Egyptian Informatics Journal 21 (2020) 13–22

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

An efficient methodology for discovering both of gene-environment interactions and gene-gene interactions causing genetic diseases

Mohamed A. El-Rashidy

Computer Science and Engineering Department, Faculty of Electronic Engineering, Menoufiya University, Egypt

ARTICLE INFO

Article history:
Received 27 July 2019
Revised 13 September 2019
Accepted 8 October 2019
Available online 17 October 2019

Keywords:
Genetic diseases
Feature selection methods
Relief algorithm
Genetic algorithm
Classification decision tree

ABSTRACT

Genetic diseases are one of the most critical diseases facing human societies, their risk lies in the transmission of genetic characteristics from one generation to another, where the imbalance of these characteristics leads to an unhealthy offspring, which negatively affects the effort of this offspring and its services to society. Genetic disease is caused by a mutation in the Deoxyribonucleic Acid (DNA), these genetic mutations are generated by nonlinear interactions between two or more genes and / or environmental exposures. The aim of this paper is to discover both of gene-environment interactions and gene-gene interactions causing a genetic disease, the proposed methodology is based on both of the filter and wrapper feature selection methods, it uses the filter method using a Relief algorithm to detect the gene-environment interactions, wrapper method using genetic algorithm to discover gene-gene interactions, and classification decision tree algorithm to generate the conditional rules of gene-gene interactions. It has been evaluated using many different classifier models on four benchmark databases, and compared its performance with an Apriori algorithm for generating rules of gene-gene interactions, the proposed methodology achieved the highest performance and better classification accuracy on all databases containing patients affected by gene-environment interactions or gene-gene interactions or both of gene-environment and gene-gene interactions.

1. Introduction

Genetic diseases have a major impact on health care systems and an effective factor in increasing the mortality rate of newborns, children and adults. Genes are the building blocks of heredity, so that they carry the Deoxyribonucleic Acid (DNA) and instructions related to the manufacture of proteins. Sometimes one or more genes may be exposed to a certain mutagenic change, which affects the instructions of the involved genes manufacturing the proteins, this prevents the protein from functioning properly or may lose its ability to function fully, which leads to genetic diseases as cancer, nearly 13 percent of the world deaths are caused due to cancer diseases. Genetic disease is defined as any disorder caused by abnormalities in a person's genetic material.

These diseases are very complex diseases, which many phenomena can lead to genetic mutation including nonlinear interactions between two or more genes "gene-gene interactions" and / or environmental exposures "gene-environment interactions" such as exposure to x-rays, radium, ultraviolet radiations and chemicals, smoking, and excess diet. Awareness of these interactions plays an important role in better understanding the development of these complex diseases. So recently, Genome-Wide Association Studies (GWAS) have been significantly working to study and understand the genome-wide interaction studies. Since then, there have been several studies aimed to describe the effects of gene-gene (G-G) interactions and gene-environment (G-E) interactions that lead to genetic disease [1]. There are many challenges associated with an understanding of these interactions, the major challenge is the high dimensionality of genetic data that needs to be analyzed, which leads to high computational complexity of traditional statistical approaches that analyze large-scale genetic data [2].

E-mail address: mohamed.elrashedy@el-eng.menofia.edu.eg

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2019.10.001

Recently, many researches that based on surrogate modelling and machine learning approaches have been done to overcome the challenges of the different fields of science as autonomic machine learning platform [3], reservoir-based surrogate modeling of dynamic user equilibrium [4], the application of machine learning to studies on plant response to radio-frequency [5], designing mesoscale structure of Li-Ion battery electrodes using machine learning approaches [6], developing surrogate indicators for predicting suppression of halophenols formation potential and abatement of estrogenic activity during ozonation of water [7], drug development [8] and medical diagnosis using machine learning approaches [9,10], integration of artificial intelligence (AI) approaches to tackle the challenges of scalability and high dimensionality data in cancer genetic testing and diagnostics [11].

The researches that based on machine learning approaches to overcome the challenges of understanding the mutagenic changes divided into two types, the first type is interested in selecting the subset of genes associated with gene-environment interactions. New artificial bee colony approach has been proposed to find the best subset of genes related to gene-environment interactions, it is based on independent component analysis approach to reduce the size of data and artificial bee colony approach to optimize the reduced genes [12]. Recursive Memetic Algorithm (RMA) model for gene selection has been developed, it is a variant of Memetic Algorithm (MA) and performs much better than MA as well as genetic algorithm [13]. Bimodal unsupervised dimension reduction algorithm (BOUNDER) has been implemented to identify the best subset of genes for downstream analysis considering the variety and redundancy across all the samples using bimodal modelling before it feeds into the learning method [14]. A novel Markov blanket-embedded genetic algorithm (MBEGA) for gene selection has been proposed, it based on memetic operators to add or delete genes from a genetic algorithm (GA) solution to quickly improve the solution and fine-tune the search [15]. A new combination of Teaching Learning-based Optimization (TLBO) and Simulated Annealing (SA) algorithm with Support Vector Machine to identify the subset of the most informative genes that can contribute to the accurate detection of cancer [16].

The other type of the researches is concerned to discover the rules of gene-gene interactions. Ant colony optimization approach has been used for the discovery of gene-gene interactions in genome-wide association study data [17,18]. Clustering and association rule mining approaches have been developed for analysing Alzheimer's disease gene expression to uncover gene-gene interactions [19]. A novel parallel association rules extractor from SNPs has been implemented as a multi-threaded version of an optimized version of the Frequent Pattern Growth (FP-Growth) algorithm for the extraction of association rules from Omics datasets [20]. A new computational method using deep learning models to predict enhancer-promoter interactions based on sequence-based features only [21]. A deep neural network was improved by avoiding overfitting for an intensive search of gene-gene interactions [22].

The different types of previous researches are interested in the diseases that affected by one type of the mutagenic changes, which select the subset of genes related to gene-environment interactions, or discover the rules of gene-gene interactions, they ignored the complex genetic diseases that affected by both of gene-gene (G-G) and gene-environment (G-E) interactions. The aim of this research is to discover both of gene-environment interactions and gene-gene interactions causing a disease with the lowest complexity time and highest accuracy, which helps to better understand the complex genetic diseases that affected by all the types of mutagenic changes, including G-G interactions, or G-E interactions, or both of G-G and G-E interactions, it takes into consideration all the reasons leading to the genetic diseases without ignoring some of them. The proposed methodology consists of three stages; data pre-processing, discovery of gene-gene and gene-environment interactions causing a disease, and classification of genetic disease cases considering all mutagenic changes. The proposed methodology is based on a Relief algorithm to select the subset of genes relating to gene-environment interactions, genetic algorithm to discover the subset of genes relating to gene-gene interactions, and categorical classification decision tree algorithm to generate the conditional rules of gene-gene interactions.

