Available online at www.sciencedirect.com

ScienceDirect

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

Original Research Article

A hybrid gene selection method for microarray recognition

Alok Kumar Shukla *, *Pradeep Singh*, *Manu Vardhan*

Department of Computer Science & Engineering, NIT, Raipur, India

ARTICLE INFO

Article history:
Received 27 April 2018
Received in revised form
29 July 2018
Accepted 14 August 2018
Available online 5 September 2018

Keywords:
Accuracy
Ensemble
Adaptive genetic algorithm
Gene selection
Support vector machine

ABSTRACT

DNA microarray data is expected to be a great help in the development of efficient diagnosis and tumor classification. However, due to the small number of instances compared to a large number of genes, many of the computational learning methods encounter difficulties to select the low subgroups. In order to select significant genes from the high dimensional data for tumor classification, nowadays, several researchers are exploring microarray data using various gene selection methods. However, there is no agreement between existing gene selection techniques that produce the relevant gene subsets by which it improves the classification accuracy. This motivates us to invent a new hybrid gene selection method which helps to eliminate the misleading genes and classify a disease correctly in less computational time. The proposed method composes of two-stage, in the first stage, EGS method using multi-layer approach and f-score approach is applied to filter the noisy and redundant genes from the dataset. In the second stage, adaptive genetic algorithm (AGA) work as a wrapper to identify significant genes subsets from the reduced datasets produced by EGS that can contribute to detect cancer or tumor. AGA algorithm uses the support vector machine (SVM) and Naïve Bayes (NB) classifier as a fitness function to select the highly discriminating genes and to maximize the classification accuracy. The experimental results show that the proposed framework provides additional support to a significant reduction of cardinality and outperforms the state-of-art gene selection methods regarding accuracy and an optimal number of genes.

1. Introduction

Biological data, like microarray, contain many irrelevant and redundant genes. Nowadays, DNA microarray has gained considerable attention due to its ability to measure the expression levels of hundreds or thousands of genes in a single experiment [1]. Finally, it produces gene expression data that contain valuable statistics on genomics, and prognosis for researchers [2]. Therefore, there is a need to identify the essential genes that contribute to predict the state of cancers [3]. At the abstract level, the main problem of this

* *Corresponding author at*: Department of Computer Science & Engineering, NIT, Raipur, Chhattisgarh 492010, India.
E-mail address: akshukla.phd2015.cs@nitrr.ac.in (A.K. Shukla).

https://doi.org/10.1016/j.bbe.2018.08.004

experiment is thousands number of genes compared to the small number of instances, misleading genes, and noisy data [4]. To resolve this limitation, researchers have been used prominent gene selection method to select a subset of relevant genes that maximize the capability of the classifier to classify instances more accurately. The instances can be unlike from many points of view like genotype, phenotype or other relevant biological or clinical record, respectively. In the field of bioinformatics, feature selection is also known as gene selection [5].

The gene selection (GS) method tries to select the critical genes among all the features available in the data, which are useful for the application of learning algorithms. Furthermore, it is an essential application of data reduction to avoid challenges, such as over-processing, the high cost of the calculation, and the low interpretability of the final model [6]. Furthermore, the main problem in high-dimensional data sets is the "*curse of dimensionality*." To solve this problem, researchers have introduced a large number of gene selection methods, many of them derived from the need to analyze microarray data to select the best discriminating gene, called as "*biomarker*" [7]. In previous studies, gene selection methods are widely used before data classification in bioinformatics domains [8]. The gene selection approaches are efficient and straightforward, and has several merits:

1. It can retain or enhance the classification performance.
2. It can diminish the data dimensionality.
3. It can discard meaningless and irrelevant genes.
4. It can reduce the computational complexity while performing experiments.

