Expert Systems With Applications 121 (2019) 233–243

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets

Sabah Sayed, Mohammad Nassef*, Amr Badr, Ibrahim Farag

Faculty of Computers and Information, Department of Computer Science, Cairo University, 5 Dr. Ahmed Zewail St., Orman, Giza, Egypt

ARTICLE INFO

Article history:
Received 18 March 2018
Revised 18 October 2018
Accepted 14 December 2018
Available online 14 December 2018

Keywords:
Microarray gene expression
DNA Methylation
Colon cancer
Lung cancer
Machine learning
Genetic algorithm
Feature selection
Support Vector Machine

ABSTRACT

Cancer is a dangerous disease that causes death worldwide. Discovering few genes relevant to one cancer disease can result in effective treatments. The challenge associated with the Microarray datasets is its high dimensionality; the huge number of features compared to the modest number of samples in these datasets. Recent research efforts attempted to reduce this high-dimensionality using different feature selection techniques. This paper presents an ensemble feature selection technique based on *t*-test and genetic algorithm. After preprocessing the data using *t*-test, a Nested Genetic Algorithm, namely Nested-GA, is used to get the optimal subset of features by combining data from two different datasets. Nested-GA consists of two Nested Genetic Algorithms (outer and inner) that run on two different kinds of datasets. The Outer Genetic Algorithm (OGA-SVM) works on Microarray gene expression datasets, whereas the Inner Genetic Algorithm (IGA-NNW) runs on DNA Methylation datasets. Nested-GA is performed on a colon cancer dataset with 5-fold cross validation. After applying Nested-GA, the Incremental Feature Selection (IFS) strategy is used to get the smallest optimal genes subset. The genes subset has been validated on an independent dataset resulting in 99.9% classification accuracy. Consequently, the biological significance of the resulting optimal genes is validated using Enrichment Analysis. Moreover, the results of Nested-GA have been compared to the results of other feature selection algorithms that have been run on either Gene Expression or DNA Methylation datasets. From the experimental results, Nested-GA showed the highest classification performance with a small optimal feature subset compared to the other algorithms. Furthermore, by running Nested-GA on lung cancer datasets that contain two different cancer subtypes, it resulted in significantly better classification accuracy (98.4%) compared to the accuracy of a previous research (84.6%) that utilized lung cancer DNA-Methylation data only.

1. Introduction

There are many types of cancer that can be caused by either genetic or epigenetic changes. Generally, cancer is a complex disease that is resulting from interactions of biological processes. Several measurement platforms have been developed and implemented in bioinformatics to understand these cancer processes (Gentle, Härdle, & Mori, 2010). One basic goal of bioinformatics cancer systems is to infer the malignant drivers of those biological processes.

Microarrays are one of the well-established tools used to identify and analyze the biological data. One function of the Microarray experiments is to monitor the expression level of genes on the genome scale (Whitworth, 2010). Results of those experiments could be formed in a matrix, called Gene Expression matrix, where each row corresponds to a particular gene and each column represents an experimental condition.

DNA Methylation (DNAm) is a common epigenetic mechanism, which controls the regulation of Gene Expression and is useful for early detection of cancer. There are many databases, that serve as, repositories of huge experimental data, such as the Gene Expression Omnibus (GEO) and ArrayExpress. Those databases contain data from Microarray experiments on a wide range of samples and under a variety of experimental conditions (Vazquez, de la Torre, & Valencia, 2012). The International Cancer Genome Consortium ICGC (http://icgc.org/) and the Cancer Genome Atlas TCGA (http://cancergenome.nih.gov/) projects developed cancer-specific repositories that contain complete genotypes. For cancer genome studies, those repositories are considered the main reference that offers the opportunity to test new approaches with real data (Vazquez et al., 2012).

The growing size of biomedical data resulted in many research challenges for the analysis of data and offer more opportunities to

* Corresponding author.
E-mail addresses: s.sayed@fci-cu.edu.eg (S. Sayed), m.nassef@fci-cu.edu.eg (M. Nassef), amr.badr@fci-cu.edu.eg (A. Badr), i.farag@fci-cu.edu.eg (I. Farag).

https://doi.org/10.1016/j.eswa.2018.12.022
0957-4174/

discover new knowledge from this data. Biomedical markers detection, diseases diagnosis, drug design and classification of high-dimensional data are some of these research trends.

Discovering few number of genes relevant to one cancer disease could derive in effective treatments. The challenge with Microarray datasets is its high dimensionality. Unfortunately, a Microarray dataset consists of a small number of observations but with many genes. The noise and variability of Microarray data add complications to the Microarray data analysis. With the high ratio between the number of features (genes) and the number of samples, classifying cancer subtypes becomes a complex process. It is common that a huge number of genes may be uninformative for the classification because they are either irrelevant or redundant (Abusamra, 2013). So, dimensionality reduction and feature selection techniques may be very useful for such a problem.

From the large number of genes in Microarray gene expression dataset, only a small number of genes strongly correlates with the targeted disease. More studies suggested that only a small number of genes can be sufficient markers for a specific disease (Li & Yang, 2002; Xiong, Fang, & Zhao, 2001), where the genes biological relationship with respect to the target disease can be easily identified. Those few genes are called biomarker genes. Using only biomarker genes in decision making reduces the computational effort and increases the classification accuracy. Selecting an effective and more representative gene subset is called a biomarker problem. In a Microarray dataset, there are many genes that are highly correlated. Those genes are considered redundant genes. In other words, if a biomarker gene set contains redundant genes, then this genes' set is not a comprehensive representation of the characteristics of the target disease. Redundant genes limit the efficiency and generality of the biomarker genes set (Ding & Peng, 2005), and so, the issue of gene redundancy should be solved in biomarker problem.

From the aforementioned problems, it is obvious that applying Feature Selection (FS) techniques in bioinformatics has become an important prerequisite step for model building rather than being an optional choice. Moreover, most of the pattern recognition techniques were not designed to deal with huge number of irrelevant features, so combining them with FS techniques results in more efficient solutions (Saeys, Inza, & Larrañaga, 2007). Feature selection refers to selecting the most relevant features from the original feature space (Abusamra, 2013).

One important objective of feature selection techniques is to avoid over-fitting and improve model performance, i.e., higher prediction accuracy for supervised classification and better cluster detection for unsupervised classification. Improving a model performance means providing faster and more cost-effective model. There are many FS techniques that are differing in the way each technique copes with the feature space to form a feature subset. In the classification problem context, according to how FS techniques combine the feature selection search with the construction of the classification model, they can be divided into three categories; filter methods, wrapper methods, and embedded methods.

The Filter techniques select relevant features independently of the selection model. They measure the relevance of features by only using the properties of the data then order the features according to the calculated relevance score. The Filter techniques are simple and fast but they neglect the features dependencies. Filter techniques are performed as a pre-processing step in the selection model and can be followed by one or more classification algorithms. The second category of FS techniques are the Wrapper techniques where selecting the features is dependent on the selection model. They define feature subsets and evaluate each by using a classification algorithm. Then, they select the feature subset with the high evaluation measure. The Wrapper techniques take into consideration the feature dependencies, but they are slower and computationally intensive. In the third category, namely the Embedded techniques, the feature selection is built into the search step done by classification algorithm. They consider the feature dependencies but with less computations than Wrapper techniques. For more details about feature selection, the reader can refer to (Saeys et al., 2007).