It has been evaluated using four benchmark databases of gene microarrays that available on the genomics data repository of the NCBI databases, the proposed methodology has been implemented with the values of all dataset genes, values of selected subset genes relating to gene-environment interactions and discovered rules of gene-gene interactions. The efficiency of proposed methodology has been verified by using different classification algorithms SVM, Naïve Bayes, decision tree, and k nearest neighbour algorithms to construct the classifier models of gene-environment interactions and compared with Apriori algorithm for generating rules of gene-gene interactions. The proposed methodology achieved the best classification accuracy on all databases containing samples affected by gene-environment interactions or gene-gene interactions or both of gene-environment and gene-gene interactions.

The rest of paper describes the proposed methodology, presents and discusses the experimental results, Finally, the last section concludes this research.

2. Proposed methodology

The identification of genetic mutations causing a disease is based on the relationships of genotype-phenotype, which are represented in the interactions between genetic variants and environment, there include one or both of gene-environment and gene-gene interactions.

The aim of the proposed methodology is to discover both of gene-environment and gene-gene interactions causing a disease, which helps to better understand and achieves a higher classification accuracy of complex genetic diseases. The proposed methodology consists of three stages as shown in Fig. 1, the first stage prepresses the data to be quality data, the second stage reduces the data dimensions to reduce overfitting, improve classification accuracy, and reduce training time, it also analyses the selected gene values with the output predictions to discover both of gene-environment and gene-gene interactions, the third stage classifies the testing samples considering all mutagenic changes.

2.1. Stage I: Data preprocessing

Gene expression databases *DD* are extremely vulnerable to noisy and incomplete data due to their integration from multiple and heterogeneous sources, and their huge size, noisy and incomplete data will lead to low-quality results of data analysis and knowledge discovery. There are several data pre-processing techniques included data cleaning and transformation processes are used to achieve high quality data.

Data cleaning process is used to analyze the data discrepancy, which applied to detect and remove outliers, and fill incomplete data as shown in Fig. 2.

Outliers are detected and removed by clustering technique, where similar data objects are grouped into clusters, data objects that fall outside of the clusters are considered outliers and removed. Incomplete data is handled by regression technique, that is compiled the values of data objects to a linear function and used one or more genes to predict the other incomplete gene values.

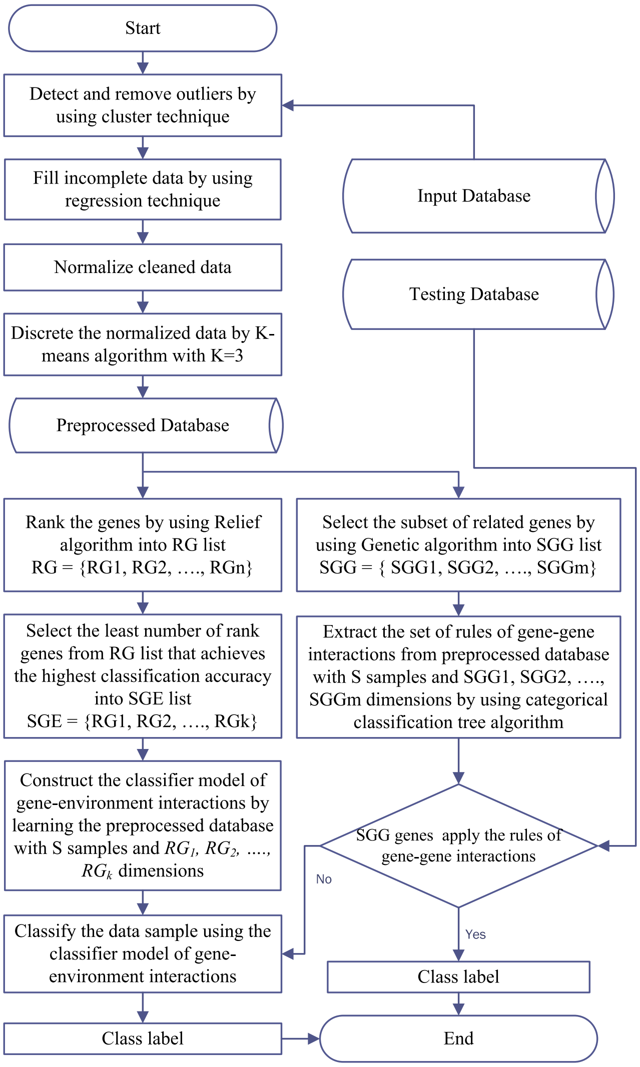

Start
Detect and remove outliers by using cluster technique
Input Database
Fill incomplete data by using regression technique
Normalize cleaned data
Testing Database
Discrete the normalized data by K-means algorithm with K=3
Preprocessed Database
Rank the genes by using Relief algorithm into RG list RG = {RG1, RG2, …., RGn}
Select the subset of related genes by using Genetic algorithm into SGG list SGG = { SGG1, SGG2, …., SGGm}
Select the least number of rank genes from RG list that achieves the highest classification accuracy into SGE list SGE = {RG1, RG2, …., RGk}
Extract the set of rules of gene-gene interactions from preprocessed database with S samples and SGG1, SGG2, …., SGGm dimensions by using categorical classification tree algorithm
Construct the classifier model of gene-environment interactions by learning the preprocessed database with S samples and RG_1, RG_2, …., RG_k dimensions
SGG genes apply the rules of gene-gene interactions
No
Yes
Classify the data sample using the classifier model of gene-environment interactions
Class label
Class label
End

Fig. 1. Flowchart of the proposed methodology.

Data transformation is used to make the values of genes more accurate, efficient and appropriate to the inputs of mining algorithms, data normalization applied on the cleaned data set *CD* to give the values of all genes an equal weight. The normalized values of genes "*ND* dataset" are continuous numbers, where the Apriori algorithm cannot be applied to continuous attributes that have infinite values. Therefore, the transformation process is used to convert genes data from continuous into discrete values. Clustering technique is a popular data discretization method [23], which takes into consideration the distribution of gene values instead of hypothesized values for partitioning boundaries of the other discretization techniques. Each gene has two possible alleles {*A*, *B*}, which lead to three possible genotypes (*AA, AB/BA, and BB*) [1]. Therefore, K-means algorithm is applied to partition the values of each gene into three clusters and each gene value of cluster number one, two and three are replaced to labels *L1, L2* and *L3* respectively.