Generally, gene selection techniques are categorized into three important phases: filter [9], wrapper [10], and hybrid [11]. In the first phase, the filter approach is the first technique to select subsets of genes before applying the inductive algorithm. On the other hand, the wrapper method [12] uses inductive learning as a fitness function and looks for the optimal subset of genes in the space of all characteristics. The irrelevant and noisy genes are produced by the high computational effort and also decrease the classification performance. Meanwhile, these genes increase the dimensionality of datasets thereby classifier gives the bias results. Besides filter and wrapper approach, the hybrid method takes the advantages of individual approaches such as filter and wrapper. Over the few decades, several hybrid approaches have been developed primarily, an integration of filter and wrapper method to select the useful genes for accurate diagnosis [13,14]. It takes advantage of both approaches by integrating the complementary strengths [15]. In general, the gene selection method uses a fitness function with a search strategy to attain the subsets of characteristics. The fitness function tries to measure the ability to discriminate a gene and to classify the different labels. In the current literature, there are five measures available, i.e., distance, information or uncertainty, dependence, consistency and correlation [16].

A large number of gene selection methods are available on high dimensional datasets for a tumor or cancer classification, over the past decades [17]. There are, in general, two methods for gene selection is used such as filter and wrapper. The filter methods have been incredibly increased since it can be used to define various relationships between pairs of features to select informative genes from gene classification datasets [18]. For instance, a gene decided in an earlier step cannot be reconsidered in later phases; however, a feature in the gene subset referred as relevant may lose its importance when the gene subset is updated with new feature sets. To address this limitation, researchers have used the powerful nature-inspired optimization technique namely GA [19] which was introduced by John Holland in early 1975, and another algorithm is PSO [20] used as wrapper approach. The improved version of the genetic algorithm is used in this study called adaptive genetic algorithm (AGA). The AGA can improve the growth of the solution quality by adjusting the values of the regulation parameters and controlling the premature convergence, and stagnation.

During the previous years, several machine learning (ML) approaches have been used an ensemble learning model for better prediction. It is a way toward building multiple logical models which convey us to solve the gene selection (or classification) problem with the help of machine learning approaches. Ensemble learning has affected classification problems, on the other hand, it can be used as gene selection for enhancing additional machine learning capacity [21]. The primary cause of inaccuracy in the machine learning is due to noise, bias, and variance. Ensemble strategy helps to minimize these factors. This method is designed to improve the stability and the accuracy of machine learning algorithms. The combinations of multiple gene selection methods may decrease variance, especially in the case of unstable techniques and produce a more reliable representation than an individual gene selection method.

The ensemble can be formed in many ways, but we are using ensemble framework in the form of multi-layer approach which is based on ranking of gene selection method such as "minimum redundancy-maximum relevance" (mRMR) [22], Relief-F [23], chi-square (CS) [24], joint mutual information (JMI) [25], and information gain (IG) [26]. A large number of techniques have widely applied to gene expression datasets for classification as shown in Table 1. Still, there is no agreement on which technology is better than others; particular gene selection method can accomplish better correlated with class than others in respect of specified dataset, while a further approach can outperform the others when dealing with different datasets. This uncertainty about which methods produce an optimal gene subset; is overcome in this research by introducing a new gene selection method that able to exploit the potential of existing gene selection methods.

Availability of statistics in bioinformatics fields is a critical problem toward selecting the relevant genes and overlook the extraneous genes that contribute to a carcinomatous stage. In this paper, we have developed a hybrid model, combination of ensemble gene selection (EGS) and AGA algorithm on biological datasets for identifying the informative features and reduces the computational cost of the learning algorithm. The proposed method uses EGS as a filtering approach to select highest ranked genes that will be passed to AGA algorithm. Furthermore, AGA is combined with SVM and NB to search for the most top-rated genes obtained from EGS genes to find the most revealing genes that will satisfy cancer classification accurately. This process continues until a sufficiently reached

Table 1 – Classification performance in gene expression dataset.

