-
IG + BBO-RF [34]	92.34 (11)	-	-	-	-
10-fold CV					
IG-SVM [35]	90.32	96.08	100	-	-
10-fold CV					
[36]	87 (4)	-	-	-	88.66 (6)
Leave one out CV	91.95 (5)				
FLD-NRS [37]	88.0 (6)	80 (4)	-	-	-
MOCEPO [38]	96.74	-	-	-	-
10-fold CV					
IG+micro GA -SVM [18]	-	-	-	-	92.86 (29)
IG-MBKH [19]	96.47	-	-	-	90.34
10-fold CV					
MRMR+CHS+KNN[20]	80.64 (72)	-	96.10 (61)	-	-
10-fold CV					
MOCS-EO [21]	-	-	-	-	76.30 (952)
10-fold CV					
GA(Pearson)-SVM	95.00 (20)	97.00 (275)	100.00 (101)	89.23 (216)	93.33 (28)
GA(Relief-F)-SVM	98.33(33)	99.00(210)	100.00 (214)	90.77(227)	95.00(100)
GA(MI)-SVM	98.33 (49)	97.00 (437)	100.00(38)	90.00 (396)	90.00 (119)
GA(Pearson)-KNN	88.33 (3)	94.00 (8)	96.25 (2)	90.00 (304)	76.67 (3)
GA(Relief-F)-KNN	86.67 (1)	86.00 (1)	100.00 (82)	85.38 (24)	81.67 (196)
GA(MI)-KNN	86.67 (1)	85.00 (1)	97.50 (2)	83.85 (274)	73.33 (1)

The classification results obtained with the 10-fold cross-validation protocol for all performance metrics are presented in Tables 5–7. According to these results, the proposed methods improve all performance metrics for all datasets. They particularly increase accuracy, F1 score and decrease the number of selected genes. For example, for CNS dataset, highest F1 score obtained with traditional wrapper methods is 0.57, while the proposed methods increase F1 score to 0.92. Similarly, in the breast cancer dataset results in Table 6, it is observed that F1 score is increased from 0.69 to 0.80. In addition, the proposed methods decrease the number of selected genes by an average of 95% compared to the traditional methods.

Table 5. Classification results of CNS, colon cancer, and DLBCL datasets.

Datasets	Methods	Acc.%	Pre.%	Sens.%	Spec.%	F1	NPR%	Number of genes
CNS	GA-KNN	65.00	50.00	33.33	82.05	0.40	69.57	1945
	GA(Pearson)-KNN	76.67	70.59	57.14	87.18	0.63	79.07	3
	GA(Relief-F)-KNN	81.67	77.78	66.67	89.74	0.71	83.33	196
	GA(MI)-KNN	73.33	66.67	47.62	87.18	0.55	75.56	1
	GA-SVM	73.33	64.71	52.38	84.62	0.57	76.74	1807
	GA(Pearson)-SVM	93.33	94.74	85.71	97.44	0.90	92.68	28
	GA(Relief-F)-SVM	95.00	100.00	85.71	100.00	0.92	92.86	100
	GA(MI)-SVM	90.00	89.47	80.95	94.87	0.85	90.24	119
Colon cancer	GA-KNN	83.87	80.00	72.73	90.00	0.76	85.71	171
	GA(Pearson)-KNN	88.33	85.71	81.82	92.11	0.84	89.74	3
	GA(Relief-F)-KNN	86.67	88.89	72.73	94.74	0.80	85.71	1
	GA(MI)-KNN	86.67	88.89	72.73	94.74	0.80	85.71	1
	GA-SVM	91.67	90.48	86.36	94.74	0.88	92.31	208
	GA(Pearson)-SVM	95.00	91.30	95.45	94.74	0.93	97.30	20
	GA(Relief-F)-SVM	98.33	100.00	95.45	100.00	0.97	97.44	33
	GA(MI)-SVM	98.33	100.00	95.45	100.00	0.97	97.44	49
DLBCL	GA-KNN	86.25	100.00	81.67	100.00	0.89	64.52	1853
	GA(Pearson)-KNN	96.25	98.31	96.67	95.00	0.97	90.48	2
	GA(Relief-F)-KNN	100.00	100.00	100.00	100.00	1.00	100.00	82
	GA(MI)-KNN	97.50	98.33	98.33	95.00	0.98	95.00	2
	GA-SVM	98.75	100.00	98.33	100.00	0.99	95.24	1814
	GA(Pearson)-SVM	100.00	100.00	100.00	100.00	1.00	100.00	101
	GA(Relief-F)-SVM	100.00	100.00	100.00	100.00	1.00	100.00	214
	GA(MI)-SVM	100.00	100.00	100.00	100.00	1.00	100.00	38

The convergence graphs of GA(Relief-F)-SVM method over different datasets are shown in Figures 2 and 3. The graphs are obtained by sample runs of the method for corresponding dataset. It is seen that the accuracy increases as the number of generations increases, and the number of selected genes is generally in a downward trend. The fluctuations in the number of selected genes indicate that there is genetic diversity in the population, and the algorithm is in the global and local search process. Despite the fluctuations, the number of selected genes decreases and converges to a certain value for all datasets within 40–60 number of generations.