Using one FS technique does not guarantee obtaining a universally optimal feature subset (Yeung, Bumgarner, & Raftery, 2005). So, an ensemble FS approach runs different FS techniques where each technique produces a separate feature subset. Then, the ensemble FS approach combines the resulting feature subsets to form a final feature subset as its outcome. Ensemble FS approaches differ from each other in how they combine features. They may use averaging over multiple separate feature subsets (Levner, 2005; Li & Yang, 2002) that result from performing different runs of the same technique (for example, genetic algorithm) to assess the importance of each feature (Li, Umbach, Terry, & Taylor, 2004; Li, Weinberg, Darden, & Pedersen, 2001), and using a collection of decision trees as random forests to assess the relevance of each feature (Díaz-Uriarte & De Andres, 2006; Jiang et al., 2004). Ensemble FS approaches improve the robustness, stability, and generality but they require additional computations. The development of ensemble frameworks is a promising trend for improving the gene selection problem and the feature selection process in general. That's because the characteristics of an ensemble framework are more flexible and efficient in dealing with high dimensional data (Chin et al., 2015).

Genetic Algorithms (GAs), inspired by John Holland during the 1970s, are a class of evolutionary algorithms motivated by the biological theory of evolution, made popular (Scrucca et al., 2013). A Genetic Algorithm (GA) is used in search and optimization problems utilizing the "survival of the fittest" concept as known in evolutionary biology. A GA mimics the natural selection process in producing sets of solutions (population). Each solution, called chromosome, consists of a set of features (genes) that represent a candidate solution for the underlined problem. GA repeatedly generates solutions, evaluates their fitness and terminates when the given objective is achieved or when some stopping criteria is met. Genetic operators and fitness function characterize the implementation of a genetic algorithm. Fitness function is considered the main guide to the selection of the features. It is used to assign a probability to each chromosome in the population which reflects the quality of that chromosome and controls keeping that chromosome to the next generation. Genetic operators are used to investigate the entire search space and to avoid the local minima. The commonly used operators are crossover and mutation. Crossover is a mechanism for swapping genes between two randomly chosen chromosomes producing two new chromosomes for the next generation. Crossover can be performed on different kinds of representations (like binary or floating-point encodings). It also can be performed at single or multiple crossover points between chromosomes (Coley, 1999).

The mutation operator is the mechanism of flipping one or more gene in a randomly chosen chromosome according to a predefined probability. Altering gene values in mutation guarantees investigating all the search space of the underlying problem and causing variations in the resultant solutions. Mutation can be Binary encoding mutation, Value encoding mutation, or Permutation encoding mutation (Malhotra, Singh, & Singh, 2011). Elitism is an optional operator in a GA's implementation that allows retaining some chromosomes with high fitness values to the next generations.

Support Vector Machine (SVM) is a supervised classification algorithm developed by Cortes and Vapnik (1995). SVM is used to classify high-dimensional and noisy data. SVM was originally designed for binary classification problems, but recently several variants of SVM have been introduced to deal with multiclass prob-

lems. The basic idea of SVM is to generate high dimensional feature space according to the attributes of features in the data. Then, it defines a hyperplane (decision boundary) to separate the features into two portions where each one contains the data points of one class. The shortest distance between the first and second samples that are next to the hyperplane is defined as the margin of the hyperplane (Karatzoglou, Meyer, & Hornik, 2006). Support vector learning can be presented as the problem of identifying a hyperplane with the largest margin to split one class from the other. It depends on the maximum-margin principle which means maximizing the margin between the decision hyperplane and the closest training samples. The hyperplane with a large margin avoids the noise effect more than the hyperplane with a small margin.

An Artificial Neural Networks (ANN) consists of multiple layers of interconnected neuron units. The first layer is the input layer to match the feature space. The input layer is followed by one or more layers of nonlinearity then the last layer is a linear regression or classification layer to match the output space. For the intermediate layers, output from one layer is forwarded as an input to the next layer. Each layer has some weights that are used with the input to form its output.

Each time the training samples are used by an ANN to adapt the weights in order to minimize the classification error. Adapting the weights is called the learning process. Deep-learning (LeCun, Bengio, & Hinton, 2015) architecture involves multiple levels of nonlinearity. So, the basic framework of multi-layer neural networks can be used to perform deep learning tasks (Arora, Candel, Lanford, LeDell, & Parmar, 2015).

The most important advantage of the deep-learning is that learning is not done by human engineers and it does not require human interference. Alternatively, the layers weights are learned from data using general-purpose learning techniques. Deep-learning is applicable to many domains of science and has exhibited high performance on complex and high dimensional data. So, it is more suitable for Microarray data.

Various research studies have been attempted to apply miscellaneous feature selection techniques over Microarray data. One study aimed to identify the biomarker genes for glioma-alerted by integrating the Monte Carlo simulation with singular value decomposition (SVD) (Han, Lai, Xie, Li, & Zhu, 2014).

An ensemble-based feature selection technique was proposed by Cai et al. (2015) to classify different lung cancer subtypes by using DNA Methylation data. This technique integrates Multi-category Receiver Operating Characteristic (Multi-ROC), Random Forests (RFs) and Maximum Relevance and Minimum Redundancy (mRMR) methods followed by machine learning methods.

A study conducted by Abusamra (2013) aimed to compare some of state-of-the-art feature selection and classification methods on gene expression data of glioma using five-fold cross-validation to evaluate the classification performance. The used feature selection methods including: information gain, twoing rule, sum minority, max minority, gini index, sum of variances, *t*-statistics, and one-dimension Support Vector Machine. Feature selection methods are used with three classification algorithms; SVM, KNN, and random forest.

A feature selection and classification framework is offered by Valavanis, Pilalis, Georgiadis, Kyrtopoulos, and Chatziioannou (2015). The proposed framework uses evolutionary algorithms and Gene Ontology (GO) tree and is applied on 450k human methylation data of breast cancer and B-cell lymphoma.

Le, Uy, Dung, Binh, and Kwon (2013) tried to identify the associations between diseases and protein complexes. Firstly, a protein complex network is constructed where two protein complexes are connected by using their shared genes. Then, random walk with restart (RWR) algorithm is applied on that network in order to rank the protein complexes based on their relative importance to known disease protein complexes. The performance of that method is evaluated by the leave-one-out cross-validation method. That method is applied on the breast cancer dataset.

González and Belanche (2013) applied an algorithm for feature selection using Simulated Annealing and discrete multivariate joint entropy on five public domain Microarray gene expression data samples aiming to find the small subsets of highly relevant genes.

Luque-Baena, Urda, Subirats, Franco, and Jerez (2013) compared between genetic algorithm with constructive neural networks and the classical Stepwise Forward Selection (SFS) algorithm in predicting the cancer outcome. Welch *t*-test filtering method is embedded into the two algorithms. Those two algorithms are applied on six cancer gene expression datasets.

A method for prediction biomarker mining is introduced by Popovic, Sifrim, Pavlopoulos, Moreau, and De Moor (2012). A genetic algorithm with a novel fitness function and a bagging-like model averaging scheme is applied on three independent publicly available Microarray datasets for colon cancer; one for training and one for testing and the last one for external validation. Ingenuity Pathway Analysis (IPA) is used as a functional analysis to estimate the biological relevance of the resulting gene signature.

Prostate cancer biomarker genes are identified by Raza and Jaiswal (2013) by constructing a gene regulatory network using two-stage filtering approach *t*-test and fold-change measure. After identifying significant genes by using the two filtering methods, Pearson correlation coefficient is used to compute regulatory relationships between the identified genes.

Jirapech-Umpai and Aitken (2005) used genetic algorithm as a wrapper feature selection for predicting gene markers for the leukemia disease. They assessed the performance using a low variance estimation technique and presented an analysis of the predicted genes. they concluded that the choice of feature selection criteria have a significant effect on the classification accuracy.