Authors	Year	Method	Datasets	Performance
Zawbaa et al. [5]	2018	ALO-GWO	NCI	70%
Mafarja and Mirjalili [2]	2018	WOA	Leukemia	98.20%
Nakariyakul [3]	2018	IGIS	TOX_171	69.97%
Li et al. [37]	2017	GWO-KELM	Breast cancer	94.90%
Lu et al. [38]	2017	MIMAGA	Colon cancer	83.41%
Mediated et al. [46]	2017	SVM-RFE	Breast cancer	78.94%
Armanfard et al. [47]	2017	lLFS	Leukemia	88.50%
Bashir et al. [36]	2016	HM-BagMoov	Breast cancer	73.45%
Naghibi et al. [42]	2015	mRMR + COBRA	NCI	73.67%
Chaurasia and Pal [35]	2014	SMO	Breast cancer	94.20%
Maulik and Chakraborty [41]	2014	FPRS	MLL	95.61%
Liao et al. [49]	2014	LSLS	SRBCT	97.50%
Xue et al. [40]	2013	MOPSO	Lung cancer	53%
Lavanya and Usha Rani [39]	2012	CART	Breast cancer	74.47%
Chandra and Gupta [44]	2011	ERGS	ALL_AML	97.22%
Xu et al. [48]	2010	Correlation coefficient	SRBCT	77.43%
Sun et al. [43]	2010	Local-learning FS	Breast cancer	79%
Ding and Peng [45]	2005	MID	Lymphoma	92%

accuracy is achieved with limited gene subsets. However, central shortcomings of the wrapper method are that its computation requirement is difficult [27], mainly if the original gene set is large.

The difficulties of the existing approach of filter, wrapper, and hybrid methods have been investigated in [28,29], in addition these approaches have shown the advantage of the simplicity of the filter approach for initial gene filtering and then make use of the wrapper approach to optimize classification accuracy from the filter-out genes. However, there is no agreement on which gene selection method produces the significant gene subsets to discriminate the samples or genes. This motivates us to develop a hybrid framework to identify the cancerous or tumor samples. The central contributions of this paper are shortened as follows:

- An efficient rank-based gene selection is an ensemble which is comprised of IG, JMI, mRMR, Relief-F, and CG.
- Proposed ensemble approach uses multi-layer approach and f-score measure for final gene ranking.
- In this study, we have used a new version of a genetic algorithm called an adaptive genetic algorithm.
- This paper mainly focuses on classifying microarray gene data and to find discriminative gene subsets. Additionally, we compare the proposed approach with existing state-of-art methods to prove the superiority of the proposed framework.

In this study, features are exposed into two-stage, which forms the gene subsets with the lowest cardinality, i.e., according to the corresponding rank of the features by using multi-layer and f-score method, the minimum number of features is selected when f-score value of each gene is greater than the mean of all f-score value, otherwise features will be removed from the gene ranking. Wrapper algorithm as AGA is addressed in the last stage for optimal gene subsets selection which is selected one by one using fitness function to discriminate the samples from available gene expression data sets and evaluate the classification performance. Furthermore, the fact that different hybrid can select different genes makes it doubtful whether the genes selected by a particular classifier are accurate biomarkers. In order to resolve the limitation, we used two classifiers as SVM and NB fitness function on six gene expression datasets, including binary-class and multi-class datasets. It is vital for the cancer classification as well as selecting informative genes.

The remainder of the composition is structured as follows. Sections 2 and 3 provide the related work on existing gene selection methods. Section 4 introduces the proposed work to take the significant features from the gene datasets and distinguish the samples. Sections 5 and 6 discuss the different classification techniques and show the experimental result on the data sets. Section 7 presents the conclusion.

2. Related works

In the recent studies, there are still some shortcomings available on filter and wrapper approach to identify the different type of a tumor or cancer-based decision system. For example, (i) interrelationship among diagnosis factors have not considered, (ii) dynamic alters in sickness direction after some time have been disregarded while effective diagnosis technique could cause better control of the circumstance and make better choices over the long run. The earlier decision system has shown difficulties to analyze the full gene space attempt at a time in the clinical laboratories [30]. In machine learning fields, most of the computational methods mention a pair of critical features for tumor classification [31]. At that opinion, these recommended essential features may enable the researcher to better with a more pre-learning about the reason for disease and give the quickest solution to recover the infected patients as right on time as can be expected under the circumstances. The main aim of the gene selection method is to select a meaningful gene subset from the biological gene datasets based on some criteria, i.e., redundancy, relevancy, and consistency. Our related work focuses on the gene selection and highlights the limitations of previously studied

methods with the aim of selecting a significant subset of features that maximize the performance of a given classifier. For example, Al-Rajab et al. [32] have investigated the embedded particle swarm optimization (PSO) and SVM with gene subset, consisting of the most relevant genes from the microarray datasets. An extension of this work is introduced in [5], used the GA, tabu search (TS), and SVM for optimal gene subset, consisting of the most important genes. Furthermore, several gene selection algorithms are also available for cancer classification, in the literature [33,34]. Zhang et al. [35] have applied the AGA algorithm to optimize the lowest regulator that requires the satisfaction of various static and dynamic operational requirements. The results showed that AGA has significantly improved the performance of the genetic algorithm.