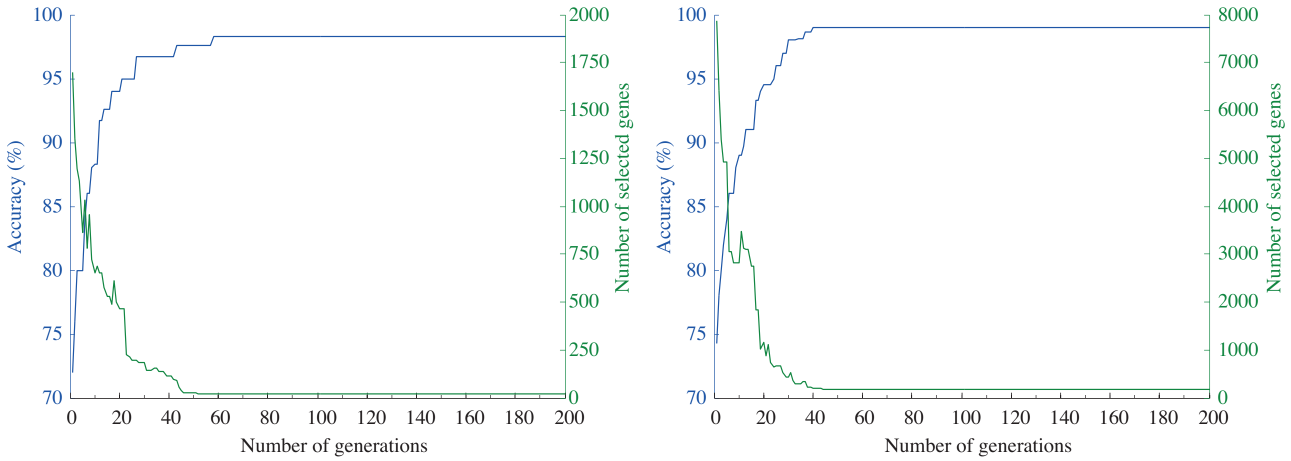
The convergence graphs also show that the accuracy tends to increase as the number of selected genes decreases. The fluctuations in the number of selected genes do not disrupt the increasing trend of accuracy. Moreover, different gene configurations with the same number of genes can result in different accuracy rates as the algorithm continues to search for the optimal solution. Therefore, accuracy continues to increase after the number of selected genes converges to a value as seen in Figures 2a, 2b, and 3b.

Table 6. Classification results of breast cancer, breast cancer (2), and prostate cancer datasets.

Datasets	Methods	Acc.%	Pre.%	Sens.%	Spec.%	F1	NPR%	Number of genes
Breast cancer	GA-KNN	70.00	69.77	63.83	75.47	0.66	70.18	8977
	GA(Pearson)-KNN	77.00	77.27	72.34	81.13	0.74	76.79	22
	GA(Relief-F)-KNN	77.00	87.50	59.57	92.45	0.70	72.06	20
	GA(MI)-KNN	80.00	80.00	76.60	83.02	0.78	80.00	9
	GA-SVM	58.00	60.87	29.79	83.02	0.40	57.14	9080
	GA(Pearson)-SVM	71.00	68.75	70.21	71.70	0.69	73.08	21
	GA(Relief-F)-SVM	67.00	75.00	44.68	86.79	0.56	63.89	376
	GA(MI)-SVM	82.00	80.85	80.85	83.02	0.80	83.02	352
Prostate cancer	GA-KNN	80.00	84.44	74.51	85.71	0.79	76.36	4064
	GA(Pearson)-KNN	94.00	94.12	94.12	93.88	0.94	93.88	8
	GA(Relief-F)-KNN	86.00	84.91	88.24	83.67	0.86	87.23	1
	GA(MI)-KNN	85.00	84.62	86.27	83.67	0.85	85.42	1
	GA-SVM	89.00	90.00	88.24	89.80	0.89	88.00	3931
	GA(Pearson)-SVM	97.00	98.00	96.08	97.96	0.97	96.00	275
	GA(Relief-F)-SVM	99.00	100.00	98.04	100.00	0.99	98.00	210
	GA(MI)-SVM	97.00	98.00	96.08	97.96	0.97	96.00	437
Breast cancer (2)	GA-KNN	81.54	79.25	97.67	50.00	0.87	91.67	21,338
	GA(Pearson)-KNN	90.00	91.01	94.19	81.82	0.92	87.80	304
	GA(Relief-F)-KNN	85.38	86.02	93.02	70.45	0.89	83.78	24
	GA(MI)-KNN	83.85	84.95	91.86	68.18	0.88	81.08	274
	GA-SVM	83.08	85.56	89.53	70.45	0.87	77.50	21,103
	GA(Pearson)-SVM	89.23	90.00	94.19	79.55	0.92	87.50	216
	GA(Relief-F)-SVM	90.77	91.11	95.35	81.82	0.93	90.00	227
	GA(MI)-SVM	90.00	90.11	95.35	79.55	0.92	89.74	396