Ooi and Tan (2003) applied genetic algorithm to the problem of multi-class prediction. A GA-based gene selection scheme is presented to predict the marker gene group, as well as the optimal group size, which maximized classification success using a maximum likelihood (MLHD) classification method. The GA/MLHD-based approach is applied to The NCI60 gene expression dataset contains the Gene Expression profiles of 64 cancer cell lines. the approach achieved higher classification accuracies than other published predictive methods on the same multi-class test dataset.

García and Sánchez (2015) presented a two-stage classification model based on combining feature selection with the dissimilarity based representation paradigm. The ReliefF algorithm was used in the first stage to generate a subset of top-ranked genes, whereas, in the second stage, a dissimilarity space formed by the samples of the selected genes was used in constructing a classifier. The performance of the dissimilarity-based models was analyzed by means of a collection of experiments to classify eight Microarray gene expression datasets using an Artificial Neural Network, a Support Vector Machine and the Fishers linear discriminant classifier built on the gene space, and the same classifiers built on the dissimilarity space. The experimental results showed that the dissimilarity-based classifiers outperform the feature-based models.

A research was conducted by García, Sánchez, Cleofas-Sánchez, Ochoa-Domínguez, and López-Orozco (2017) to analyze the effect of high-dimensional data on the classification of gene expression datasets. Gain ratio and ReliefF were used as gene ranking methods with six classifiers on four biomedical datasets. The results showed that regardless of the used gene ranking algorithm and classifier, the highest classification performance was achieved by using a very small number of genes (less than the fifth of the total amount of genes).

A dynamic relevance-based gene selection method (DRGS) was introduced by Sun et al. (2013) to identify a gene subset from

Table 1
Datasets description.

Dataset type	Number of variables	Dataset function	Sample type	Number of samples
TCGA gene Expression data	17,815 genes	46 samples for 5-fold cross validation training and 200 samples for testing	Normal Cancer	21 225
GEO gene Expression data	17,815 genes	174 samples for Independent test	Normal Cancer	19 155
DNA Methylation	27,578 CpG sites	276 samples for 5-fold cross validation training	Normal Cancer	42 234

high dimensional gene expression Microarray data for cancer classification and diagnosis. This method aimed to use a target-based scheme for relevance, interdependence and redundancy analysis to retain the useful functional gene groups. This is done by updating the relevance between each gene and target dynamically when a new gene is selected. The proposed method was validated against Information Gain, mRMR, ReliefF, and Significance Analysis of Microarrays (SAM) on six gene expression Microarray datasets. The results showed that, compared to the other selectors, DRGS selected fewer genes with higher classification accuracy.

In this study, an ensemble feature selection approach based on a Nested Genetic Algorithm is proposed to select the optimal Microarray genes subset that represents the biomarker genes of one cancer type by combining the information from two types of Microarray data; gene expression data and DNA Methylation data. The Nested Genetic Algorithm (Nested-GA) utilizes both Filter and Wrapper feature selection methods. For filter feature selection, *t*-test is used as a preprocessing step. Then, a Nested Genetic Algorithm composed of two genetic algorithms, one with a Support Vector Machine (SVM) and the other with a Neural Network, are used as the Wrapper feature selection technique. Incremental Feature Selection (IFS) is then used as an ensemble approach to present the biomarker genes as its outcome.

2. Materials and methods

2.1. Datasets

The results presented in this paper are based on the colon cancer gene expression data downloaded from The Cancer Genome Atlas (TCGA) https://tcga-data.nci.nih.gov/tcga/ and TCGA DNA Methylation dataset based on the IHM-27k platform for running the Nested-GA algorithm. The colon cancer gene expression data from Gene Expression Omnibus (GEO) from NCBI has been used as a dataset for independent testing. Table 1 shows more details of the used datasets.

2.2. The proposed algorithm

The pipeline of the proposed method, as shown in Fig. 1, starts by preprocessing for both the Gene Expression and the DNA Methylation datasets before applying feature selection. After that, feature filtering is applied using *t*-test to select a subset of the top ranked Genes and CpG sites from Gene Expression and DNA Methylation data. The filtered gene subset is fed as an input to the OGA-SVM with SVM fitness function, while the filtered CpG sites subset is fed as input to the IGA-NNW with *N*-Net fitness function. Finding the relation between genes and CpG sites is important step that is used in the initialization stage of each IGA-NNW and OGA-SVM. After determine number of runs of OGA-SVM, we get number of solutions *N*. We rank the genes in the *N* solutions in descending order based on their frequency. Next, we incrementally append genes with high rank producing *M* subsets of top ranked genes, models. SVM is used to evaluate the *M* models to get the optimal model. At the end, the optimal model's genes are validated.

2.2.1. Preprocessing step

It is common to have some genes with missing expression values in most of the gene expression datasets. These genes should play no role in building the final classifier, and so, they should be excluded. The same exclusion process can be applied to exclude the CpG sites with missing values from the DNA Methylation datasets.

2.2.2. Filter feature selection

Within the huge gene expression data, there are hundreds genes that are irrelevant or redundant. So, it is significant to reduce the number of genes in order to get a good accuracy for the classification task. The Student's *t*-test is one of the most successful filter feature selection methods in terms of the quality of the ranked features (Huerta, Duval, & Hao, 2010). We apply Student's *t*-test on gene expression dataset using *t.test()* R function as follows:

1. Divided samples into two classes; normal and tumor.
2. Calculate *p*-value for each feature reflecting how this feature is effective at separating classes.
3. Order all features according to *p*-value ascending.
4. Select the best features (with low *p*-value).

For gene expression dataset we select the first 3000 gene and in DNA Methylation dataset we select the first 10,000 CpG site.

2.2.3. Proposed wrapper feature selection (Nested-GA)

A simple GA starts with initializing a population and runs in multiple iterations. Each iteration consists of some steps, which are known as GA operators; selection, crossover and mutation. At the end of each iteration, a new generation is created to be inputted to the next iteration. The algorithm terminates when reaching the maximum number of iterations or finding the best solution.

The proposed Nested-GA consists of two Nested Genetic Algorithms; Inner and Outer. The Outer one (OGA-SVM) is the main algorithm that has gene expression data as input and outputs the best subset of genes evaluated by SVM as a fitness function. The Inner one (IGA-NNW) takes DNA Methylation data as input and outputs the best subset of CpG sites by utilizing deep-learning for fitness function. We have two-way update from IGA-NNW to OGA-SVM and from OGA-SVM to IGA-NNW. Each iteration of OGA-SVM, a complete run of IGA-NNW is performed firstly to output a subset of CpG sites used as guidance in forming the population of the OGA-SVM in this iteration. So instead of initializing the population randomly, we initialize the population with genes related to best CpG sites resulting from the IGA-NNW which improves the ability of OGA-SVM in finding the best genes. The same thing is done in IGA-NNW by initializing the population with CpG sites related to best genes resulted in the previous iteration of OGA-SVM. So the final solution of Nested-GA is resulting from combining the information from Gene Expression and DNA Methylation data. The flowchart of the Nested-GA algorithm is depicted in Fig. 2, and its pseudocode is illustrated in Listing 1.