2.2. Stage II: Model construction

Model construction stage consists of two steps, the first step is used to select the subset of genes interacting with environment and construct the classification model of gene-environment interactions. The second step is applied to extract the subset of genes interacting with the others and discover the conditional rules of gene-gene interactions.

2.2.1. Step1: Model construction of gene-environment interactions

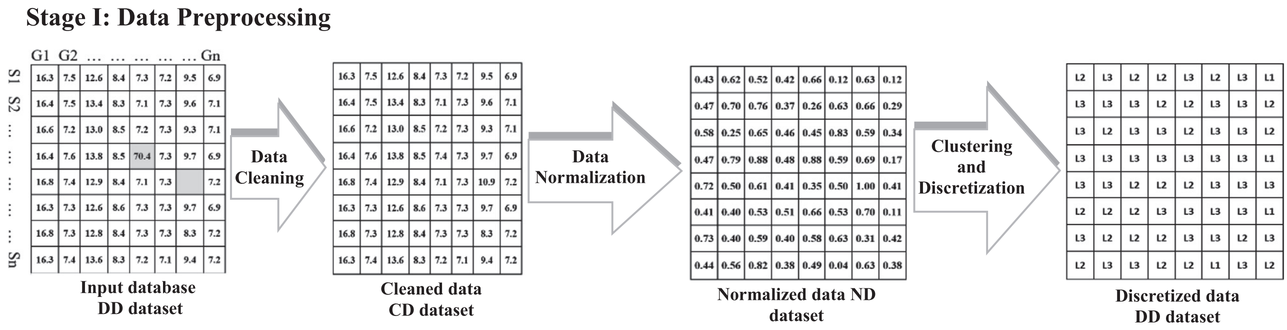
Genes interacting with an environment "G-E Interactions" are not related to each other, but each of them related to external environmental effects, which lead to each of them individually affected to the output prediction. Therefore, filter approach of the feature selection process is used to select the gene-environment interactions that more contribute to the output prediction, it depends on the importance of each gene to the output prediction, which gives a score for each gene expression, the higher the score value indicates to more relevant to the output prediction values.

Relief algorithm is one of the most common filter techniques for feature selection due to its simplicity and effectiveness [24], it is applied to rank the genes *G* depending on their importance to the output prediction values, it computes a gene weight *Wg* for each gene *g* to estimate gene relevance to the output prediction that can range from –*S* (worst) to +*S* (best), where S is the number of samples, it finds the nearest two neighbors of a sample S using Euclidian distance, one from the same class called the nearest hit *NH* and the other from the opposite class called the nearest miss *NM*, the weight *Wg* of each gene is computed as:

$$\mathrm{Wg} = \mathrm{Wg} + \mathrm{diff}(\mathrm{Sg} - \mathrm{NMg}) - \mathrm{diff}(\mathrm{Sg} - \mathrm{NHg}) \tag{1}$$

$$\mathrm{diff}(\mathrm{Ag} - \mathrm{Bg}) = \begin{cases} 0; & value(Ag) = value(Bg) \\ 1; & otherwise \end{cases} \tag{2}$$

where *diff(Ag – Bg)* is the distance between the values of gene *g* to the two samples *A* and *B*. Weight of each gene *g* is computed by iterating over all samples and started with *Wg = 0* [24]. Therefore, the complexity time of it is O(G*S), where G is the number of genes and S is the number of samples. The genes are ranked depending into *Wg* into *RG* list, where *RG = {RG_1, RG_2, ..., RG_n}*, RG_1 is the gene



Input database DD dataset

	G1	G2	…	…	…	…	…	Gn
S1	16.3	7.5	12.6	8.4	7.3	7.2	9.5	6.9
S2	16.4	7.5	13.4	8.3	7.1	7.3	9.6	7.1
…	16.6	7.2	13.0	8.5	7.2	7.3	9.3	7.1
…	16.4	7.6	13.8	8.5	70.4	7.3	9.7	6.9
…	16.8	7.4	12.9	8.4	7.1	7.3		7.2
…	16.3	7.3	12.6	8.6	7.3	7.3	9.7	6.9
…	16.8	7.3	12.8	8.4	7.3	7.3	8.3	7.2
Sn	16.3	7.4	13.6	8.3	7.2	7.1	9.4	7.2

Cleaned data CD dataset

16.3	7.5	12.6	8.4	7.3	7.2	9.5	6.9
16.4	7.5	13.4	8.3	7.1	7.3	9.6	7.1
16.6	7.2	13.0	8.5	7.2	7.3	9.3	7.1
16.4	7.6	13.8	8.5	7.4	7.3	9.7	6.9
16.8	7.4	12.9	8.4	7.1	7.3	10.9	7.2
16.3	7.3	12.6	8.6	7.3	7.3	9.7	6.9
16.8	7.3	12.8	8.4	7.3	7.3	8.3	7.2
16.3	7.4	13.6	8.3	7.2	7.1	9.4	7.2

Normalized data ND dataset

0.43	0.62	0.52	0.42	0.66	0.12	0.63	0.12
0.47	0.70	0.76	0.37	0.26	0.63	0.66	0.29
0.58	0.25	0.65	0.46	0.45	0.83	0.59	0.34
0.47	0.79	0.88	0.48	0.88	0.59	0.69	0.17
0.72	0.50	0.61	0.41	0.35	0.50	1.00	0.41
0.41	0.40	0.53	0.51	0.66	0.53	0.70	0.11
0.73	0.40	0.59	0.40	0.58	0.63	0.31	0.42
0.44	0.56	0.82	0.38	0.49	0.04	0.63	0.38

Discretized data DD dataset

L2	L3	L2	L2	L3	L2	L3	L1
L3	L3	L3	L2	L2	L3	L3	L2
L3	L2	L3	L3	L2	L3	L3	L2
L3	L3	L3	L3	L3	L3	L3	L1
L3	L3	L2	L2	L2	L3	L3	L3
L2	L2	L2	L3	L3	L3	L3	L1
L3	L2	L2	L2	L3	L3	L2	L3
L2	L3	L3	L2	L2	L1	L3	L2

Fig. 2. Data preprocessing stage.

expression with the highest weight, RG_n is the gene expression with the lowest weight, and n is the number of genes. The subset of genes interacting with environment $SGE = \{RG_1, RG_2, \ldots, RG_k\}$, where k is the least number of rank genes in RG list that achieves the highest classification accuracy.