Similarly, Wei [36] has obtained the Genetic Quantum method and integrated with an enhanced Self-Adaptive behavior, is applied for solving electromagnetic optimization problems. Chen et al. [36] have developed a prognostic algorithm based on support vector regression [37] and AGA. One of the new challenging fields in the GS method is the verification of kinship where Alirezazadeh et al. [37] have studied the actual and discriminatory characteristics using the genetic kinship algorithm and then satisfied kinship verification. Furthermore, the proposed method was tested and analyzed in the KinFace W-I and KinFace W-II standards and large data series, and verification rates of 81.3% and 86.15% were obtained. Different gene selection methods are reported in Table 1, detection of cancer and tumor classification based on gene selection approaches has been investigated by only a few researchers to overcome the limitation.

The various literatures are available in the field for gene selection as seen from Table mentioned above. The new technique for cancer classification has introduced by Pes et al. [38], which is related to ensemble-based gene selection and assess the effects of ensemble method. In the context of ensemble method, Emmanuella et al. [39] have presented a new methodology by using three well-known evolutionary techniques such as particle swarm Optimization, Ant Optimization, and genetic algorithm to select the subsets of genes for the individual components of ensembles. Similarly, Ebrahimpour et al. [40] have investigated the new ensemble method based on mRMR and Hesitant Fuzzy Set method that elect genes by using ensemble algorithm and similarity measures. On the other hand, Moradi et al. [41] have proposed HPSO-LS method to integrate the local search strategy and particle swarm Optimization to select the less correlated and useful gene subsets. As stated above related work, the variety of research has done with gene datasets. From Srinivas and Patnaik [42] point of view, the adaptive genetic algorithm has to explore the massive search space in the benchmark gene datasets for getting the best quality of solutions.

3. Existing gene selection method

The critical point of gene selection method is to choose m relevant genes from the d original dataset where $m < d$ on some criteria as correlation, redundancy, similarity, and inconsistency to identify the early stage tumor detection and cancer discovery. In this article, the problem is solved by a two-stage gene selection approach for choosing significant genes from the original dataset to correctly classify the samples. The overall process of the proposed method is shown in Fig. 2. The gene selection process for classification is applied to the following six gene datasets: leukemia, colon cancer, diffuse-large B-cell lymphoma (DLBCL), breast cancer, SRBCT, and lung cancer.

Let £ = {£1, £2, . . ., £n} is a set of genes and S = {s_1, s_2, . . ., s_m} a set of instances. The vector representation of the DNA microarray gene dataset is expressed as $Z = \{Z_{i,j} | 1 \le i \le m, 1 \le j \le n\}$, where m indicates the number of samples; and n number of genes. Before transferring data value in methods, we standardized the gene expression level of each dataset in interval 0 to 1 by using the following formulation:

$$Z' = \frac{Z_{i,j} - \min(Z_{i,j})}{\max(Z_{i,j}) - (Z_{i,j})}$$

where min and max correspond to the minimum and maximum gene value for gene £ in all samples. In the proposed method, the first stage uses the EGS method, and the second stage uses the wrapper approach as adaptive genetic algorithm search strategy. The goal of the adaptive genetic algorithm is to investigate the gene subsets as solutions to reduce a large number of genes to be later classified.

3.1. *Gene selection*

Gene selection is a useful tool in the fields of machine learning and bioinformatics, by which reduction is done on available high dimensional datasets. Thanks to gene selection method which decreases the computational cost and also increases classification performance when growing datasets explosively regarding samples and attributes. As reported by Ang et al. [1], have presented the different categories of gene selection approaches, namely filter, wrapper, and hybrid methods for cancer classification. They have observed that improved results can be obtained by filter-based gene selection which chooses the informative features from the gene dataset for the better diagnosis. This paper briefly described the most popular gene selection techniques, i.e., Relief-F, mRMR, Information gain, joint mutual information, and chi-square with the aim of selecting an informative subset of features that maximize the prediction performance.

