Improved gene selection-based classification for DNA microarray data using multi-gene genetic programming

Sara Sfaksi 1* , Leila Djerou 2

12 LESIA Laboratory, Computer science Department, Mohamed Khider University.

P.O Box: 145RP- 07000 Biskra –Algeria
1* sara.sfaksi@univ-biskra.dz, 2 l.djerou@univ-biskra.dz

*Corresponding Author:
Sara Sfaksi,
PHD student,
LESIA Laboratory,
Computer science Department,
Mohamed Khidher University
P.O Box: 145RP- 07000 Biskra –Algeria
E-mail: sara.sfaksi@univ-biskra.dz
Tel: +213 676124919

Abstract

DNA microarray is a technology that allows researchers to measure and analyze the expression levels of large numbers of genes in a specific tissue under various conditions. The main objective of microarray analysis is to classify biological samples and predict treatment efficacy or resistance in certain diseases, such as cancer. However, the curse of dimensionality makes it difficult to create prediction models utilizing gene expression patterns. Thus, finding useful genes has always been crucial in interpreting microarray data because there are many irrelevant and inconsequential genes. This paper presents a new hybrid model that integrates feature selection in DNA microarrays with multi-gene genetic programming through an interactive complexity-efficiency trade-off that characterizes the multi-gene genetic programming technique. The proposed model can find the best combination of predictor gene expression in the datasets by learning from the DNA microarray dataset. Four well-known cancer datasets and six-prediction accuracy metrics (Correlation Coefficient $\boldsymbol{R^2}$*, RMSE, SSE, MAE and MAXE) were used to evaluate the performance of the proposed method. The results show good prediction efficiency with four induction methods: SVM, KNN, Ada-boost and Naive bays.*

Keywords: *DNA microarray, Gene expression, Feature selection, Cancer classification, Machine learning, Multi-gene genetic programming.*

1. Introduction

Cancer is a disease that occurs due to changes in DNA [1]. These changes can contribute to the uncontrolled multiplication of cells and the subsequent development of a tumour. Since most large genes contain several areas where mutations can occur, it cannot be

easy to construct a test to identify them. Therefore, to find or tailor a case-specific diagnosis, laboratories are turning to genomic analysis, which is used to study and characterize genomic information in the cell's DNA. This study will make it possible to predict the efficacy or resistance of treatment, in particular by chemotherapy and/or targeted therapy, the production of an accurate diagnosis by characterizing the genomic signature, the prediction of the evolution of the disease and the prescription of a targeted treatment when available. DNA microarray analysis is one of the fastest-growing new technologies in the genetic research plot [2]. The scientific community uses this tool because of its great potential to simultaneously measure the expression level of a large number of genes in tissue samples [3], allowing researchers to obtain a global view of the gene regulatory network and determine which ones are expressed in a specific tissue under different conditions [4]. The study of cancer development has made extensive use of microarrays. This technique allows genotyping of several areas of a genome or measuring the expression levels of a large number of genes simultaneously. In addition, gene copy number changes and methylation patterns are identified [5]. In addition, clinicians use these findings in personalized medicine to select cancer treatment based on the patient's cancer genetic profile.

Classification and analysis of microarray data can accelerate diagnosis, prognosis, and treatment regimens [8]. However, the unique characteristics of microarray data present the statistics and data mining community with a real challenge. Given that several thousand to tens of thousands of gene expression levels are measured for a relatively small number of experiments, the high dimensionality of microarray databases is the main cause of this problem [4]. Indeed, among all these genes, many of them are irrelevant, insignificant or redundant for the discrimination problem [7].

Therefore, to analyse or classify microarray data, it is necessary to minimize their dimensionality by selecting a collection of appropriate genes that maintain or increase the initial classification accuracy. In DNA microarrays, this problem is known as gene selection [9, 10]. It is, therefore, both essential and useful to identify discriminating genes. The best genes can be examined to confirm recent advances in cancer research or offer new directions for future research in biology and medicine [6]. The gene selection problem can be defined as an NP-hard optimization problem. It consists in selecting a subset of genes from a large number n of genes. Indeed, selecting the "right" subset of genes requires the consideration of 2n-1 potential subsets.

Optimization metaheuristics, such as evolutionary algorithms, are frequently used to solve this problem because they can explore a large solution space efficiently and handle noisy and incomplete data, as well as nonlinear and non-convex optimization problems, which are commonly encountered in gene selection problems. Moreover, metaheuristics can also identify gene inter- actions and relationships that may be missed by other methods. This is particularly important in complex diseases and conditions where multiple genes may be involved. Overall, metaheuristics are a powerful tool for solving gene selection problems in DNA microarrays, offering an efficient and effective approach for identifying relevant genes associated with specific diseases or conditions. Metaheuristic algorithms are frequently used as wrapper methods to guide the search process for gene selection [11]. In this process, varying classification algorithms are frequently used as a fitness evaluation to deter- mine the subset of genes. In some cases, the filter method was introduced first as a pre-processing step to filter noise [11]. However, the selected genes cannot improve classification performance when applying another classifier that is different from the one used as fitness evaluation.