Classifier model of gene-environment interactions is constructed by learning the training dataset, which each sample of the training dataset is represented by the values of k-dimensional genes $RG_1, RG_2, \ldots, RG_k$ and associated with an output label Positive "*P*" or Negative "*N*". The output model is represented by mathematical formulae, or decision tree, or artificial neural network depending on the applied classification algorithm of the learning process.

2.2.2. Step2: Rules extraction of gene-gene interaction

Genes affected by gene-gene interaction type "G-G Interactions" are related to each other and inferred to the output prediction. Filter approach of the feature selection process depends on the importance of each individual gene to the output prediction without considering the importance of each related subset of genes to the output prediction values. The search process about the subset of related genes from the tremendous probabilities of the subsets *NPS* Eq. (3) is NP-hard problem.

$$NPS = \sum_{i=1}^{n} \frac{n!}{(n-i)!} \tag{3}$$

Genetic algorithm GA is a search and optimization technique for numerous non-deterministic polynomial-time hard NP-hard of combinatorial optimization problems. Therefore, genetic algorithm is used as a wrapper method to select the subset of related genes that are most closely related to the output prediction from the tremendous probabilities of the subsets *NPS*. A simple GA uses a population of chromosomes to solve a given problem. Each subset of genes is encoded as a chromosome, where it is a possible solution for the problem. A genetic algorithm starts generating an initial population formed by a random group of chromosomes, which can be seen as first guesses to solve the problem. The initial population is evaluated by fitness function and for each chromosome a fitness value is given reflecting the quality of the solution associated to it. At each step, the genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the children for the next generation. Over successive generations, the population evolves toward an optimal solution. Fitness function for gene-gene interactions is focused on the subsets of related genes that are most closely related to the output prediction, the quality of the subset of genes selection is measured by the classification error rate metric eq. (4), where *FS* is the number of incorrect classified samples. The subset of genes $SGG = \{SGG_1, SGG_2, \ldots, SGG_m\}$ that has the lowest classification error rate is selected as the optimized solution.

$$\text{Classification Error Rate} = \text{FS}/\text{total samples} \tag{4}$$

The set of conditional rules "rules of G-G interactions" that related to the output predicted values will be extracted from *SGG* dataset, which each sample of the training dataset is represented by the values of m-dimensional genes $SGG_1, SGG_2, \ldots, SGG_m$ and associated with an output label "*P*" or "*N*", each conditional rule consists of left-hand side called a dimensional conditions and right-hand side called a predictor Eq. (5),

$$\begin{aligned} &SGGx(C1), SGGy(C2), \ldots \rightarrow \text{Output}(O) \\ &\text{Dimensional conditions} \rightarrow \text{Predictor} \end{aligned} \tag{5}$$

where x and $y \in m$-dimensional, *C1* and $C2 \in$ discretized labels {*L1, L2, L3*} and $O \in$ output labels {*P, N*}, these rules reflect the closed relationship between each frequent subset of m-dimensional genes "dimensional conditions" to the output prediction "predictor".

Apriori algorithm is used in many recent researches [19,25,26] to extract all frequent values of dataset genes related to the output labels, and set minimum confidence equals one for minimizing the percentage error of the classification process and indicating to the closed relationship between the frequent set and the output prediction. Apriori algorithm is applied to compute the support and confidence values for every mined possible rule and filter the mined rules to apply the constrained values of support and confidence, "rules of G-G interactions" are extracted from the filtered rules relating to the output predictor values. Apriori algorithm is prohibitively expensive because there are exponentially many rules that can be extracted from a data set. More specifically, the total number of possible rules mined from a data set that contains *U* unique values is [27]

$$\textit{possible rules} = 3^u - 2^{u+1} + 1 \tag{6}$$

The number of unique values for each *m* dimensions of the *SGG* dataset is three, which each value of the m dimensions ∈ discretized labels {*L1, L2, L3*}, the number of unique items for output attribute is two, which each value of it ∈ output class labels {*P, N*}. Therefore, the total number of unique values for the *SGG* dataset is *(3*m + 2)*, and according to Eq. (6) the complexity time to extract the set of rules of "G-G interactions" using Apriori algorithm is

$$O\left(3^{(3*m+2)} - 2^{(3*m+3)} + 1\right) \tag{7}$$

the operational method of Apriori algorithm using to extract the rules of gene-gene interactions in recent researches is based on the concept of frequent values of genes with the two output *P* and *N*. This concept is mined all possible rules, without considering the distinctively frequent values of related genes to the outputs, where the change in values of dimensional conditions "*C1, C2,……*" may not affect to the predictor value. So, the numerous mined rules using Apriori algorithm are not reliable to discover the distinctive rules achieving the meaning of gene-gene interactions.

Therefore, the proposed methodology in this step is elaborated to find the distinctive rules between the two outputs indicating to gene-gene interactions. It is based on the categorical classification tree to construct the decision tree of the *SGG* dataset and generate the distinctive rules "rules of gene-gene interactions" from the constructed decision tree. The complexity time of categorical classification tree is *O(Smd)*, where *S* is the number of training samples, *m* is the number of dimensions, and *d* is the depth of classification tree. In the best case of the balanced classification tree, *d* will equal *logS*, but the classification tree splits training data without considering about balancing. This means that the worst case of *d* will equal *N*. So, the time complexity of classification tree is in between *O(SmlogS)* and *O(S2m)*. The proposed methodology uses the categorical classification tree to extract the rules of gene-gene interactions, which is better performance than the Apriori algorithm in the complexity time and accuracy, where it is considered the distinctive values of *m* dimensions between the two classification outputs of *SGG* dataset as a conditions into the splitting data process for constructing a classification tree model. The proposed methodology uses support and confidence metric as a constrained value to filter the conditional rules of classification tree, which extracts the rules that minimize the percentage error into the stage III, the filtered rules are ordered by confidence then support values to the list of rules of gene-gene interactions.

2.3. Stage III: Model construction

In the third stage, the constructed classifier model of gene-environment interactions and discovered rules of gene-gene interactions from the second stage are used for classifying the test data.