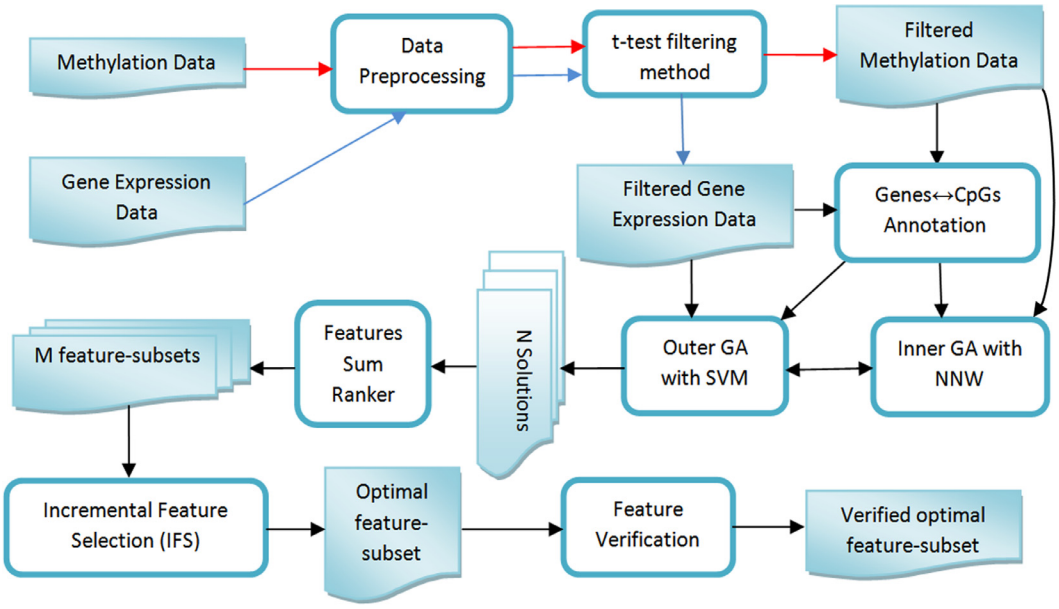

Fig. 1. Pipeline of the proposed method.

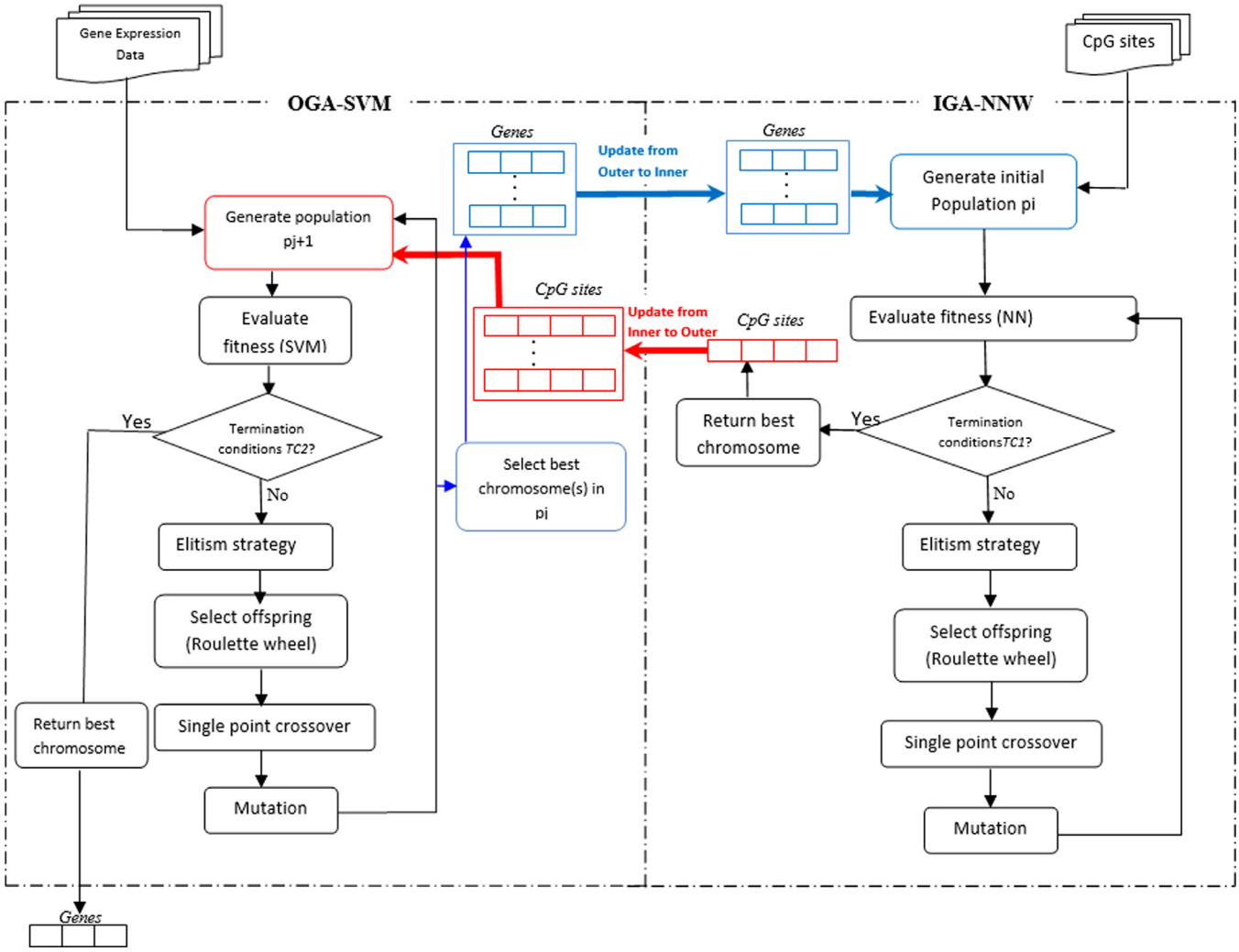

Fig. 2. Nested-GA flowchart.

Listing 1. Nested-GA pseudocode.

```
1:Outer-Ch = initializeRandom(1,M)
2:FOR(int i = 0; i<Outer-maxIter; i++)
3:   Inner-p = IntializePopulationInner(Outer-Ch, M)
4:   FOR(int j= 0; j<Inner-maxIter; j++)
5:      Inn-value  = NN-fitness(Inner-p)
6:      IF(Inn-value >= Required-value)
7:         break
8:      ELSE
9:         Inner-E = elitism()
10:        srw = selectRouleteWheel()
11:        Inner-C= performCrossover(srw)
12:        Inner-U = performMutation(srw)
13:        Inner-p = ReplacePopulation(Inner-E,Inner-C, Inner-U )
14:     EndIF
15:  EndFOR
16:  Outer-p= IntializePopulationOuter(Inner-p, N)
17:  Out-value = SVM-fitness(Outer-p)
18:  IF(Out-value >= SVMRequired-value)
19:     Outer-Ch= selectBest()
20:     break
21:  ELSE
22:     Outer-E = elitism()
23:     srw = selectRouleteWheel()
24:     Outer-C= performCrossover(srw)
25:     Outer-U = performMutation(srw)
26:     Outer-p = ReplacePopulation(Outer-E,Outer-C, Outer-U )
27:  EndIF
28:EndFOR
29:return Outer-Ch
```

457	23	7098	. . .	5001	5

Fig. 3. Chromosome structure. Each gene in the chromosome refers to an index of either a gene (Outer GA) or a CPG site (Inner GA).

Nested-GA Chromosome Structure A chromosome ch with n features is represented as $ch = (x_1, x_2, ..., x_n)$. These n features are randomly selected from the reduced feature set F produced from the previous stage. Each feature x_i is represented as an integer value that refers to the index of this feature in F. The chromosome structure shown in Fig. 3 is used for IGA-NNW chromosome *CpG-InCH* and OGA-SVM chromosome *Gene-OutCH*.