Unlike all previous methods based on classifier-dependent metaheuristics, we propose in this paper, a new method based on multi-gene genetic programming (MGGP) as an evolutionary learning technique inspired by natural evolution, allowing to identify a set of dissimilar genes that are most closely correlated to the target class, without using any classification model. From a training set, MGGP finds good gene combination models by simultaneously optimizing two competing goals: fitting the models with respect to their correlation with target class and complexity of their structure. The remainder of this paper is organized as follows: Section 2 reviews previous gene selection approaches, Section 3 introduces multi-gene genetic programming (MGGP); Section 4 describes the proposed method for gene selection using MGGP; Section 5 discusses experimental results, and Section 5 concludes the paper.

2. Related work

The classification of cancer DNA microarray data is an important and complex task in bioinformatics and medical research. The main goal of this task is to accurately predict the presence or absence of cancer based on gene expression level obtained from microarray data [12]. Ensuring a robust and reliable classification results require a careful consideration of the data, an appropriate selection of methods and algorithms and rigorous validation. One of the main challenges in classification of cancer in DNA microarray is dealing with the high dimensionality of the data which can lead to overfitting of the data and affect the accuracy of classification. Addressing these challenges need to pass by feature selection step in order to identify a subset of genes that are most relevant for classification.

In the last decade, the problem of feature (gene) selection has been the subject of much research in the field of microarray data analysis [9]. These studies are mainly based on filter and wrapper-based algorithms [10]. Filtering methods evaluate genes using statistical measures without a learning algorithm [8, 10]. Therefore, these techniques are fast, easy to understand and implement [9], however, they are based on fixed criteria and cannot adapt to changes in the data or problem. The wrapper approach evaluates the relevance of the selected gene set using learning algorithms. This approach has shown better results than the first one despite the long selection process due to its extensive computation [10] but also it can be computationally intensive and may suffer from overfitting or underfitting. . In the literature, various filter-based and wrapper-based algorithms are used for feature selection in many research areas. Filter-based approaches include Fisher score (FS) [13], Laplacian score (LS) [14] and RELIEF [15]. In addition, some wrapper-based algorithms include Improved Salp Swarm Algorithm (ISSA) [16], Binary Weight Sallow Swarm Optimization (BWSSO) [17] and Quantum Whale Optimization Algorithm (QWOA) [18].

In contrast, while filter and wrapper methods are useful techniques for gene selection, metaheuristics such as evolutionary algorithms are preferred for their efficiency, adaptability, and ability to handle complex, large-scale datasets Several feature selection methods address the problem of selecting informative genes for the classification of microarray data. The classical Bat algorithm is extended with improved formulations, efficient multi-objective operators, and new search techniques in [4], which proposes a new bio-inspired multi-objective algorithm for gene selection.

In [7], a group of four extensions of recursive machine feature elimination (SVM-RFE) called (MSVM-RFE) was proposed to solve the multi-class gene selection problem. A new gene selection method based on community detection and node centrality techniques was

proposed in [8] to select a set of different and most correlated genes. The testing data is not used to evolve the models and serves to give an indication of how well the models generalise to new data. The study in [9] presented a new gene selection approach based on kernel Fisher discriminant analysis (KFDA). To filter out significant genes, the authors of [19] suggested a hybridization of the Adaptive Elastic Net (AEN) algorithm with conditional mutual information (AEN-CMI), which improves AEN by incorporating conditional mutual information into the gene selection process. The methodology proposed in [20] combines both information gain (IG) and standard genetic algorithm (SGA) such that attributes (genes) are selected by IG and then reduced by GA. The authors in [21] introduced a new criterion, LS Bound, to address the problem of gene selection. This method can be seen as hybridization between the filter and wrapper methods. On the one hand, the LS Bound measure is derived from the leave-one-out procedure of LS-SVM. Authors in [22] proposed a novel method for gene selection in Random Forest-Based classification problems.

3. Multigene genetic programming (MGGP)

MGGP, a novel subset of genetic programming (GP), operates according to Darwin's principles of natural selection, which favour the best and eliminate the worst [23] . The main advantage of this evolutionary technique is its ability to present the relationship between a collection of inputs and their associated outputs in a simple mathematical form that is accessible to users. All individuals in the population are represented as an empirical mathematical model in the form of a weighted linear combination of several GP trees, combining the ability of the GP norm to build the model structure with the ability of traditional regression in estimating the parameters.

Each multi-gene individual consists of one to Gmax genes, denoted as a structure of "traditional" GP tree [23] with a max depth Dmax, that receives a set of input terminals X_j where j = 1, . . . J (expression values in our case) mapped via mathematical operators inserted in nodes to predict an output variable $\hat{y}$. Moreover, each prediction of the output variable $\hat{y}$ is formed by the weighted output of each of the trees/genes in the multi-gene individual plus a bias term [23]. The mathematical form of the multi-gene representation is shown as follows:

$$\sum_{i=0}^{n} d_i \; G_i \; + \; d_0 \qquad (1)$$

Where:

- n is the number of genes contained in the multi-gene individual
- G_i is the i^{th} gene value
- d_i is the i^{th} gene weight, and d_0 is the bias term.