The values of genes for each test data sample are tested using the rules of gene-gene interactions to classify it, if the values of SGG genes do not apply the rules of gene-gene interactions, then the values of SGE genes is classified using the constructed classifier model of gene-environment interactions.

3. Results and discussion

Four benchmark binary classification databases of gene microarrays are used to evaluate the proposed methodology, including the GSE25070, GSE9476, GSE10950, and GSE6919. The data is available in the genomics data repository of the NCBI databases, Table 1 shows the detailed description of these databases.

Stage I of the proposed methodology has been applied on the four databases for preprocessing data to achieve high quality data free of outliers and missing data, the preprocessed databases have been discretized to appropriate the stage II inputs of the proposed methodology. The performance of proposed methodology has been measured by partitioning each discretized database into independent training and testing datasets using 5-fold cross validation technique, testing datasets have been used to evaluate predictive accuracy of the proposed methodology, accuracy metric is one of the most commonly used metrics for the classification performance, which measures the percentage of correctly classified samples of the testing datasets. Number of incorrectly classified samples have also been measured to indicate the number of complex disease cases that need more developing approaches to discover it, where the decrease in this number indicates the development of the predictive model for detecting complex disease cases comparing with the other predictive models.

Stage II of the proposed methodology has been applied to the training datasets for constructing a classifier model of gene-environment interactions and discovering the rules of gene-gene interactions. It consists of two steps, first step uses Relief algorithm to rank the genes of training dataset depending on their importance to the output prediction values, the classification accuracy percentage of each subset of the consecutive genes in the ranked list is different, which can be seen from Figs. 3–6.

The proposed methodology selects the subset of least number of consecutive genes in the ranked list *RG* achieving the highest classification accuracy, which it represents to the subset of genes interacting with an environment and responsible to output change, the least number of consecutive genes achieving the highest classification accuracy for each database listed in Table 2. Classifier model of gene- environment interactions has been constructed training by the values of selected genes.

Stage III of the proposed methodology has been applied onto the testing datasets, the proposed methodology has been evaluated and compared with the different classifier models training on the values of all dataset genes, the subset of genes relating to gene-environment interactions, and the rules of gene-gene interactions.

Table 1
Description of the databases.

Database	Number of Genes	Number of Samples	Description
GSE25070	24,526	52	Gene expression dataset of 26 colorectal tumor patients and 26 normal donors, public on Jun 03, 2011.
GSE9476	22,283	64	Gene expression dataset of 38 normal donors and 26 Acute myeloid leukemia (AML) patients, public on Nov 01, 2007.
GSE10950	22,184	48	Gene expression dataset of 24 normal donors and 24 colon tumor patients, public on Mar 29, 2008
GSE6919	12,625	146	Gene expression dataset of 81 normal donors and 65 Prostate tumor patients, public on Jan 30, 2007.

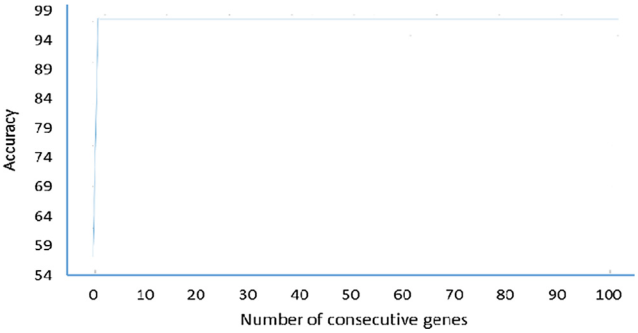


Fig. 3. Classification accuracy percentage of the different subsets of consecutive genes in RG list on GSE25070 database.

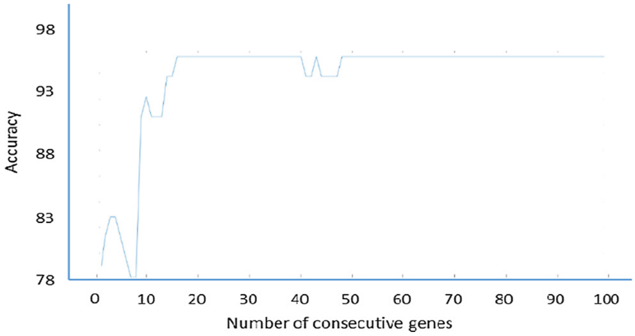


Fig. 4. Classification accuracy percentage of the different subsets of consecutive genes in RG list on GSE9476 database.

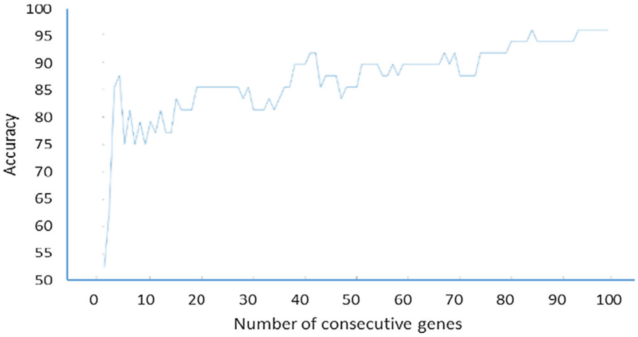


Fig. 5. Classification accuracy percentage of the different subsets of consecutive genes in RG list on GSE10950 database.

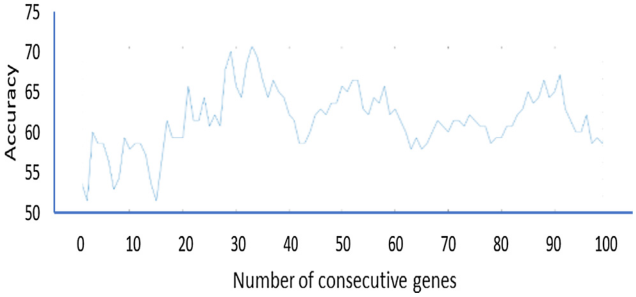


Fig. 6. Classification accuracy percentage of the different subsets of consecutive genes in RG list on GSE6919 database.

Table 2
Least number of consecutive genes in the ranked list RG achieving highest classification accuracy.

Database	Highest accuracy achieved by the subset of ranked genes	Number of ranked genes
GSE25070	98.08%	2
GSE9476	95.31%	16
GSE10950	95.75%	85
GSE6919	71.92%	38

The efficiency of proposed methodology has been verified by using different classification algorithms SVM, Naïve Bayes, decision tree, and k nearest neighbor algorithms to construct the classifier models.