Steps of Nested-GA are as follow:

1. Initialize the OGA-SVM population of chromosomes:
   (a) Initialize the IGA-NNW initial population p_i with Y chromosomes each contains y CpG sites selected from the filtered CpG sites (Fc) produced from the previous stage. Each chromosome is represented as an array of y indices from 1 to Fc that refer to the selected CpG sites. In first iteration of Nested-GA, IGA-NNW chromosomes are randomly initialized. For iteration i ($i = 2, ..., x$) of Nested-GA, IGA-NNW chromosomes are initialized by using best OGA-SVM chromosomes from previous iteration *i-1*. Genes in OGA-SVM chromosome are mapped to CpG sites by using the *minfi* and *IlluminaHumanMethylation27kanno:ilmn12:hg19* R packages. Each gene can be mapped to h ($h = 0 : 50$) CpG sites.
   (b) Calculate the fitness value f_i for each chromosome in the current IGA-NNW generation using deep-learning neural network as fitness function. Using 5-fold cross-validation, train the neural network with the training data producing weights used in classifying samples in the test data. Those weights are used in calculating a score f_i for each chromosome reflecting the quality of this chromosome to be selected in the next generation.
   (c) Check if the termination conditions have been reached. The algorithm terminates with two conditions; reaching a solution with a predefined fitness value or reaching a predetermined number of iterations. In this case the algorithm outputs the best solution (subset of CpG sites) which is the chromosome with the highest fitness value in the current generation. Otherwise continue with the following steps.
   (d) To improve the performance, we select the best chromosomes in the current generation to be persisted in the next generation with no change (elitism mechanism). To avoid trapping in local peaks, we chose to perform elitism for 9 consecutive generations and cancel it for the 10th generation and repeat that for all generations.
   (e) Apply Roulette Wheel Selection (Sastry, Goldberg & Kendall, 2005) to select subset W for crossover with length l_c. Steps of selection are as the following:
      i. Generate random number r between 0 and sum of fitness values.
      ii. For each chromosome in the current generation, check if the chromosome's fitness is less than r then pick this chromosome to be in W and return to step 1. Otherwise, check another chromosome.
      iii. Repeat step 1 and 2 till l_c chromosomes are selected.
   (f) Apply crossover by randomly select two parent chromosomes to create two new chromosomes. Crossover is applied as in Fig. 4.
   (g) Randomly select set of chromosomes with length *lm* for mutation with mutation rate *Pm*. Perform a random single point mutation on these chromosomes by altering their genes values to ensure that a sufficient portion of the parameter space is explored.
   (h) Replace old generation with the new generation contained all chromosomes produced from elitism, crossover, and mutation algorithms.
   (i) Repeat from step (b).
2. Calculate the fitness value $svmf_i$ for each chromosome in the current OGA-SVM generation using SVM method from *e1071* package as fitness function. Using 5-fold cross-validation, train the SVM algorithm with the training data then used it to produce the accuracy of classifying samples in the testing data.
3. Check for the termination conditions (as illustrated in Step (C)), and then output the fittest solution (subset of Genes) in this generation. Otherwise continue with the following steps.
4. Form a subset of chromosomes by applying elitism for 9 consecutive generations and cancel it for the 10th generation and repeat that for all generations.
5. Apply Roulette Wheel Selection (Sastry et al., 2005) to select subset w for crossover with length *lc* and subset h for Mutation.
6. Apply crossover by randomly select two parent chromosomes to create two new chromosomes.
7. Perform a random single point mutation.
8. Select best chromosomes from current generation in OGA-SVM and perform step 1 (IGA-NNW).
9. Get the best subset of CpG sites and get their related Microarray genes to produce k OGA-SVM chromosomes.
10. Generate new OGA-SVM generation from combining chromosomes in *k, h, w, e*.
11. Replace old generation with the new generation contained all chromosomes produced from elitism, crossover, and mutation algorithms.
12. Go back to step 2.

The parameters used in Nested-GA are described in Table 2.

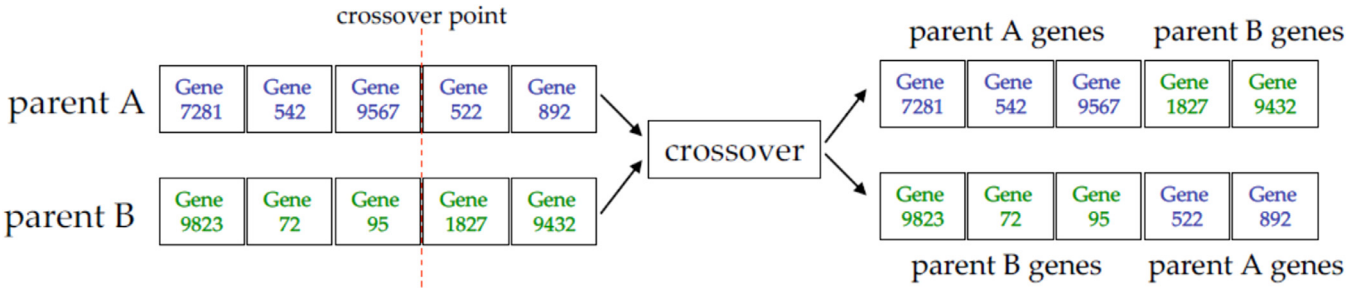

Fig. 4. Crossover mechanism.

Table 2
Nested-GA parameters description.

Parameter	Description	Value
N	number of Microarray genes (features in OuterGA)	17,814
M	number of CpG sites (features in InnerGA)	27,578
Fg	reduced features (Microarray genes) by *T*-test	3000
Fc	reduced features (CpG sites) by *T*-test	10,000
Outer-PSize	number of chromosomes in OuterGA population	15
Inner-PSize	number of chromosomes in InnerGA population	20
n	number of genes in a chromosome for OuterGA	20
m	number of genes in a chromosome for InnerGA	50
Outer-maxIter	maximum number of iterations for OuterGA	100
Inner-maxIter	maximum number of iterations for InnerGA	100
J	counter increased with each iteration in InnerGA	1 : Outer-maxIter
I	counter increased with each iteration in OuterGA	1 : Inner-maxIter
Outer-Pc	probability of crossover for OuterGA	0.5
Inner-Pc	probability of crossover for InnerGA	0.5
Outer-Pm	probability of mutation for OuterGA	0.1
Inner-Pm	probability of mutation for InnerGA	0.1
Outer-E	elitism selected chromosomes for OuterGA	1
Inner-E	elitism selected chromosomes for InnerGA	2
Outer-C	offspring (chromosomes) produced from crossover for OuterGA	7
Inner-C	offspring (chromosomes) produced from crossover for InnerGA	10
Outer-U	chromosomes produced from mutation for OuterGA	1
Inner-U	chromosomes produced from mutation for InnerGA	2
R	number of Nested-GA runs	100
goalF	required fitness value used in Nested-GA	95%
k	number of folds in cross-validation for Nested-GA	5
TC1: *j*	maxIter or the desired accuracy from NN (Neural Network) classifier is achieved	Inner-maxIter OR goalF
TC2: *i*	maxIter or the desired accuracy from SVM (Support Vector Machine) classifier is achieved	Outer-maxIter OR goalF

2.2.4. Feature sum ranker

N feature subsets are produced after repeating Nested-GA *N* times. Unique features are picked from those *N* subsets, and then sorted in descending order based on their cumulative frequencies over all the *N* subsets. The more frequent a feature, the higher its rank is.

2.2.5. Feature subsets evaluator

Ranked features list *L* is produced from previous stage. Incremental Feature Selection IFS is applied to produce *S* feature sets. Starting with the three top ranked features, the first feature set is constructed. The remaining features are added one by one incrementally to produce new feature sets. So each new set is the previous set with a new feature added. Finally, *S* feature sets are constructed where the *i*th feature set is: $si = (f_1, f_2, ..., f_{i+2})$ where $(1 < i < S+2)$

Each feature set is evaluated with SVM using 5-fold cross-validation. The feature subset of choice is the one with the highest classification accuracy and the lowest number of features. Classification accuracy is calculated as in the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where *TP*, *TN*, *FP*, and *FN* represent the true positive, true negative, false positive, and false negative respectively.