3.1. MGGP parameters

The capacity of the model to be developed by MGGP is affected by the selection of some control parameters. Some of these parameters are the followings: the number of iterations, the maximum number of allowed genes in an individual G_{max}, maximum tree depth D_{max}, crossover and mutation event probabilities and selection method and mathematical functions (+, -, cos, sin) which can represent the nodes of subtrees. The choice of these control factors has an impact on the model that MGGP can generate: The number of

individuals in the population is fixed by the population size, the number of generations indicates how many times the algorithm is used before it succeeds, the complexity of the problems is frequently correlated with the size of the population and the number of generations. Increasing the G_{max} value and D_{max} value increases the training data's fitness value, while the test data's fitness value decreases due to the overfitting of the training data.

3.3. MGGP process

After defining control parameters, the MGGP process will be implemented in two steps: initialization and evolution. In the first step, an initial population of individuals is randomly created based on user-defined functions and variables. In the second step, a process of a few steps will be repeated until reaching specific criteria: calculate the fitness value of each individual, select the best ones as parents, reproduce new individuals by genetic operators (crossbreeding, mutation and selection) and finally, replace weaker parents with stronger ones.

4. Proposed method

In this study, the MGGP technique was used to pick out the best attribute/- gene selection model. The MGGP algorithm is executed on the training data and checked on the test data. The optimal model is selected based on its simplicity and performance on the learning data. So, applying MGGP in gene selection requires the description of the learning data, the representation of a solution and objective functions to be optimized. *Figure 1* shows the block diagram of the proposed method for choosing the best gene subset from a dataset. Below is a presentation of the suggested algorithms' specifics.

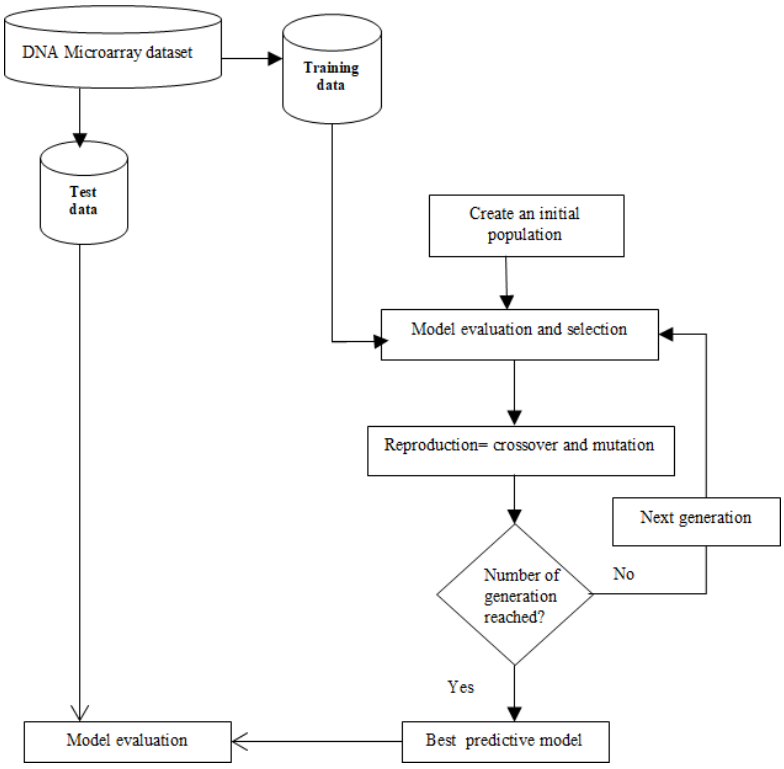

Figure 1. Block diagram of the proposed method

4.1. Learning data

For the training data, we were interested in microarray datasets describing the level of gene expression measured on two types of tissue:

- Normal samples have a y = 1 label
- Cancerous samples have a y = −1 label.

4.2. Objective functions

Two objective functions (f_1 and f_2) are used to choose the best gene selection model:

1. Minimizing model complexity f_1 is defined as the sum of the nodes of a tree's subtrees.

2. Maximizing model performance, defined by its quality of adjustment, which is represented by the coefficient R^2 where:

$$\boldsymbol{R^2 = \frac{\sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=0}^{n} (y_i - \bar{y})^2}} \qquad (\mathbf{2})$$

where: n is the number of the total example,
y_i is the example value,
$\hat{y}_i$ is the predicted value,
$\bar{y}$ is the example average.

Or minimizing f_2, where:

$$\boldsymbol{f_2 = 1 - R^2} \qquad (\mathbf{3})$$

4.3. Initial population

After learning data and objective functions definition, an initial population of N solution (model) is generated randomly. For each model, node functions and gene numbers (attributes) representing the leaves are randomly chosen from a set of operators defined by the user and a set of gene numbers in the dataset, respectively.

4.4. Evaluation and selection of models, reproduction and stopping criteria

Based on the NSGA-2 Multi-objective Optimization Algorithm (Non- Dominated Sorting Genetic Algorithm 2), solutions (models) are ranked according to their dominance ranks to choose those in the first fronts as parents. A new population of parents (p_t +1) is formed by adding the entire first fronts (first front F_1, second front F_2) as long as these do not exceed half the population size (N/2).

If the number of individuals present in (p_t+1) is less than (N/2), a crowding procedure is applied on the first following edge (F_i), not included in (p_t +1). The purpose of this operator is to insert the (N/2 - |(p_t + 1|) best individuals that are missing in the population (p_t +1). The

new population of children is then produced by crossover and mutation operators. The MGGP process is stopped after a predefined iteration number.