Table 7. Classification results of COAD and BRCA datasets.

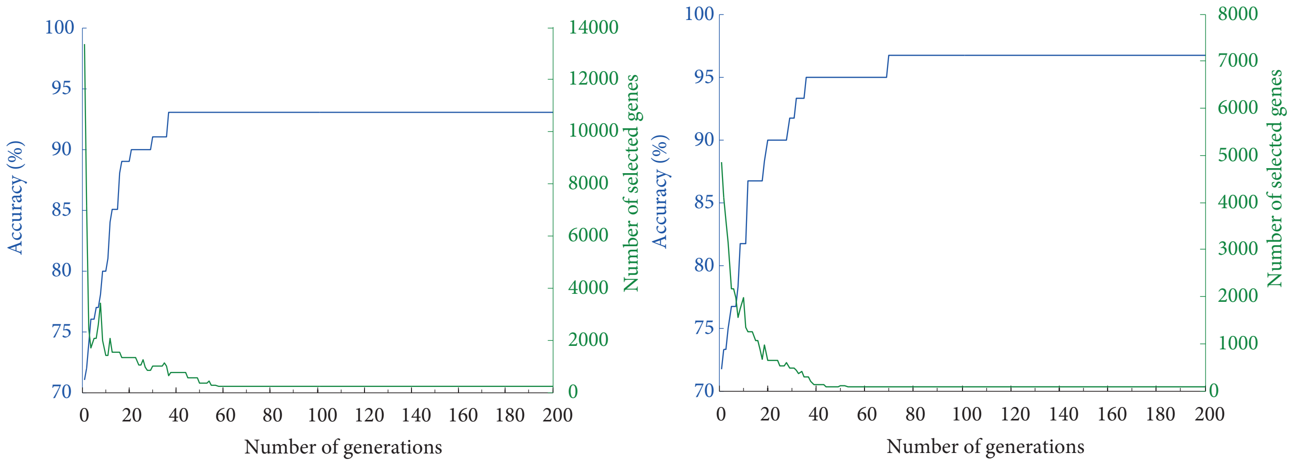
Datasets	Methods	Acc.%	Pre.%	Sens.%	Spec.%	F1	NPR%	Number of genes
COAD	GA-KNN	96.51	97.47	98.72	75.00	0.98	85.71	28,901
	GA(Pearson)-KNN	98.37	98.98	99.23	90.00	0.99	92.31	19
	GA(Relief-F)-KNN	97.91	98.47	99.23	85.00	0.98	91.89	2
	GA(MI)-KNN	98.14	98.97	98.97	90.00	0.98	90.00	1
	GA-SVM	99.07	99.74	99.23	97.50	0.99	92.86	28,877
	GA(Pearson)-SVM	99.30	99.49	99.74	95.00	0.99	97.44	219
	GA(Relief-F)-SVM	99.07	99.74	99.23	97.50	0.99	92.86	124
	GA(MI)-SVM	97.91	98.47	99.23	85.00	0.98	91.89	1
BRCA	GA-KNN	90.91	93.94	96.17	39.82	0.95	51.72	1892
	GA(Pearson)-KNN	97.44	98.28	98.91	83.19	0.98	88.68	8
	GA(Relief-F)-KNN	98.18	99.09	98.91	91.15	0.99	89.57	30
	GA(MI)-KNN	98.68	99.27	99.27	92.92	0.99	92.92	433
	GA-SVM	99.09	99.82	99.18	98.23	0.99	92.50	28,773
	GA(Pearson)-SVM	99.01	99.73	99.18	97.35	0.99	92.44	529
	GA(Relief-F)-SVM	99.34	100.00	99.27	100.00	0.99	93.39	625
	GA(MI)-SVM	99.01	99.63	99.27	96.46	0.99	93.16	629



(a) Colon cancer dataset

(b) Prostate cancer dataset

Figure 2. The convergence graphs of the GA(Relief-F)-SVM method in terms of accuracy and the number of selected genes for colon cancer and prostate cancer datasets.