2.2.6. Enrichment Analysis

Gene Ontology (GO) (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2011) are the most famous Enrichment Analysis tools. Gene and gene product features across all species are represented in GO. KEGG pathway is used for mapping genes to pathways. There are three categories for the GO terms, which are Biological Process (BP), Cellular Component (CC) and Molecular Function (MF).

3. Results and discussion

This section presents the experimental setup and discusses the different evaluation techniques of the proposed Nested Genetic Algorithm (Nested-GA) based on the colon cancer datasets detailed in Section 2.1 and Table 1. The pipeline depicted in Fig. 1 has been followed in all the conducted experiments considering the number of feature subsets equals 100. Moreover, the parameter values of the proposed algorithm are listed in Table 2.

In order to evaluate the proposed Nested-GA approach, its accuracy has been compared to the accuracies of other multiple feature selection algorithms. In all the conducted experiments, the input features are the features selected by *t*-test filtering method. Moreover, an independent colon cancer dataset from GEO was used to further validate the strength of the classification model.

Table 3
SVM Accuracy on the testing dataset and on the independent dataset.

Num. of genes	GA-SVM accuracy		GA-NNW accuracy		Nested-GA accuracy	
	Testing data	Independent testing data	Testing data	Independent testing data	Testing data	Independent testing data
2	0.9461538	0.9482759	0.9461538	0.9425287	0.9384615	0.9425287
3	0.9769231	0.9770115	0.9561538	0.9561538	0.9923077	0.9942529
4	0.9769231	0.9827586	0.9923077	0.9942529	0.9846154	0.9885057
5	0.9846154	0.9885057	0.9769231	0.9827586	0.9923077	0.9942529
6	0.9923077	0.9942529	0.9846154	0.9885057	0.9999999	0.9999999
7	0.9923077	0.9942529	0.9923077	0.9942529	0.9999999	0.9999999
8	0.9999999	0.9999999	0.9999999	0.9999999	0.9999999	0.9999999

Table 4
Accuracy of OGA-SVM, IGA-NNW, KNN, RF and Nested-GA on the DNA Methylation dataset.

Num. of CPG sites	GA-SVM	GA-NNW	KNN	RF	Nested-GA
2	0.9778761	0.8539823	0.9292035	0.8628319	0.9384615
3	0.9955752	0.920354	0.9778761	0.8318584	0.9923077
4	0.9955752	0.9778761	0.9778761	0.9159292	0.9846154
5	0.9955752	0.9911504	0.9867257	0.9823009	0.9923077
6	0.9955752	0.9955752	0.9867257	0.9867257	0.9999999
7	0.9955752	0.9955752	0.9911504	0.9911504	0.9999999
8	0.9955752	0.9955752	0.9867257	0.9867257	0.9999999
9	0.9955752	0.9955752	0.9955752	0.9911504	0.9999999
10	0.9955752	0.9955752	0.9999999	0.9955752	0.9999999
11	0.9955752	0.9999999	0.9999999	0.9955752	0.9999999
12	0.9955752	0.9999999	0.9955752	0.9911504	0.9999999
13	0.9999999	0.9999999	0.9955752	0.9911504	0.9999999

At first, Nested-GA has been compared to a non-nested Genetic algorithm with SVM as its fitness function (GA-SVM) and to a non-nested Genetic algorithm with deep-learning neural network fitness function (GA-NNW) that both run over the colon cancer gene expression (CC-GE) dataset. Table 3 lists the SVM performance measurement (Accuracy) for the three experiments on the testing dataset and on the independent dataset, respectively.

After that, Nested-GA has been compared to GA-SVM, GA-NNW, KNN, and RF that all run over colon cancer DNA Methylation (CC-DM) dataset. It is worth to note that Nested-GA uses both the CC-GE and CC-DM datasets. Table 4 shows the accuracies of the five algorithms over the testing dataset. Although GA-SVM had better accuracy compared to Nested-GA when using two or three features, the accuracy of Nested-GA was noticeably better than GA-SVM when using more features.

Based on the results listed in Tables 3 and 4, it is clear that Nested-GA improves the classification accuracy by using fewer genes (six genes).

Based on the proposed algorithm, the colon cancer biomarkers can be accurately discovered by selecting the smallest satisfactory optimal feature set to represent the Microarray gene markers. Using this criterion, a feature set including six Microarray genes is selected. The genes are "*DAB2IP*", "*KLRB1*", "*NUP155*", "*NPC1L*", "*CDKN2A*" and "*SEC61A2*". As a step towards the validation of the resultant biomarkers, Fig. 5 depicts the heatmap generated for the six biomarkers genes (rows) with respect to the experimental samples (columns). It is clear from the heatmap that these six genes are cooperatively indicating high discrimination ability between the normal and cancerous samples. Gene *DAB2IP* has the highest discrimination ability, whereas gene *CDKN2A* has the lowest one.

As a second step towards the validation of the resultant biomarkers, they have been substituted by six less important ones ("*MGC10701*", "*CCDC85B*", "*BOC*", "*METTL7B*", "*KIAA2013*" and "*LAMB3*") resulting in accuracy of 0.553 for the testing dataset and 0.501 for the independent dataset. This means that replacing the six resultant genes by less important ones leaded to deterioration of the classification accuracy.

3.1. Enrichment Analysis

The top GO terms and KEGG pathways are derived from DAVID and StRAnGER2. Inputs of DAVID and StRAnGER2 are the genes where the minimum number of genes corresponding to GO terms and KEGG pathways is 2. Most of the resultant genes have been validated in previous research. More specific, *DAB2IP* (Kibriya et al., 2011), *NUP155* (Ancona et al., 2006; Bianchini et al., 2006; Chakraborty, 2014; Halvey, Zhang, Coffey, Liebler, & Slebos, 2011; Zhao, 2013), *CDKN2A* (Bandrés et al., 2006; Barat & Ruskin, 2015; Exner et al., 2015; Kibriya et al., 2011; Lundemo, Pettersen, Berge, Berge, & Schønberg, 2011; Luo et al., 2016), *NPC1L* (Shi et al., 2010), *SEC61A2* (Lundemo et al., 2011) have been reported to have effect on colon cancer. Moreover, it has been proved in uniport (http://www.uniprot.org/) that *DAB2IP* and *CDKN2A* are tumor-related genes.

Furthermore, a Copy Number Variation (CNV) dataset of colon cancer has been used to explore any possible association between its tumor CNV segments and the resultant NestedGA six genes. The CNV dataset was downloaded from (http://firebrowse.org/?cohort=COAD). It consisted of 918 samples: 453 tumor samples and 465 normal samples. The following steps have been implemented as mentioned in (Guttery et al., 2018) in order to reach the genes that are intersecting with CNV segments in tumor samples. First, the Probes meta file from (ftp://ftp.broadinstitute.org/pub/GISTIC2.0/hg19_support/) was used to differentiate between the normal and tumor CNV segments. After that, a dataset for all the human genes (hsapiens gene ensembl) from the host (grch37.ensembl.org (hg19)) was used as a reference dataset with CNV colon dataset to get the genes overlapped with CNV segments. This step was done by utilizing the two *R* packages entitled *biomaRt* and *GenomicRanges*. At the end, the resultant 554 genes have been intersected with the six genes resulted from the Nested-GA approach. The *NUP155* gene appeared to be a common gene between the two gene sets. This means that the CNV segments falling inside the *NUP155* gene might play an important role in altering its normal expression level.