3.1.1. *Filter based gene selection*

3.1.1.1. Relief-F. In the existing literature, one of the dominant gene selection approach as Relief-F is used for getting a suitable final ranking of genes. It is an improved version of Relief [43] and also helps us to select the accurate gene sets which can better to predict a target label. It can distinguish conditional dependencies between attributes which present in candidate dataset and provide a unified view of the attribute assessment in classification. It also examines the optimal gene subsets based on evaluation function that is highly correlated with the sample class and uncorrelated with other class. The Relief-F methods are striking because they may be successfully applied in all situations. However, Relief does not explicitly reduce the relevancy of selected genes.

3.1.1.2. *Joint mutual information*. Joint mutual information (JMI) is a simple gene selection algorithm that works on the concept of information theory to evaluate the worth of features which is present in the original dataset. In addition, it helps us to find the stingy subsets which has relevant characteristics and which are highly correlated with the class and uncorrelated with each other. The outcome of optimal gene subsets can ignore some relevant features because they have a smaller correlation with the class. According to mutual information (MI) base, here objective of JMI is to choose a subsets S with k features $X = \{x_1, x_2, \ldots, x_k\}$. To calculate both relevancy and redundancy between two variables such as X and c, MI have applied to assess the mutual dependency between them. Mutual information is updated according to Eq. (1).

$$I(y; X) = H(y) - H\left(\frac{y}{X}\right) \tag{1}$$

where $H(y)$ and $H\left(\frac{y}{X}\right)$ represent as entropy and conditional entropy (CE) between the class and variable. Here, we use joint mutual information (JMI) to shrink the redundancy between data attributes X and class y. The relevance of inputs attributes defined by the JMI as shown in Eq. (2):

$$M_{JMI}(X) = \sum_{x_j \in S} I(x_k; x_j; y) \propto \sum_{x_j \in S} I\left(y; \frac{x_k}{x_j}\right) \tag{2}$$

where $I(x_k; x_j; y)$ represents the MI between original features set x_k and selected features x_j with respect to target class y.

3.1.1.3. *Minimal-redundancy-maximal-relevance (mRMR)*. The pre-processing task is very crucial in gene selection to identify the important or relevant genes from the high dimensional data. In this paper, most popular filter method as mRMR is used, and observed that the minimization of max-dependency and maximal relevance on the gene sets is hard to understand. To handle these kind of problem, more efficient method as ''minimal-redundancy-maximal-relevance'' (mRMR) to gene selection is introduced which employed as filtering approach to finding initial solutions with high ranked genes for gene selection problems [44]. It also tries to find the most relevant features based on its correlation with the class label and to minimize redundancy of the features themselves [45]. To quantify both relevancy and redundancy, mutual information (MI) is applied to estimate the mutual dependency of two variables such as X and y as seen in Eq. (1). By using mutual information (MI) approach, researcher have designed a gene selection with the aim to choose a subsets S with N genes with maximum dependency on the target class c; so-called max dependency, is formulated as Eq. (3).

$$\max\ w(X, y) = I(y; x_1, x_2, \ldots, x_N) = H(y) - H\left(\frac{y}{x_1, x_2, \ldots, x_N}\right) \tag{3}$$

As shown in Eq. (3), the dependency among features X is estimated it can be large value. The relationship between redundancies between features is expressed in Eqs. (4) and (5).

$$\min\ Z(X, c) = 1/s^2 \sum_{x_j \in s} I(x_j; x_k) \tag{4}$$

$$\text{Max}\ \varnothing(w, Z) = w - Z \tag{5}$$

The integration of Eqs. (4) and (5) is known as ''minimal-redundancy-maximal-relevance'' (mRMR) which describes in Eq. (6):

$$j_{mRMR}(\varnothing) = I(c; X) - 1/s^2 \sum_{x_j \in s} I(x_j; x_k) \tag{6}$$

where x_j is selected subset of gene S and x_k is original genes set.