(a) Breast cancer (2) dataset

(b) CNS dataset

Figure 3. The convergence graphs of the GA(Relief-F)-SVM method in terms of accuracy and the number of selected genes for breast cancer (2) and CNS datasets.

4. Conclusion

In this study, the genetic algorithm is improved using filter feature selection methods to form GA-KNN and GA-SVM hybrid methods for classification of the gene expression datasets. The Pearson correlation coefficient, Relief-F, and mutual information are used as the filter methods to improve and specialize the selection and crossover operations of the genetic algorithm. The classification performances of the proposed methods are evaluated by using eight different gene expression datasets. It is observed that the proposed methods improve all performance metrics. Moreover, the highest accuracy rates and F1 score values for all datasets are obtained with the proposed methods. In the six proposed methods, the highest results are obtained using the methods that include SVM as the classifier. The GA(Relief-F)-SVM method achieves the highest accuracy rates for five among eight datasets. The Relief-F method outperforms the other filter methods.

In future studies, different optimization methods and classifiers can be used to analyze and compare the results. Other metaheuristic methods such as particle swarm optimization can be evaluated instead of the genetic algorithm.

References

[1] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: Bergadano F, De Raedt L. (editors). Machine Learning: ECML-94. Berlin, Germany: Springer, 1994, pp. 171-182.

[2] Chen XW. Margin-based wrapper methods for gene identification using microarray. Neurocomputing 2006; 69 (16-18): 2236-2243. doi: 10.1016/j.neucom.2005.07.007

[3] Alba E, Garcia-Nieto J, Jourdan L, Talbi EG. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In: IEEE Congress on Evolutionary Computation; Singapore, Singapore; 2007. pp. 284-290. doi: 10.1109/CEC.2007.4424483

[4] Gunavathi C. Premalatha K. Performance analysis of genetic algorithm with kNN and SVM for feature selection in tumor classification. International Journal of Computer and Information Engineering 2014; 8 (8): 1490- 1497. doi: 10.5281/zenodo.1096103

[5] Kar S, Sharma KD, Maitra M. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications 2015; 42 (1): 612-627. doi: 10.1016/j.eswa.2014.08.014

[6] Li Y, Wang G, Chen H, Shi L, Qin L. An ant colony optimization based dimension reduction method for high-dimensional datasets. Journal of Bionic Engineering 2013; 10: 231-241. doi: 10.1016/S1672-6529(13)60219-X

[7] Lee CP, Lin WS, Chen YM, Kuo BJ. Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Systems with Applications 2011; 38 (5): 4661-4667. doi: 10.1016/j.eswa.2010.07.053

[8] Arunkumar C, Sooraj MP, Ramakrishnan SMP. Finding expressed genes using genetic algorithm and extreme learning machines. In: International Conference on Advanced Computing and Communication Systems; Coimbatore, India; 2017. pp. 1-4. doi: 10.1109/ICACCS.2017.8014609

[9] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics 2015. doi: 10.1155/2015/198363

[10] Guyon I, Weston J, Barhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002; 46: 389-422. doi: 10.1023/A:1012487302797

[11] Mundra PA, Rajapakse JC. SVM-RFE with MR filter for gene selection. IEEE Transactions on Nanobioscience 2010; 9 (1): 31-37. doi: 10.1109/TNB.2009.2035284

[12] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 2005; 3 (2): 185-205. doi: 10.1142/S0219720005001004

[13] Turgut S, Dağtekin M, Ensari T. Microarray breast cancer data classification using machine learning methods. In: Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT); Istanbul, Turkey; 2018, pp. 1-3. doi: 10.1109/EBBT.2018.8391468

[14] Luo K, Wang G, Li Q, Tao J. An improved SVM-RFE based on F-statistic and mPDC for gene selection in cancer classification. IEEE Access 2019; 7: 147617-147628. doi: 10.1109/ACCESS.2019.2946653

[15] Shreem SS, Abdullah S, Nazri MZA, Alzaqebah M. Hybridizing reliefF, MRMR filters, and GA wrapper approaches for gene selection. Journal of Theoretical and Applied Information Technology 2012; 46 (2): 1034-1039.