4.5 Evaluation of the models obtained

The model's performance is then evaluated in terms of its ability to maximize the coefficient of determination R_2 and reduce the test error RMSE. A P-value statistical measure is used to evaluate the result (model obtained by MGGP). In addition, the frequency of these genes in the best models is applied to prove the selected genes' importance in the developed population.

5. Experiments and discussion

An open-source software platform for symbolic data mining in MATLAB, GPTIPS 2 [24], was used to develop the proposed MGGP-based model for feature selection in DNA Microarrays.

Four well-known public datasets (*Table* 1) were selected for learning data. These datasets have been used in many works concerning microarray data analysis (datasets description can be founded on [25, 26]).

MGGP parameters were determined experimentally by observing the convergence of the objective functions over the generations. The parameters presented in (*Table* 2) are considered the best after their application on all datasets: prostate cancer (Pc102), Acute leukemias (AMLALL), Diffuse Large B-cell Lymphoma (DLBCL) and B-cell Lymphomas (BCL7129).

After 200 iterations on each dataset, four prediction models were selected as the best. (*Table* 3) lists the number of genes of the microarray datasets after and before MGGP model.

Table 1. Microarray datasets description

Dataset	Description	Ngenes	Nsamples	Nclasses
Pc102	Prostate Cancer dataset: 51 normal samples and 52 tumor samples.	12600	102	2
AMLALL	Acute Leukemias dataset: acute myeloid leukemia (AML, 25 samples) and acute lymphoblastic leukemia (ALL, 47 samples).	7129	72	2
DLBCL	Diffuse Large B-cell Lymphoma: 24 samples of germinal center B-like type and 23 samples of activated B-like type.	4026	47	2
BCL7129	B-cell Lymphomas (BCL): 58 samples of Diffuse Large B-cell Lymphoma (DLBCL) and 19 samples of Follicular Lymphoma (FL).	7129	77	2

Table 2. Optimal MGGP control parameters values.

Parameters	Value
Population size	300
Number of generations	200
Gmax	08
Dmax	04
Learning data	90%
Test data	10%
Crossover probability	0.8
Mutation probability	0.1
Pareto tournament	0.3

Table 3. Number of gene selected by MGGP models.

Dataset	gBefore	gAfter
PC102	12600	24
AMLALL	7129	25
DLBCL	4026	31
BCL7129	7129	26

For the used datasets, the accurate predictive models are presented in (*Table* 4). *Figure* 2 depicts the population of evolved models in each dataset, based on their complexity and fitness value, in blue circles, with green circles considered optimal. The red circles represent the best models.

Table 4. MGGP predictive models.

Dataset	MGGP Model
PC102	6.1*x2624 - 1.4*x2360 - 1.4*x160 + 0.42*x3572 - 5.4*x4587 - 0.54*x5862 + 1.3*x6592 + 0.42*x7230 - 1.1*x7757 + 0.69*x7857 - 4.0*x8507 + 1.4*x9598 + 0.69*x11845 - 0.72*x12513 + 1.6*x900*x1482 + 1.6*x1444*x2624 - 1.6*x1444*x4587 - 1.6*x2624*x4866 - 1.4*x2624*x5862 + 1.6*x4587*x4866 + 1.4*x4587*x5862 - 5.7*x2624*x8507 + 10.0*x4587*x8507 - 1.4*x2624*x11600 - 1.2*x4866*x9465 - 1.6*x4587*x10249 + 1.6*x5765*x9228 + 1.4*x4587*x11600 - 0.27*x2624^2 - 1.7*x900*x8507*x10249 - 1.7*x900*x10249*x11600 - 1.4*x4866*x8507*x9598 - 0.054
AMLALL	0.56*x2725 - 1.3*x1455 - 0.63*x2288 - 0.63*x1152 - 2.1*x4141 -