3.1.1.4. *Information gain*. In gene selection process, the most critical evaluation method is used known as information gain which is applied for finding the optimal informative genes rendering to the information gain values, in consideration of a single gene at a time. As shown in Eqs. (7)–(9), measures the amount of each gene by the information gain concerning the class.

$$IG = H(X) - H\left(\frac{Y}{X}\right) \tag{7}$$

$$H(X) = -\sum_{X}^{x} P(x) \log(P(x)) \tag{8}$$

$$H\left(\frac{Y}{X}\right) = -\sum_{X}^{x} P(x) \sum_{Y}^{y} P\left(\frac{y}{x}\right) \log P\left(\frac{y}{x}\right) \tag{9}$$

where X and Y feature; $P(x)$ and $P\left(\frac{y}{x}\right)$ are the probabilities distribution of x and y.

3.1.1.5. *Chi-square* (χ^2). Chi-square (χ^2) is computing the value of each gene by the statistic test concerning the class and shows a better performance in the gene selection. The mathematical formulation of χ^2 represents in Eq. (10) for calculating the value for two adjacent intervals.

$$\text{Chi-square}\ (\chi^2) = \sum_{2}^{j=1} \sum_{l}^{k=1} \frac{(\alpha_{j,k} - \beta_{j,k})^2}{\beta_{j,k}} \tag{10}$$

where l is the number of classes, $\alpha_{j,k}$ is the number of instances in the jth interval with class k, R_j is the number of instances in the jth interval, l_k is the number of instances of class k in the two intervals, N is the total number of instances in the two intervals, and $\beta_{j,k}$ is the expected frequency of $\alpha_{j,k} = R_j * l_k/N$.

3.1.2. *Wrapper-based gene selection*

3.1.2.1. *Adaptive genetic algorithm*. Firstly, the genetic algorithm (GA) [46] is introduced by John Holland in 1975, which is inspired by the natural selection process and worked on parallel search heuristic. GA solves the optimization problem based on the process of natural genetic schemes. A genetic algorithm has two primary operations, namely crossover, mutation, and two tuning factors: (a) *Pc* crossover probability and (b) *Pm* mutation probability.

In the GA algorithm, *Pc* and *Pm* are static variables that are corrected in the GA search process. When the *Pc* is too big, the global search is too difficult, and the optimal solution can be lost. When *Pc* is too small, the search may stop at the local minimum. At this stage, we discuss an alternative approach to a suitable assessment for gene selection, based on NB and SVM classifier, as defined in [44]. When the value of *Pm* is too big, GA is similar to random search algorithms; and when *Pm* is smaller, the search scan ability is crushed.

To find the most suitable value for *Pc* and *Pm*, a multiple cross-over may be required. A more reasonable approach is to allow GA to regulate the *Pc* and the *Pm* during the searching process, which is termed adaptive genetic algorithm (AGA). In AGA, the values of *Pc* and *Pm* can be modified according to Eqs. (11) and (12):

$$Pc = \begin{cases} k_1 \cdot \left[\dfrac{(f_{\max} - f')}{(f_{\max} - f_{avg})}\right] - k_2, & f' \geq f_{avg} \\ k_3, & f' < f_{avg} \end{cases} \tag{11}$$

$$Pm = \begin{cases} k_4 \cdot \left[\dfrac{(f_{\max} - f)}{(f_{\max} - f_{avg})}\right] - k_5, & f \geq f_{avg} \\ k_6, & f < f_{avg} \end{cases} \tag{12}$$

where $f_{\max}$ represents the maximum of all chromosome fitness when AGA do a search operation, f_{avg} represents the average fitness, f' represents the maximum fitness of the parents in cross-over and k_1, k_3, k_4, k_6 signify the four control variables ranged from [0,1], k_2 and k_5 are the constant. The overall AGA optimization process is revealed in Fig. 1.

The AGA may improve the convergence of the solution by adjusting the values of tuning parameters. The adaptability of AGA makes it more robust and therefore enhances the likelihood to find the optimal global solution. Due to fast convergence, AGA is applied to examine the relevant genes from high dimensional datasets. Additionally, several AGA approaches work as gene selection which is getting a good solution in machine learning algorithms because of the low computational cost [47].

4. NB and SVM as fitness function

Most of the wrapper approaches [48] have been applied in microarray datasets to reduce the computational burden using a learning algorithm and to evaluate the performance regarding classification accuracy. The proposed method is not combined with specific classifiers. We used two different classifiers namely NB and SVM classifier; due to its simplicity and lower complexity, and in the SVM classifier; due to its tremendous performance and robustness to microarray data. These are cited in the top ten data mining algorithms [49,50].