[16] Leung Y, Hung Y. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2010; 7 (1): 108-117. doi: 10.1109/TCBB.2008.46

[17] Lee CP, Leu Y. A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing 2011; 11 (1): 208-213. doi: 10.1016/j.asoc.2009.11.010

[18] Pragadeesh C, Jeyaraj R, Siranjeevi K, Abishek R, Jeyakumar J. Hybrid feature selection using micro genetic algorithm on microarray gene expression data. Journal of Intelligent & Fuzzy Systems 2019; 36 (3): 2241-2246. doi: 10.3233/JIFS-169935

[19] Zhang G, Hou J, Wang J, Yan C. Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. Interdisciplinary Sciences: Computational Life Sciences 2020; 12: 288-301. doi: 10.1007/s12539-020-00372-w

[20] Wang A, Liu H, Chen G. Chaotic harmony search based multi-objective feature selection for classification of gene expression profiles. In: IEEE 9th International Conference on Bioinformatics and Computational Biology (ICBCB); Taiyuan, China; 2021. pp. 107-112. doi: 10.1109/ICBCB52223.2021.9459222

[21] Othman MS, Kumaran SR, Yusuf LM. Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data. IEEE Access 2020; 8: 186348-186361. doi: 10.1109/ACCESS.2020.3029890

[22] Meenachi L, Ramakrishnan S. Metaheuristic search based feature selection methods for classification of cancer. Pattern Recognition 2021; 119: 108079. doi:10.1016/j.patcog.2021.108079

[23] Qaraad M, Amjad S, Manhrawy IIM, Fathi H, Hassan BA et al. A hybrid feature selection optimization model for high dimension data classification. IEEE Access 2021; 9: 42884-42895. doi: 10.1109/ACCESS.2021.3065341

[24] Khadijah, Rismiyati, Mantau AJ. Multiclass classification of cancer based on microarray data using extreme learning machine, In: 1st International Conference on Informatics and Computational Sciences; Semarang, Indonesia; 2018. pp. 159-164. doi: 10.1109/ICICOS.2017.8276355

[25] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 1999; 96 (12): 6745-6750. doi: 10.1073/pnas.96.12.6745

[26] Singh D, Febbo PG, Ross K, Jackson DG, Manola J et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002; 1 (2): 203-209. doi: 10.1016/S1535-6108(02)00030-2

[27] Shipp M, Ross K, Tamayo P, Weng A, Kutok J et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 2002; 8: 68-74. doi: 10.1038/nm0102-68

[28] Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR et al. A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene 2007; 26: 1507-1516. doi: 10.1038/sj.onc.1209920

[29] Veer LJ, Dai H, Vijver M, He YD, Hart A et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-536. doi: 10.1038/415530a

[30] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M et al. Prediction of central nervous system embryonal tumor outcome based on gene expression. Nature 2002; 415: 436-442. doi: 10.1038/415436a

[31] Yu H, Gu G, Liu H, Shen J, Zhao J. A modified ant colony optimization algorithm for tumor marker gene selection. Genomics, Proteomics & Bioinformatics 2009; 7 (4): 200-208. doi: 10.1016/S1672-0229(08)60050-9

[32] Shen Q, Shi WM, Kong W, Ye BX. Combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta 2007; 71 (4): 1679-1683. doi: 10.1016/j.talanta.2006.07.047

[33] Lu H, Chen J, Yan K, Jin Q, Xue Y et al. A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 2017; 256: 56-62. doi: 10.1016/j.neucom.2016.07.080

[34] Nikumbh S, Ghosh S, Jayaraman VK. Biogeography-based informative gene selection and cancer classification using SVM and random forests. In: IEEE Congress on Evolutionary Computation; Brisbane, QLD, Australia; 2012. pp. 1-6. doi: 10.1109/CEC.2012.6256127

[35] Gao L, Ye M, Lu X, Huang D. Hybrid method based on information gain and support vector machine for gene selection in cancer classification. Genomics, Proteomics & Bioinformatics 2017; 15 (6): 389-395. doi: 10.1016/j.gpb.2017.08.002

[36] Peng Y, Li W, Liu Y. A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification. Cancer Informatics 2006; 2: 301-311. doi: 10.1177/117693510600200024

[37] Sun L, Zhang X, Xu J, Wang W, Liu R. A Gene selection approach based on the fisher linear discriminant and the neighborhood rough set. Bioengineered 2018; 9 (1): 144-151. doi: 10.1080/21655979.2017.1403678

[38] Baliarsingh SK, Vipsita S, Muhammad K, Bakshi S. Analysis of high-dimensional biomedical data using an evolutionary multi-objective emperor penguin optimizer. Swarm and Evolutionary Computation 2019; 48: 262-273. doi: 10.1016/j.swevo.2019.04.010