	0.63*x4535 + 0.56*x5024 - 0.26*x5170 - 1.9*x2482*x5107 + 0.56*x4494*(2.0*x5107 + x4535*x5392) - 0.63*x5107*(2.0*x557 + x3566) - 0.94*x2482*(2.0*x5107 + x98*x833*x3550*x5107*x6901*(x5024 + x5170 - 1.9)) + 4.0*x1455*x4535*x5107 + 1.9*x557*x4141*(x2482 + x3545 + x4*x1455) - 0.2*x4535*x5107^2*x6901*(x3289 + x4999 + x6423 + x6917 - 9.6) + 3.1
DLBCL	0.42*x130 + 0.42*x172 + 0.42*x188 + 0.68*x267 + 0.42*x721 + 0.41*x951 + 1.1*x1276 + 0.82*x1705 + 0.41*x1709 - 0.59*x2200 + 0.42*x2325 + 0.39*x2460 - 0.42*x3146 + 0.42*x3379 + 0.42*x3675 + 0.42*x3734 + 0.42*x3829 - 0.42*x3922 + 0.27*x267*x1276 - 0.27*x267*x2460 + 0.26*x1276*x3379 - 0.26*x2460*x3379 - 0.68*x674*x2200^2 - 0.41*x1674*x2200^2 - 0.41*x2200^2*x3539 - 1.6*x238*x816*x1246 + 1.6*x238*x1246*x2200 + 0.1*x267*x1276*x3379 - 0.1*x267*x2460*x3379 - 0.41*x1674*x2200*x2460 - 0.41*x2200*x2460*x3539 - 30.0*x205*x238*x267*x654*x951*x1009*x1674*x3279 - 8.7*x205*x238*x267*x951*x1009*x1674*x3279*x3852 - 2.2
BCL7129	- 2.0*x4741*x2149^3 + 0.26*x355 - 0.26*x566 - 0.23*x613 + 0.48*x1073 + 0.48*x1185 - 0.46*x1302 - 0.23*x2677 + 0.029*x3629 + 0.26*x4571 - 0.23*x4741 - 2.0*x5784 - 3.1*x5983 + 0.48*x6015 + 0.26*x6719 - 0.23*x250*x3290 + 1.5*x355*x5983 - 0.69*x540*x5983 + 1.5*x786*x5983 + 0.69*x1204*x5983 + 0.84*x4486*x5983 + 2.6*x1302*x1642*x3027 - 2.6*x1302*x3027*x3632 - 0.23*x1420*x4486*x4741 + 2.6*x1204*x1302*x1642*x3027 + 1.1

For each prediction model, we calculate the following evaluation metrics to prove the effectiveness of our proposed method in the gene selection problem: R^2, RMSE (Root mean Square Error), MSE (mean squared error), SSE (Sum of squared errors), MAE (Mean absolute error) and MAXE (Max absolute error) on the learning data and also on the test data (*Table* 5).

Figure 2. Population of evolved models in Pareto terms of complexity and fitness.

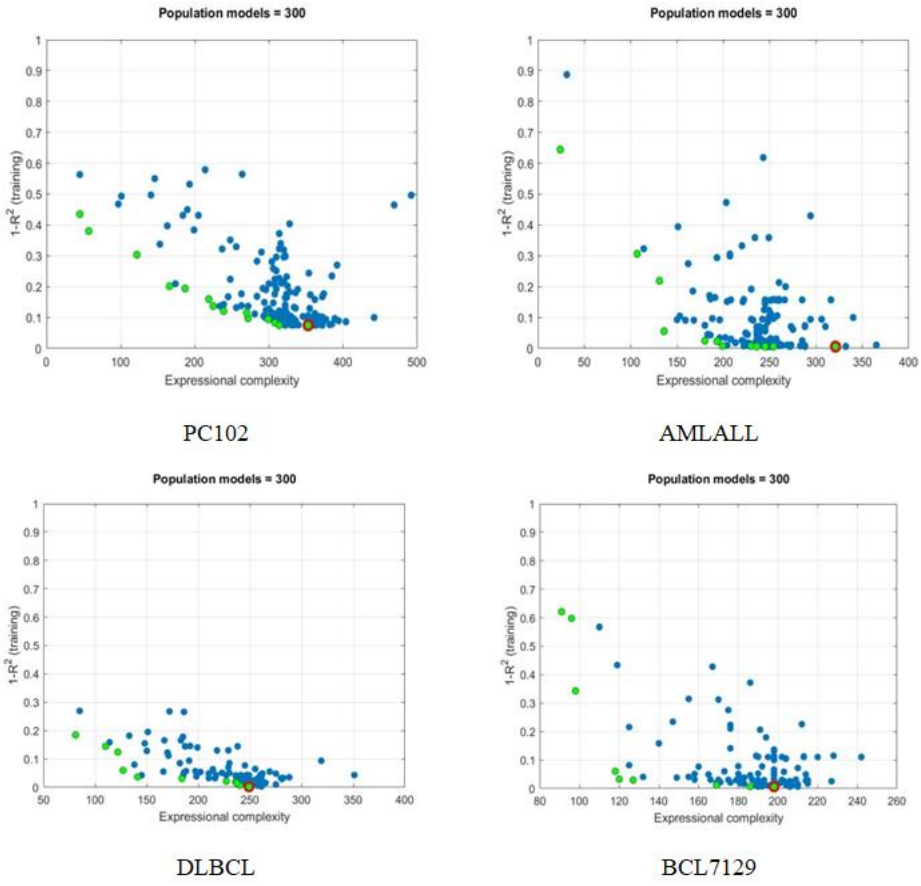

Table 5 . Evaluation metrics values.

Datasets	R^2	RMSE	MSE	SSE	MAE	MAXE
PCa102-Train	0.92	0.27	0.07	6.10	0.17	1.60
PCa102-Test	0.76	0.47	0.22	12.01	0.36	1.21
AMLALL -Train	0.99	0.07	0.005	0.22	0.06	0.14
AMLALL -Test	0.73	0.49	0.24	6.89	0.37	1.49
DLBCL -Train	0.99	0.04	0.002	0.05	0.03	0.08
DLBCL -Test	0.74	0.50	0.25	0.25	0.39	1.01
BCL712-Train	0.99	0.06	0.004	0.21	0.05	0.26
BCL712-Test	0.78	0.39	0.15	4.56	0.26	1.26

It can be noted that evaluation metrics values are very interesting for the five databases (learning and testing) since the R^2 values are close to 1 and the RMSE, MSE and MAE values close to 0. Also, whenever R^2 is maximized, and RMSE is minimized, we get a good correlation between real and predicted values (*Figure* 3).