Tables 3–6 show the accuracy of classification algorithms training with the values of all dataset genes and another experiment for training these algorithms with the values of selected genes of G-E interactions for each database. The constructed classifier models have been evaluated using testing datasets, the highlighted bold font of the results shows that the classification algorithms training with the values of selected genes of G-E interactions obtained higher classification accuracy than training these algorithms with the values of all dataset genes.

The results also indicate the existence of incorrectly classified instances, although there is an enhancement into the classification accuracy compared to the classifier models basing on all dataset genes, these indications demonstrate that only dependence on selected genes relating to G-E interactions are not adequate in the classification process. The pathological case of the patient may be changed due to G-E or G-G interactions. Therefore, the proposed methodology stages based on both of G-E and G-G interactions to overcome incorrectly classified samples as much as possible.

The second step of stage II has been applied to discover the rules of gene-gene interactions, it uses a genetic algorithm to select the subset of related genes that are most closely related to the output prediction values, Table 7 shows the number of selected genes that is responsible of gene-gene interactions.

The proposed methodology is based on categorical classification decision tree algorithm to generate the conditional rules of gene-gene interactions from the set of selected genes *SGG*, Tables 8–11 show the comparison between Apriori and decision tree algorithms with different support values for each database.

Number of generated rules, testing samples that verify the generated rules, and incorrectly classified samples have been measured to evaluate these algorithms for generating G-G interaction rules, where the decrease number of generated rules with increasing number of testing samples that verify these rules indicates to increase the performance with decreasing the computational complexity of the model. It can be seen from the results of the comparison, the decision tree is better performance than Apriori algorithm, where the number of generated rules using decision tree is excessively less than the number of generated rules of Apriori algorithm.

Not only the great improvement on the complexity, but it also achieved enhancement on the classification accuracy obtained by applying the generated rules of decision tree rather than Apriori algorithm with all different support values. The results of the comparison also clarify the effectiveness of support values, where show that increasing support value leads to decrease the number of incorrectly classified samples. Therefore, the proposed methodology sets the support value equals 50% of the training samples for each predictor value, which reflects the robustness relationship of the generated rule with the output prediction values.

Tables 12–15 show comparison between the proposed methodology that based on both of the constructed G-E interactions model and G-G interaction rules generating using decision tree algorithm,

Table 3
Classification accuracy of the classifier models basing on all dataset genes and the classifier models basing on the selected genes of G-E interactions for database GSE25070.

Classification algorithm	Training by the data of all genes		Training by the data of selected G-E genes	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
SVM	96.15%	2	98.08%	1
Naïve Bayes	96.15%	2	96.15%	2
Decision Tree	84.62%	8	96.15%	2
k nearest	94.23%	3	96.15%	2

Table 4
Classification accuracy of the classifier models basing on all dataset genes and the classifier models basing on the selected genes of G-E interactions for database GSE9476.

Classification algorithm	Training by the data of all genes		Training by the data of selected G-E genes	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
SVM	95.31%	3	95.31%	3
Naïve Bayes	95.31%	3	95.31%	3
Decision Tree	81.25%	12	95.31%	3
k nearest	92.19%	5	96.88%	2

Table 5
Classification accuracy of the classifier models basing on all dataset genes and the classifier models basing on the selected genes of G-E interactions for database GSE10950.

Classification algorithm	Training by the data of all genes		Training by the data of selected G-E genes	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
SVM	93.75%	3	**95.83%**	2
Naïve Bayes	95.83%	2	95.83%	2
Decision Tree	93.75%	3	93.75%	3
k nearest	93.75%	3	95.83%	2

Table 6
Classification accuracy of the classifier models basing on all dataset genes and the classifier models basing on the selected genes of G-E interactions for database GSE6919.

Classification algorithm	Training by the data of all genes		Training by the data of selected G-E genes	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
SVM	65.75%	50	71.92%	41
Naïve Bayes	69.86%	44	73.29%	39
Decision Tree	54.11%	67	66.44%	49
k nearest	56.85%	63	69.18%	45

Table 7
Number of related genes responsible of G-G interactions using genetic algorithm.

Database	Number of genes
GSE25070	9
GSE9476	19
GSE10950	7
GSE6919	79

Table 8
Comparison between Apriori and decision tree algorithms to generate the rules of G-G interactions of database GSE25070.

Constraint values	Apriori			Decision tree		
	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples
Sup = 10% & Conf = 100%	275	52	5	7	8	1
Sup = 20% & Conf = 100%	138	51	5	5	8	1
Sup = 30% & Conf = 100%	69	45	4	2	4	0
Sup = 40% & Conf = 100%	31	34	4	2	4	0
Sup = 50% & Conf = 100%	7	16	1	2	4	0

Table 9
Comparison between Apriori and decision tree algorithms to generate the rules of G-G interactions of database GSE9476.

Constraint values	Apriori			Decision tree		
	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples
Sup = 10% & Conf = 100%	47,119	46	11	7	5	1
Sup = 20% & Conf = 100%	1200	31	6	6	5	1
Sup = 30% & Conf = 100%	253	26	6	5	4	1
Sup = 40% & Conf = 100%	59	19	4	5	4	1
Sup = 50% & Conf = 100%	11	5	9	1	3	0

Table 10
Comparison between Apriori and decision tree algorithms to generate the rules of G-G interactions of database GSE10950.

Constraint values	Apriori			Decision tree		
	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples
Sup = 10% & Conf = 100%	173	45	3	1	3	0
Sup = 20% & Conf = 100%	85	38	3	0	0	0
Sup = 30% & Conf = 100%	48	30	2	0	0	0
Sup = 40% & Conf = 100%	28	18	0	0	0	0
Sup = 50% & Conf = 100%	15	10	0	0	0	0

Table 11
Comparison between Apriori and decision tree algorithms to generate the rules of G-G interactions of database GSE6919.

Constraint values	Apriori			Decision tree		
	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples	No. of rules of G-G interactions	No. of samples applying extracted rules	No. of incorrectly classified samples
Sup = 10% & Conf = 100%	396	89	27	24	45	14
Sup = 20% & Conf = 100%	175	78	23	6	27	8
Sup = 30% & Conf = 100%	79	56	12	3	14	3
Sup = 40% & Conf = 100%	39	43	10	2	10	2
Sup = 50% & Conf = 100%	13	22	4	2	10	2

Table 12
Classification accuracy of the proposed methodology based on SVM classifier model for G-E interactions and two different methods for generating G-G interaction rules with support 50% and confidence 100%.