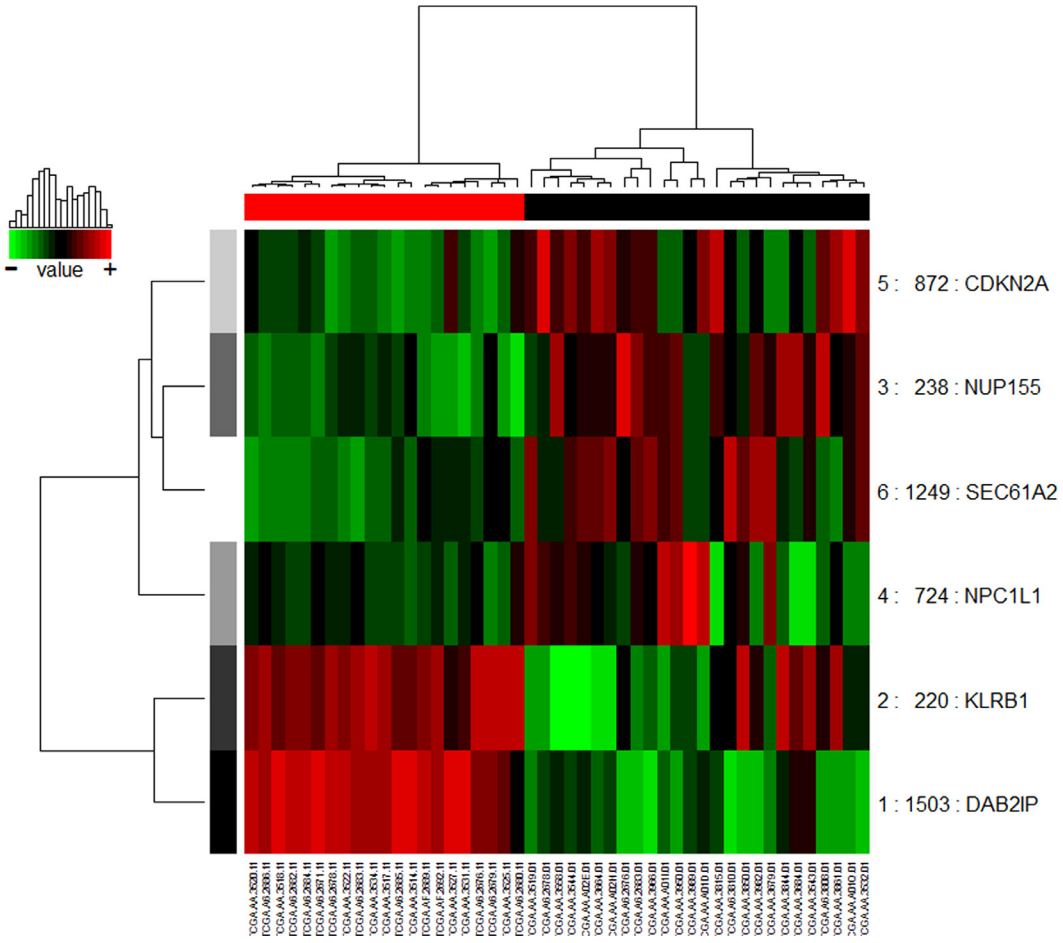

Fig. 5. A heatmap of the resultant six colon cancer biomarker genes.

Table 5
Lung cancer dataset description.

Dataset type	Number of variables	Dataset function	Sample type	Number of samples
TCGA gene Expression data	17,813 genes	70 samples for 5-fold cross validation training and 118 samples for testing	LUAD LUSC	33 155
DNA Methylation	27,578 CpG sites	311 samples for 5-fold cross validation training	LUAD LUSC	151 160

3.2. Comparative study

Following the same experimental pipeline depicted in Fig. 1 and the same Nested-GA parameter values mentioned in Table 1, a comparative study has been conducted between the proposed Nested-GA algorithm and the work presented in (Cai et al., 2015) that was based on DNA Methylation lung cancer datasets. The aim of this comparative study was to check the effectiveness of the proposed algorithm with respect to another type of cancer that has known subtypes.

Table 5 describes the Gene Expression and DNA Methylation lung cancer datasets used by Nested-GA. The accuracies of OGA-SVM and IGA-NNW running over DNA Methylation datasets have been compared to the Nested-GA accuracies with respect to different number of features (CpG sites). Table 6 lists the accuracies of OGA-SVM and IGA-NNW compared to Nested-GA. It is clear that Nested-GA resulted in noticeably better classification accuracies between the two lung cancer subtypes compared to the accuracies of OGA-SVM and OGA-NNW. Moreover, Nested-GA resulted in significantly better accuracy (98.45%) using only 16 features compared to the highest accuracy published in (Cai et al., 2015) (84.54%) using 45 features.

Table 6
Accuracies of OGA-SVM, IGA-NNW and Nested-GA over the lung cancer DNA Methylation dataset.

Num. of CPG sites	GA-SVM	GA-NNW	Nested-GA
2	0.8151659	0.7630332	0.8992248
3	0.8199052	0.7677725	0.875969
4	0.8388626	0.8151659	0.9224806
5	0.7819905	0.8246445	0.9379845
6	0.7867299	0.7772512	0.9457364
7	0.7772512	0.7914692	0.9612403
8	0.7345972	0.7298578	0.9457364
9	0.6777251	0.7582938	0.9534884
10	0.6872038	0.7725118	0.9534884
11	0.7014218	0.7725118	0.9689922
12	0.7393365	0.7725118	0.9612403
13	0.7440758	0.7772512	0.9534884
14	0.7393365	0.7867299	0.9612403
15	0.7393365	0.7819905	0.9767442
16	0.7393365	0.7772512	0.9844961

4. Conclusion

This study introduced a Nested Genetic Algorithm (Nested-GA) that consists of two genetic algorithms as an approach for feature selection by correlating different types of Microarray datasets. Nested-GA was applied on high dimensional Gene Expression and DNA Methylation Microarray data for colon cancer aiming to identify cancer biomarkers. An Incremental Feature Selection (IFS) strategy was used to select few informative and significant Microarray genes after running Nested-GA for number of times.

The results showed that the optimal Microarray genes subsets obtained by the Nested-GA produced perfect classification performance on both the testing and the independent testing datasets compared to other feature selection techniques such as KNN and RF. Additionally, the biological significance of the Microarray genes subsets has been validated using GO and KEGG pathways Enrichment Analysis. The resultant subset of six genes can be used in finding proteins related to colon cancer which are useful in determining the suitable drugs.

Moreover, Nested-GA is significantly able to differentiate between lung cancer subtypes. As a future work, other optimization techniques can be used as a fitness function for the Nested-GA's inner and outer algorithms.

CRediT authorship contribution statement

Sabah Sayed: Acquisition of data, Analysis and interpretation of data, Drafting of manuscript, Critical revision. **Mohammad Nassef:** Acquisition of data, Analysis and interpretation of data, Drafting of manuscript, Critical revision. **Amr Badr:** Acquisition of data, Analysis and interpretation of data, Drafting of manuscript, Critical revision. **Ibrahim Farag:** Analysis and interpretation of data, Critical revision.

References

Abusamra, H. (2013). A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Computer Science, 23*, 5–14.

Ancona, N., Maglietta, R., Piepoli, A., D'Addabbo, A., Cotugno, R., Savino, M., et al. (2006). On the statistical assessment of classifiers using dna microarray data. *BMC Bioinformatics, 7*(1), 387.

Arora, A., Candel, A., Lanford, J., LeDell, E., & Parmar, V. (2015). *Deep learning with H2O*. H2O. ai Inc.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics, 25*(1), 25.