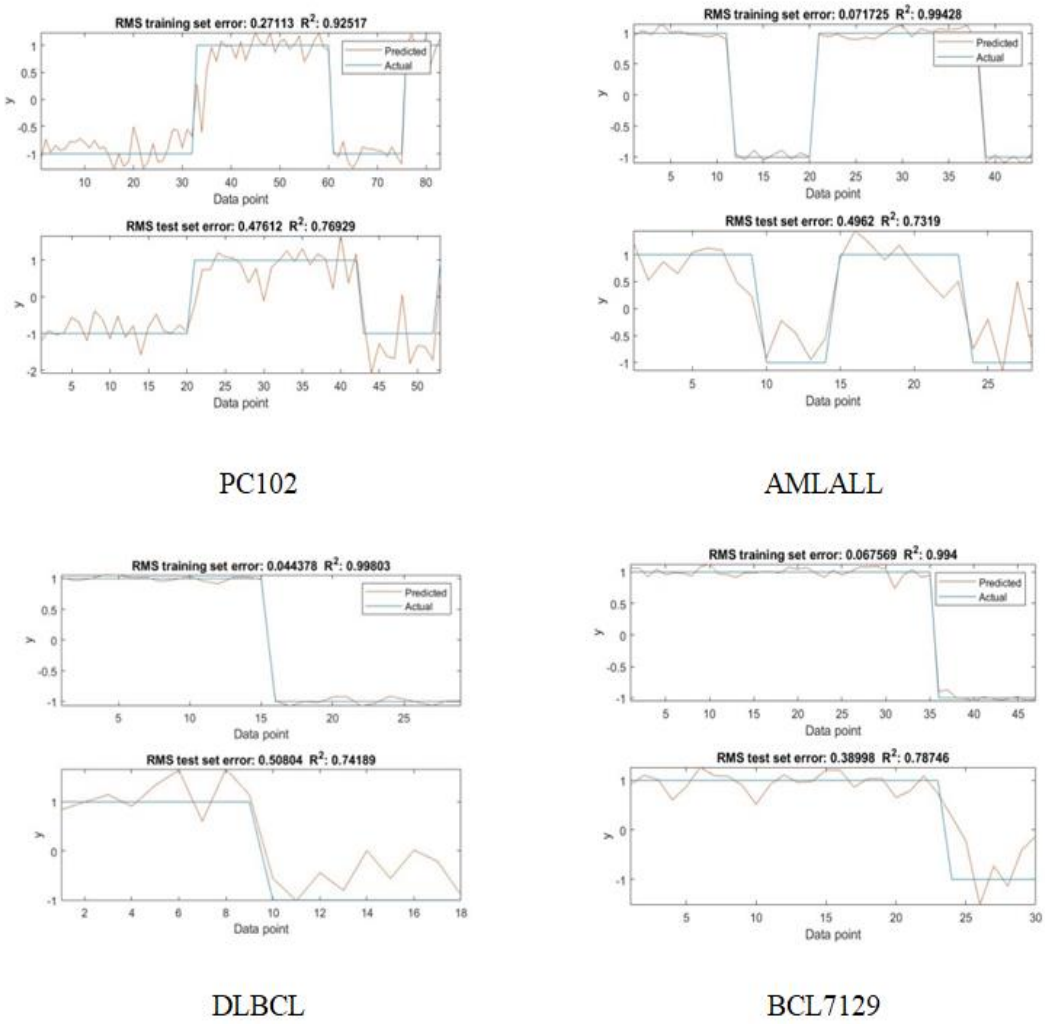

Figure 3. Correlation between predicted and actual values.

To ensure the quality of the gene subsets found by MGGP, the four datasets were classified using SVM, KNN, Ada-Boost and Naïve Bays. *Table* 6 presents classification rates of the four methods on datasets using: all dataset genes and only genes selected by MGGP.

The results indicate that classification using only MGGP-selected genes *usu*ally gives the best rate. Thus, it proves the validity of these selected subsets, whatever the induction method used and the effectiveness of the proposed method in selecting a good subset of genes that can differentiate well between the existing classes. For SVM classifier, the average of classification accuracy using only selected gene subsets for all microarrays was 97.62% which is improved by 2.59% in case of using all genes. Also, for KNN, Ada-Boost and *Naïve* bays classifiers, the average of classification using only MGGP

gene subset was respectively 12.98%, 6.89% and 58.22% better than using all genes in all datasets

Table 6 . Comparison of classification accuracy and number of gene selected by MGGP models using Different classifiers.