Database	Classifier model of G-E interactions and G-G interaction rules using Apriori algorithm		Classifier model of G-E interactions and G-G interaction rules using decision tree algorithm	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
GSE25070	96.15%	2	100%	0
GSE9476	95.31%	3	96.88%	2
GSE10950	95.83%	2	95.83%	2
GSE6919	72.60%	40	76.03%	35

Table 13
Classification accuracy of the proposed methodology based on Naïve Bayes classifier model of G-E interactions and two different methods for generating G-G interaction rules with support 50% and confidence 100%.

Database	Classifier model of G-E interactions and G-G interaction rules using Apriori algorithm		Classifier model of G-E interactions and G-G interaction rules using decision tree algorithm	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
GSE25070	94.12%	3	98.08%	1
GSE9476	95.31%	3	96.88%	2
GSE10950	95.83%	2	95.83%	2
GSE6919	73.29%	39	76.71%	34

Table 14
Classification accuracy of the proposed methodology based on decision tree classifier model of G-E interactions and two different methods for generating G-G interaction rules with support 50% and confidence 100%.

Database	Classifier model of G-E interactions and G-G interaction rules using Apriori algorithm		Classifier model of G-E interactions and G-G interaction rules using decision tree algorithm	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
GSE25070	96.15%	2	96.15%	2
GSE9476	93.75%	4	95.31%	3
GSE10950	93.75%	3	93.75%	3
GSE6919	65.07%	51	69.86%	44

Table 15
Classification accuracy of the proposed methodology based on k nearest neighbor classifier model of G-E interactions and two different methods for generating G-G interaction rules with support 50% and confidence 100%.

Database	Classifier model of G-E interactions and G-G interaction rules using Apriori algorithm		Classifier model of G-E interactions and G-G interaction rules using decision tree algorithm	
	Accuracy	Number of incorrectly classified samples	Accuracy	Number of incorrectly classified samples
GSE25070	94.23%	3	98.08%	1
GSE9476	95.31%	3	96.88%	2
GSE10950	95.83%	2	95.83%	2
GSE6919	65.07%	51	73.97%	38

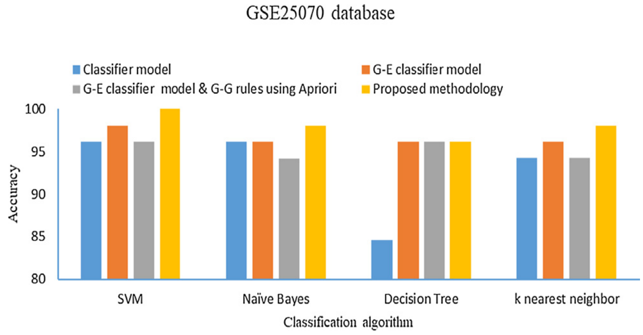


Fig. 7. Comparison between different classification methodologies on GSE25070 database.

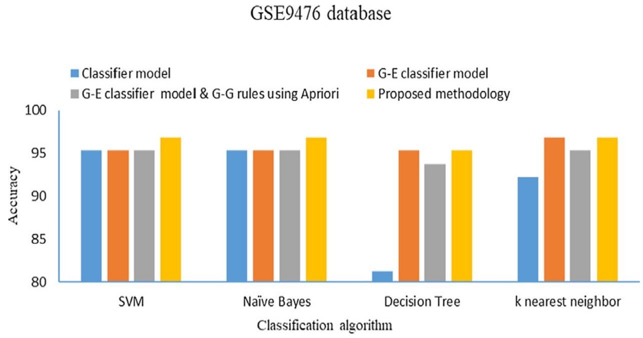


Fig. 8. Comparison between different classification methodologies on GSE9476 database.

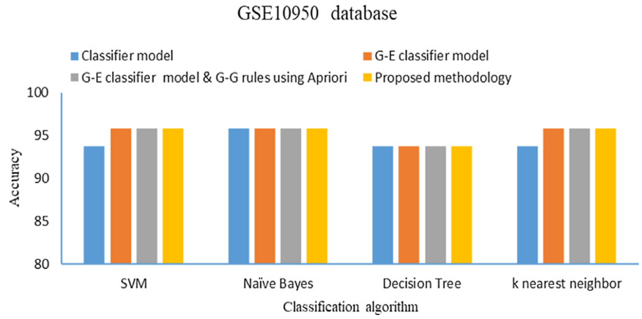


Fig. 9. Comparison between different classification methodologies on GSE10950 database.

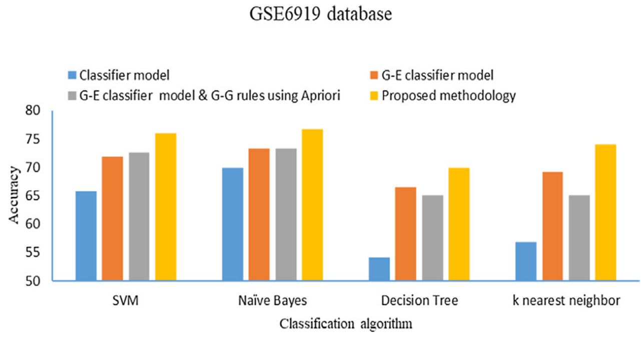


Fig. 10. Comparison between different classification methodologies on GSE6919 database.

and another experiment for the proposed methodology to generate G-G interaction rules using Apriori algorithm into stage II, two experiments set support equaled 50% and confidence equaled 100%, the proposed methodology based on decision tree for generating G-G interaction rules obtained higher classification accuracy and lowest number of incorrectly classified samples.

It achieved these results with the different classifier models on all databases, which verified the efficient of the proposed methodology with the different classification algorithms.

Figs. 7–10 show the comparison between the classification accuracy of classifier models training with the values of all dataset genes, G-E classifier models training with the values of selected subset genes of G-E interactions, G-E classifier models and G-G interaction rules using Apriori algorithm and the proposed methodology that based on G-E classifier models and G-G interaction rules using a decision tree algorithm. It can be seen from the figures, the proposed methodology achieved higher classification accuracy compared with the other techniques for databases GSE25070, GSE9476, and GSE6919, which indicated that these databases include samples affecting by gene-gene interactions. The results also show, the proposed methodology obtained an equally classification accuracy with the G-E classifier models training with the values of selected subset genes of G-E interactions on database GSE10950, which indicated that this database does not include samples affecting by gene-gene interactions. The results demonstrate that the proposed methodology achieved the best classification accuracy on all databases containing samples affected by gene-environment interactions or gene-gene interactions or both of gene-environment and gene-gene interactions.