Bandrés, E., Cubedo, E., Agirre, X., Malumbres, R., Zarate, R., Ramirez, N., et al. (2006). Identification by real-time PCR of 13 mature micrornas differentially expressed in colorectal cancer and non-tumoral tissues. *Molecular Cancer, 5*(1), 29.

Barat, A., & Ruskin, H. J. (2015). Comparative correlation structure of colon cancer locus specific methylation: Characterisation of patient profiles and potential markers across 3 array-based datasets. *Journal of Cancer, 6*(8), 795.

Bianchini, M., Levy, E., Zucchini, C., Pinski, V., Macagno, C., De Sanctis, P., et al. (2006). Comparative study of gene expression by CDNA microarray in human colorectal cancer tissues and normal mucosa. *International Journal of Oncology, 29*(1), 83–94.

Cai, Z., Xu, D., Zhang, Q., Zhang, J., Ngai, S.-M., & Shao, J. (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems, 11*(3), 791–800.

Chakraborty, S. (2014). In silico analysis identifies genes common between five primary gastrointestinal cancer sites with potential clinical applications. *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology, 27*(3), 231.

Chin, A. J., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised and semi - supervised feature selection : A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5963d*, 1–20. doi:10.1109/TCBB.2015.2478454.

Coley, D. A. (1999). *An introduction to genetic algorithms for scientists and engineers*. World Scientific Publishing Company.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*(1), 3.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology, 3*(02), 185–205.

Exner, R., Pulverer, W., Diem, M., Spaller, L., Woltering, L., Schreiber, M., Wolf, B., et al. (2015). Potential of DNA methylation in rectal cancer as diagnostic and prognostic biomarkers. *British Journal of Cancer, 113*(7), 1035.

García, V., Sánchez, J., Cleofas-Sánchez, L., Ochoa-Domínguez, H., & López-Orozco, F. (2017). An insight on the large g, small nproblem in gene–expression microarray classification. In *Proceedings of the Iberian conference on pattern recognition and image analysis* (pp. 483–490). Springer.

García, V., & Sánchez, J. S. (2015). Mapping microarray gene expression data into dissimilarity spaces for tumor classification. *Information Sciences, 294*, 362–375.

Gentle, J. E., Härdle, W. K., & Mori, Y. (2010). Springer handbooks of computational statistics.

González, F., & Belanche, L. A. (2013). Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy. arXiv:1302.1733,.

Guttery, D. S., Blighe, K., Polymeros, K., Symonds, R. P., Macip, S., & Moss, E. L. (2018). Racial differences in endometrial cancer molecular portraits in the cancer genome atlas. *Oncotarget, 9*(24), 17093.

Halvey, P. J., Zhang, B., Coffey, R. J., Liebler, D. C., & Slebos, R. J. (2011). Proteomic consequences of a single gene mutation in a colorectal cancer model. *Journal of Proteome Research, 11*(2), 1184–1195.

Han, B., Lai, H., Xie, R., Li, L., & Zhu, L. (2014). Identification of glioma cancer-alerted gene markers based on a diagnostic outcome correlation analysis preferential approach. *International Journal of Data Mining and Bioinformatics, 9*(1), 67–88.

Huerta, E. B., Duval, B., & Hao, J.-K. (2010). A hybrid lda and genetic algorithm for gene selection and classification of microarray data. *Neurocomputing, 73*(13–15), 2375–2383.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Sha, Q., Chen, J., et al. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics, 5*(1), 81.

Jirapech-Umpai, T., & Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics, 6*(1), 148.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research, 40*(D1), D109–D114.

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector algorithm in R. *Journal of Statistical Software, 15*, 1–28.

Kibriya, M. G., Raza, M., Jasmine, F., Roy, S., Paul-Brutus, R., Rahaman, R., et al. (2011). A genome-wide dna methylation study in colorectal carcinoma. *BMC Medical Genomics, 4*(1), 50.

Le, D.-H., Uy, N. Q., Dung, P. Q., Binh, H. T. T., & Kwon, Y.-K. (2013). Towards the identification of disease associated protein complexes. *Procedia Computer Science, 23*, 15–23.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Levner, I. (2005). Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics, 6*(1), 68.

Li, L., Umbach, D. M., Terry, P., & Taylor, J. A. (2004). Application of the GA/KNN method to seldi proteomics data. *Bioinformatics, 20*(10), 1638–1640.

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics, 17*(12), 1131–1142.

Li, W., & Yang, Y. (2002). How many genes are needed for a discriminant microarray data analysis. In *Methods of microarray data analysis* (pp. 137–149). Springer.

Lundemo, A. G., Pettersen, C. H., Berge, K., Berge, R. K., & Schønberg, S. A. (2011). Tetradecylthioacetic acid inhibits proliferation of human sw620 colon cancer cells-gene expression profiling implies endoplasmic reticulum stress. *Lipids in Health and Disease, 10*(1), 190.

Luo, X., Huang, R., Sun, H., Liu, Y., Bi, H., Li, J., et al. (2016). Methylation of a panel of genes in peripheral blood leukocytes is associated with colorectal cancer. *Scientific Reports, 6*, 29922.

Luque-Baena, R., Urda, D., Subirats, J., Franco, L., & Jerez, J. (2013). Analysis of cancer microarray data using constructive neural networks and genetic algorithms. In *Proceedings of the IWBBIO international work-conference on bioinformatics and biomedical engineering* (pp. 55–63).

Malhotra, R., Singh, N., & Singh, Y. (2011). Genetic algorithms: Concepts, design for optimization of process controllers. *Computer and Information Science, 4*(2), 39.

Ooi, C., & Tan, P. (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics, 19*(1), 37–44.

Popovic, D., Sifrim, A., Pavlopoulos, G. A., Moreau, Y., & De Moor, B. (2012). A simple genetic algorithm for biomarker mining. In *Proceedings of the IAPR international conference on pattern recognition in bioinformatics* (pp. 222–232). Springer.

Raza, K., & Jaiswal, R. (2013). Reconstruction and analysis of cancer-specific gene regulatory networks from gene expression profiles. arXiv:1305.5750.

Sastry, K., Goldberg, D. E., & Kendall, G. (2005). *Genetic Algorithms, Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques* (pp. 97–125). Springer: Springer.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517.

Scrucca, L., et al. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software, 53*(4), 1–37.

Shi, T., Mazumdar, T., DeVecchio, J., Duan, Z.-H., Agyeman, A., & Aziz, M. (2010). Cdna microarray gene expression profiling of hedgehog signaling pathway inhibition in human colon cancer cells. *PloS one, 5*(10), e13054.

Sun, X., Liu, Y., Wei, D., Xu, M., Chen, H., & Han, J. (2013). Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis. *Journal of Biomedical Informatics, 46*(2), 252–258. doi:10.1016/j.jbi.2012.10.004.

Valavanis, I., Pilalis, E., Georgiadis, P., Kyrtopoulos, S., & Chatziioannou, A. (2015). Cancer biomarkers from genome-scale dna methylation: comparison of evolutionary and semantic analysis methods. *Microarrays, 4*(4), 647–670.

Vazquez, M., de la Torre, V., & Valencia, A. (2012). Cancer genome analysis. *PLoS Computational Biology, 8*(12), e1002824.

Whitworth, G. B. (2010). An introduction to microarray data analysis and visualization. In *Methods in enzymology: 470* (pp. 19–50). Elsevier.

Xiong, M., Fang, X., & Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research, 11*(11), 1878–1887.

Yeung, K. Y., Bumgarner, R. E., & Raftery, A. E. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics, 21*(10), 2394–2402.

Zhao, L. (2013). *Functional characterization of the candidate colorectal cancer gene CNOT1*. University of Minnesota Ph.D. thesis.