Dataset		Classification using all genes		Classification on selected gene subset by MGGP models	
		Ngene	Acc %	Ngene	Acc%
PCa102	SVM	12600	91.30%	24	93.33%
	KNN		91.30%		100%
	Ada-Boost		89.70%		90.37%
	Naïve bays		23.52%		86.95%
AMLALL	SVM	7129	95.83%	25	97.18%
	KNN		91.66%		100%
	Ada-Boost		91.54%		100%
	Naïve bays		20.83%		95.77%
DLBCL	SVM	4026	95.65%	30	100%
	KNN		71.73%		100%
	Ada-Boost		100%		100%
	Naïve bays		73.91%		100%
BCL7129	SVM	7129	97.36%	26	100%
	KNN		93.42%		100%
	Ada-Boost		81.57%		100%
	Naïve bays		25%		93.42%

Different experiments were created to evaluate the efficiency of the pro- posed gene selection approach. Some gene selection approaches, including CDNC [8], MOEA[27], MOGA [28], ABCD [29], MOGA-Cor [30], BHAPSO [31] and AHEDL [32] were also compared to assess the performance of the strategy. *Table* 7 provides more information on these techniques.

The results are evaluated according to SVM, ada-boost and naive bays classification accuracy in (*Table 8, 9* and *10*) respectively. The obtained results demonstrate that the proposed approach consistently outperforms the other gene selection techniques. The results obtained by SVM classifier indicate that the created approach's average classification accuracy in al microarray data was 93.42%. This value is 4.73% better than the average classification accuracy for the second best technique (ie : ABCD). Also, (*Table 9* and *10*) results were consistent with (*Table 8*)' s, and the created methodology exceeded the other examined gene selection approaches in all datasets .

Table 7. Description of the different gene selection methods.

Method	Description
CDNC	A new gene selection method was proposed in [7] based on community detection and node centrality techniques;
MOEA	A wrapper approach was proposed in [24] where multi-objective evolutionary algorithm was based on Niche-based Fitness Punishing technique and Elitism.
MOGA	A Multi-Objective Genetic Algorithm (MOGA), based on NSGA-II, for gene selection problem in [25].
ABCD	Finding the ideal subset of genes for the classification problem in this study involves combining a unique gene filtering method with an optimization technique.
MOGA-Cor	NSGA-II based approach was proposed in [27]. Correlation coefficient was used to filter the most significant genes. Then, a K-NN classifier was used to execute their MOGA based on NSGA-II for classification.
MOEDA	A Multi-Objective Estimation of Distribution Algorithm (MOEDA) was proposed in [28] with a Minimum Redundancy Maximum Relevance (mRMR) criterion for gene filtering.
BHAPSO	This work integrates a binary black hole with a modified binary PSO algorithm to present an unique hybrid technique for gene selection [29].
AHEDL	Using adaptive hypergraph embedded dictionary learning, a computational gene selection technique for microarray data categorization is described in this study [30].

Table 8. Classification accuracy of different gene selection approaches on SVM classifier.

Dataset	Proposed method	CDNC	MOEA	MOGA	ABCD	MOGA-Cor	BHAPSO	AHEDL
PCa102	**93.33%**	83.91%	-	-	82.67%	-	82.81%	79.72%
AMLALL	**97.18%**	91.16%	90.00%	98.03%	88.91%	92.60%	88.13%	87.13%
DLBCL	**100%**	-	90.00%	96.05%	100%	88.00%	-	-
BCL7129	**100%**	-	-	-	100%	-	-	-

Table 9. Classification accuracy of different gene selection approaches on ada-boost classifier.

Dataset	Proposed method	CDNC	ABCD	BHAPSO	AHEDL
PCa102	**90.37%**	81.92%	81.13%	79.98%	82.19%
AMLALL	**100%**	90.16%	83.31%	86.63%	89.71%

Table 10. Classification accuracy of different gene selection approaches on naïve bays classifier.

Dataset	Proposed method	CDNC	ABCD	BHAPSO	AHEDL
PCa102	86.95%	81.72%	81.49%	79.08%	82.21%
AMLALL	95.77%	91.48%	87.38%	86.65%	89.78%

6. Conclusion

In this paper, we proposed a new strategy for feature selection, in high-dimensional microarray datasets, based on the use of MGGP metaheuristic, for the development of efficient cancer classification. By learning from the microarray cancer datasets, the MGGP can determine an explicit formulation of combination of dissimilar genes that are most closely correlated to the target class, without using any classification model. To facilitate the implementation of multigene genetic programming, we used GPTIPS 2 as an open-source software platform for symbolic data mining in MATLAB.

Several experiments have be done to determine the parameters of the evolutionary process of the MGGP. The proposed method proved its effectiveness in gene selection with three datasets of more than 12600 genes and achieved a good classification rate with three different inductions methods. One limitation of the proposed method is that it has no standard endpoint and no standard way to adjust its parameters.

Therefore, future work can provide a method to help determine the parameters of the MGGP to improve its performance.

Reference

[1] P. Patel, K. Passi and C. K. Jain, ”Improving Gene Expression Prediction of Cancer Data Using Nature Inspired Optimization Algorithms,” 2021 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2021, pp. 337-343, doi: 10.1109/CSCI54926.2021.00128.