It can also be seen from the achieved results that the accuracy of the proposed methodology varies across the different databases, this variance is due to the data characteristics of each database. DNA contains millions of genes, and the available genetic databases contain few genes compared to their total number, the genes available in the database may be affecting the incidence of the disease, and they may be part of the effect and the other parts are related to the genes that are not available in the database. Therefore, in order to achieve the best possible accuracy, genetic databases must be available with a larger number of possible genes to study and understand the gene interactions with the largest number of possible causes of the genetic diseases.

4. Conclusion

In this paper, the proposed methodology introduced to diagnosis the genetic diseases, it depended on Relief, genetic and classification decision tree algorithms to detect the subset of genes responsible of all different genetic mutations in DNA. Comprehensive study conducted to evaluate the proposed methodology, where it used SVM, Naïve Bayes, decision tree, and k nearest neighbor classification algorithms on four different genetic diseases to construct the classifier model of gene-environment interactions and compared with an Apriori algorithm for generating rules of gene-gene interactions. The proposed methodology achieved higher performance than Apriori algorithm and best classification accuracy with the different classification algorithms for all the databases containing patients affected by gene-environment interactions or gene-gene interactions or both of gene-environment and gene-gene interactions.

Declaration of Competing Interest

None.

References

[1] Uppu S, Krishna A, Gopalan R. A review on methods for detecting SNP interactions in high-dimensional genomic data. IEEE/ACM Trans Comput Biol Bioinf 2018;15(2).

[2] Koo Ch, Liew M, Saberi M, Salleh A. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. Biomed Res Int 2013. 432375.

[3] Lee K, Yoo J, Kim S, Lee J, Hong J. Autonomic machine learning platform. Int J Inf Manage 2019. In Press, Corrected Proof.

[4] Ge Q, Fukuda D, Han Ke, Song W. Reservoir-based surrogate modeling of dynamic user equilibrium. Transp Res Procedia 2019;38:772–91.

[5] Halgamug M, Davisb D. Lessons learned from the application of machine learning to studies on plant response to radio-frequency. Environ Res 2019;178:108634.

[6] Takagishi Y, Yamanaka T, Yamaue T. Machine learning approaches for designing mesoscale structure of Li-ion battery electrodes. Batteries 2019;5 (3):54.

[7] Huang Y, Cheng Sh, Wu Y, Wu J, Li Y, Huo Z, et al. Developing surrogate indicators for predicting suppression of halophenols formation potential and abatement of estrogenic activity during ozonation of water and wastewater. Water Res 2019;161:152–60.

[8] Costabal F, Matsuno K, Yao J, Perdikaris P, Kuhl E. Machine learning in drug development: characterizing the effect of 30 drugs on the QT interval using Gaussian process regression, sensitivity analysis, and uncertainty quantification. Comput Methods Appl Mech Eng 2019;348:313–33.

[9] Pan Ch, Liu J, Tang J, Chen X, Chen F, Wu Y, et al. A machine learning-based prediction model of H3K27M mutations in brainstem gliomas using conventional MRI and clinical features. Radiother Oncol 2019;130:172–9.

[10] Davatzikos Ch, Sotiras A, Fan Y, Habes M, Erus G, Rathore S, Bakas S, Chitalia R, Gastounioti A, Kontos D. Precision diagnostics based on machine learning-derived imaging signatures. Magn Resonance Imaging 2019. In Press, Corrected Proof.

[11] Xu J, Yang P, Xue Sh, Sharma B, Martin M, Wang F, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum Genet 2019;138(2):109–24.

[12] Musheer1 R, Verma C, Srivastava N. Novel machine learning approach for classification of high dimensional microarray data, Soft Computing, Methodologies and Applications, 2019.

[13] Ghosh M, Begum Sh, Sarkar R, Chakraborty D, Maulik U. Recursive Memetic Algorithm for gene selection in microarray data. Expert Syst Appl 2019;116:172–85.

[14] Damgacioglu H, Celik E, Celik N. Estimating gene expression from high-dimensional DNA methylation levels in cancer data: a bimodal unsupervised dimension reduction algorithm. Comput Ind Eng 2019;130:348–57.

[15] Zhu Z, Ong Y, Dash M. Markov blanket-embedded genetic algorithm for gene selection. Pattern Recogn 2007;40:3236–48.

[16] Shukla A, Singh P, Vardhan M. A new hybrid wrapper TLBO and SA with SVM approach for gene expression data. Inf Sci 2019;503:238–54.

[17] Sapin E, Keedwell E, Frayling T. An ant colony optimization and tabu list approach to the detection of gene-gene interactions in genome-wide association studies. IEEE Comput Intell Mag 2015;10(4):54–65.

[18] Wang Y, Liu X, Robbins K, Rekaya R. AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. BMC Research Notes 2010;3:117.

[19] Quéau B, Shafiq O, Alhajj R. Analyzing Alzheimer's disease Gene Expression Dataset using Clustering and Association Rule Mining. Proceedings of IEEE 15th international conference on information reuse and integration, 2014.

[20] Agapito G, Guzzi P, Cannataro M. Parallel extraction of association rules from genomics data. Appl Math Comput 2019;350:434–46.

[21] Singh Sh, Yang Y, Poczos B, Ma J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quantitative Biol 2019;7:122–37.

[22] Uppu S, Krishna A, An intensive search for higher-order gene-gene interactions by improving deep learning model. In: Proceedings of IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 2018.

[23] Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed, 2012.

[24] Saethang Th, Prom-on S, Meechai A, Chan J. Sample filtering relief algorithm: robust algorithm for feature selection. Proceedings of international conference on neural information processing, 2008.

[25] Mallik S, Mukhopadhyay A, Maulik U, Bandyopadhyay S. Integrated analysis of gene expression and genome-wide DNA methylation for tumor prediction: an association rule mining-based approach. Proc IEEE Comput Intelligence Bioinfor Comput Biol 2013.

[26] Al-Turaiki I, Badr Gh, Mathkour H. Trie-based Apriori motif discovery approach, international symposium on bioinformatics research and applications, 7292, 1–12, 2012.

[27] Tan P, Steinbach M, Kumar V. Introduction to Data Mining, Pearson Addison-Wesley, 1st ed., 2006.