4.1. *Naïve Bayes (NB)*

The concept of NB theorem is the extension of Bayes theorem with the hypothesis of independence of all features. It is also used in classification problem in machine learning. Let's select the frequent class of instance C from datasets and assume the random gene vector $X = (X_1, X_2, \ldots, X_l)$ by using the observed features. Let c_j represent jth class label and $x = (x_1, x_2, \ldots, x_l)$ represent a predicted gene vector. To identify the correct class label of a testing instance x, so we can use Bayes' theorem to calculate explicit probabilities as express in Eq. (13):

$$Pr(C = c_j | X = x_{1,\ldots,i}) \alpha Pr(C = c_j) \prod_{m}^{i=1} Pr(X_i = x_i | C = c_j) \tag{13}$$

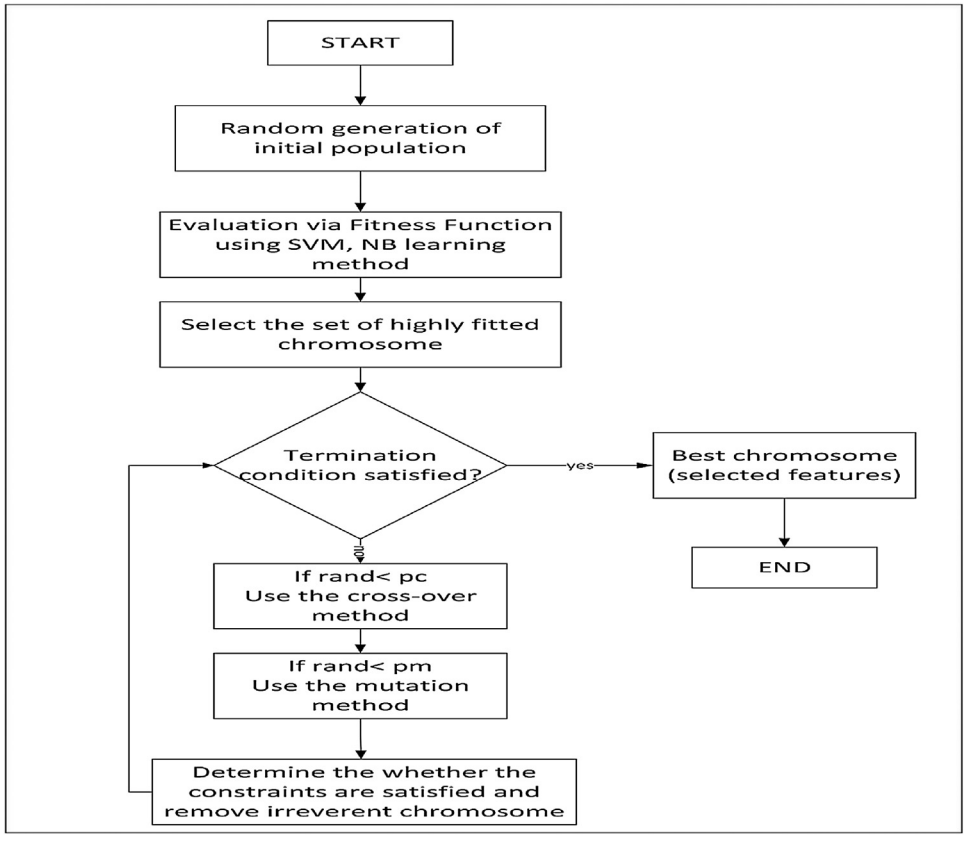

Fig. 1 – Overall procedure of AGA.

4.2. *Support vector machine (SVM)*

In existing classification algorithms, separation of the gene vector and predicting the correct label is the primary challenging task. Support vector machine classifier is a controlled power to discriminate the classes through hyper-plane and overcome the recent challenges. It has shown exceptional performance in a variety of biological classification problems [51]. In the proposed approach, it is employed as a classifier to assess the fitness of gene subsets.