[2] Houssein, E.H., Hassan, H.N., Al-Sayed, M.M. et al. Gene Selection for Microarray Cancer Classification based on Manta Rays Foraging Optimization and Support Vector

Machines. Arab J Sci Eng 47, 2555–2572 (2022). https://doi.org/10.1007/s13369-021-06102-8.
[3] Huerta, Edmundo Bonilla, Béatrice Duval, and Jin-Kao Hao. "A hybrid GA/SVM approach for gene selection and classification of microarray data." Workshops on Applications of Evolutionary Computation. Springer, Berlin, Heidelberg, 2006.
[4] Dashtban, M., Mohammadali Balafar, and Prashanth Suravajhala. "Gene selection for tumor classification using a novel bio-inspired multi-objective approach." Genomics 110.1 (2018): 10-17.
[5] Scionti, F. et al.. "Integration of DNA Microarray with Clinical and Genomic Data". In: Agapito, G. (eds) Microarray Data Analysis. Methods in Molecular Biology, vol 2401. Humana, New York, NY. (2022).
[6] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.
[7] Zhou, Xin, and David P. Tuck. "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data." Bioinformatics 23.9 (2007): 1106-1114.
[8] Rostami, Mehrdad, et al. "Gene selection for microarray data classification via multi-objective graph theoretic-based method." Artificial Intelligence in Medicine 123 (2022): 102228.
[9] Cho, Ji-Hoon, et al. "Gene selection and classification from microarray data using kernel machine." FEBS letters 571.1-3 (2004): 93-98.
[10] Tang, Chang, et al. "Gene selection for microarray data classification via subspace learning and manifold regularization." Medical biological engineering computing 56 (2018): 1271-1284.
[11] Shukla, Alok Kumar, et al. "A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges." Evolutionary Intelligence 13 (2020): 309-329.
[12] Sánchez-Maroño Noelia, Oscar Fontenla-Romero, and Beatriz Pérez Sánchez. "Classification of microarray data." Microarray Bioinformatics (2019): 185-205.
[13] Gu, Quanquan, Zhenhui Li, and Jiawei Han. "Generalized fisher score for feature selection." arXiv preprint arXiv:1202.3725 (2012).
[14] He, Xiaofei, Deng Cai, and Partha Niyogi. " Laplacian score for feature selection." Advances in neural information processing systems 18 (2005).
[15] Sun, Yijun. "Iterative RELIEF for feature weighting: algorithms, theories, and applications." IEEE transactions on pattern analysis and machine intelligence 29.6 (2007): 1035-1051.
[16] Tubishat, Mohammad, et al. "Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection." Expert Systems with Applications 145 (2020): 113122.
[17] Kalaimani, V., and R. Umagandhi. "A novel wrapper FS based on binary swallow swarm optimization with score-based criteria fusion for gene expression microarray data." Materials Today: Proceedings (2020).
[18] Agrawal, R. K., Baljeet Kaur, and Surbhi Sharma. "Quantum based whale optimization algorithm for wrapper feature selection." Applied Soft Computing 89 (2020): 106092.
[19] Wang, Yadi, Xin-Guang Yang, and Yongjin Lu. "Informative Gene Selection for Microarray Classification via Adaptive Elastic Net with Conditional Mutual Information." arXiv preprint arXiv:1806.01466 (2018).
[20] Salem, Hanaa, Gamal Attiya, and Nawal El-Fishawy. "Classification of human cancer diseases by gene expression profiles." Applied Soft Computing 50 (2017): 124-134.
[21] Zhou, Xin, and K. Z. Mao. "LS bound based gene selection for DNA microarray data." Bioinformatics 21.8 (2005): 1559-1564.

[22] Ramón Díaz-Uriarte, and Sara Alvarez de Andrés. ”Gene selection and classification of microarray data using random forest.” BMC bioinformatics 7 (2006): 1-13.
[23] Orove, J. O., N. E. Osegi, and B. O. Eke. ”A multi-gene genetic programming application for predicting students failure at school. ” arXiv preprint arXiv:1503.03211 (2015).
[24] Searson, Dominic P. “GPTIPS 2: an open-source software platform for symbolic data mining.” Handbook of genetic programming applications. Springer, Cham, 2015. 551-573.
[25] A. Cano, A. Masegosa, S. Moral, ELVIRA biomedical data set repository, 2020, Last accessed: 23-February-2023. URL: http://leo.ugr.es/elvira/ DBCRepository/ .
[26]Princeton University Gene Expression Project, Microarray databases, 2020, 23-February-2023. URL: http://genomics-pubs.princeton.edu/.
[27]Liu, Juan, and Hitochi Iba. ”Selecting informative genes using a multi- objective evolutionary algorithm.” Proceedings of the 2002 Congress on Evolutionary Computation. CEC’02 (Cat. No. 02TH8600). Vol. 1. IEEE, 2002.
[28] José García-Nieto, et al. ”Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis.” Information Processing Letters 109.16 (2009): 887-896.
[29] Coleto-Alcudia, Veredas, and Miguel A. Vega-Rodríguez. ”Artificial bee colony algorithm based on dominance (ABCD) for a hybrid gene selection method.” Knowledge-Based Systems 205 (2020): 106323.
[30] Hasnat, Abul, and Azhar Uddin Molla. ”Feature selection in cancer microarray data using multi-objective genetic algorithm combined with correlation coefficient.” 2016 International Conference on Emerging Technological Trends (ICETT). IEEE, 2016.
[31] Pashaei, Elnaz, Elham Pashaei, and Nizamettin Aydin. ”Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization.” Genomics 111.4 (2019): 669-686.
[32] Zheng, Xiao, et al. ”Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning.” Gene 706 (2019): 188- 200.