Given an instances m, the data set is supposed to be a finite set of m sample or label pairs defined as follows:

$$S = \{x_i, y_i | (x_i, y_i) \in \Phi^n$$

where $x_i \in \Phi^n$ and $y_i \in \{\pm 1\}$ represents the instances and labels of datasets. And hyper-plane is described as:

$$S = \{x_i, y_i | (x_i, y_i) \in \Phi^n$$

$$f(x) = \sum_{m}^{i} \zeta_i y_i k(x_i, x) + b$$

where k (x_i, x) represent a kernel function and sign of $f(x)$ describe that which class is belong to, with non-zero support vectors ζ_i and a correct bias b.

5. Proposed methodology

Before attempting to investigate new hybrid method, it can be sensible to find the limitations of previously developed hybrid GA method [52]. According to [52], this method requires more computational cost, due to select a near-optimal subset of genes from the high dimensional datasets. To address the limitations, our proposed hybrid method is used ensemble gene selection method and wrapper a method to choose a significant subset of genes and also tackle the issues above-mentioned. Additionally, it enhances the classification accuracy and reduces the search complexity for generating gene subset over the high dimensional datasets.

Pseudo-code 1: Heterogeneous ensemble method.

```
Data: M – the number of ranked methods
Outcome: P – Best ranking gene set
For each n from 1 to M do
Obtaining ranking Am using gene selection method M
End
A = combining ranking Am with a ranking combination method
Obtain top-ranked features
```

5.1. *Ensemble gene selection method*

In bioinformatics domains, each gene selection method has its qualities, shortcomings, and performance dependencies on the structure of datasets. However, in spite of the accessibility of recent measures of gene selection, researchers have found that no perfect method exists. The main drawback of single gene selection method is leave-out the significant genes which do not contribute in the classification process. To address the disadvantage of unique method, in this paper a hybrid method is proposed which is a combination of the two-stage processes such as filter and wrapper approach. The overall procedure of proposed method is shown in Fig. 2.

The primary aim of this integration is to discriminate the irrelevant genes from the original datasets thereby improved the classification accuracy. The first stage involves the ensemble learning with the multi-layer approach and f-score approach, and the second stage is a wrapper as AGA for early identification of the patient's disease. The EGS approach depends on two-layer combining rankings of features that contain all well-ordered elements. The results of the base selector are utilizing conglomerate, and optimal subsets of genes are selected when the f-score value of each gene is greater than the mean of all f-score value. Otherwise, features will remove from the gene ranking. By integrating the key points of ensemble gene selection method, improve the classification performance of the gene datasets.

5.1.1. *Multi-layer ensemble approach*

The ensemble model decreases variance, bias, or strengthen classification performance with the help of important features which is selected by multiple gene selection methods. The recent studies confirmed that the intensity of the ensemble gene selection method is correlated to base gene selector and also the lack of correlation between features. As compared to individual methods, it is less expensive regarding computational complexity when we integrate the two-layer approach for feature selection. In this scenario, where one gene selection method has some restriction, the other method performs well, consequently giving better performance. This uncertainty can be resolved by a proposed method which is dependent on two-layer ensemble gene selection method.

In this first stage, we propose the ensemble learning strategy of gene selection based on a two-layer approach which depends on five gene selection methods is described in Pseudo-code 1. This approach assesses the potential and shortcomings of the individual strategies. The several gene selection methods are trained using the gene datasets, and the output is then combined using a combination method.

5.1.1.1. Layer-1 for gene selection. The diversity of the gene selection method is accomplished by selecting the entirely different type of features. The mRMR is a mutual information-based method which selects the highly correlated and less irrelevant features [53]. Moreover, the mRMR method decreases the computational complexity and increase the average classification accuracy. Similarly, the joint mutual information (JMI) method also works on correlation-based to evaluate the worth of features by using mutual information function. The stingy subsets which have relevant features and they are highly correlated with the class and uncorrelated with each other. In these subsets, we can ignore some relevant features because they have the small correlation with the class. On the other hand, Relief-F is similarity-based gene selection method which has the merits of robustness, interpretability, scientific acceptance, and wide-spread accessibility [54]. Thus, all of the three-gene selected method complement each other very well. Furthermore, gene ranking outputs will be obtained from three gene selection methods in the form of gene ranking using combinator method.

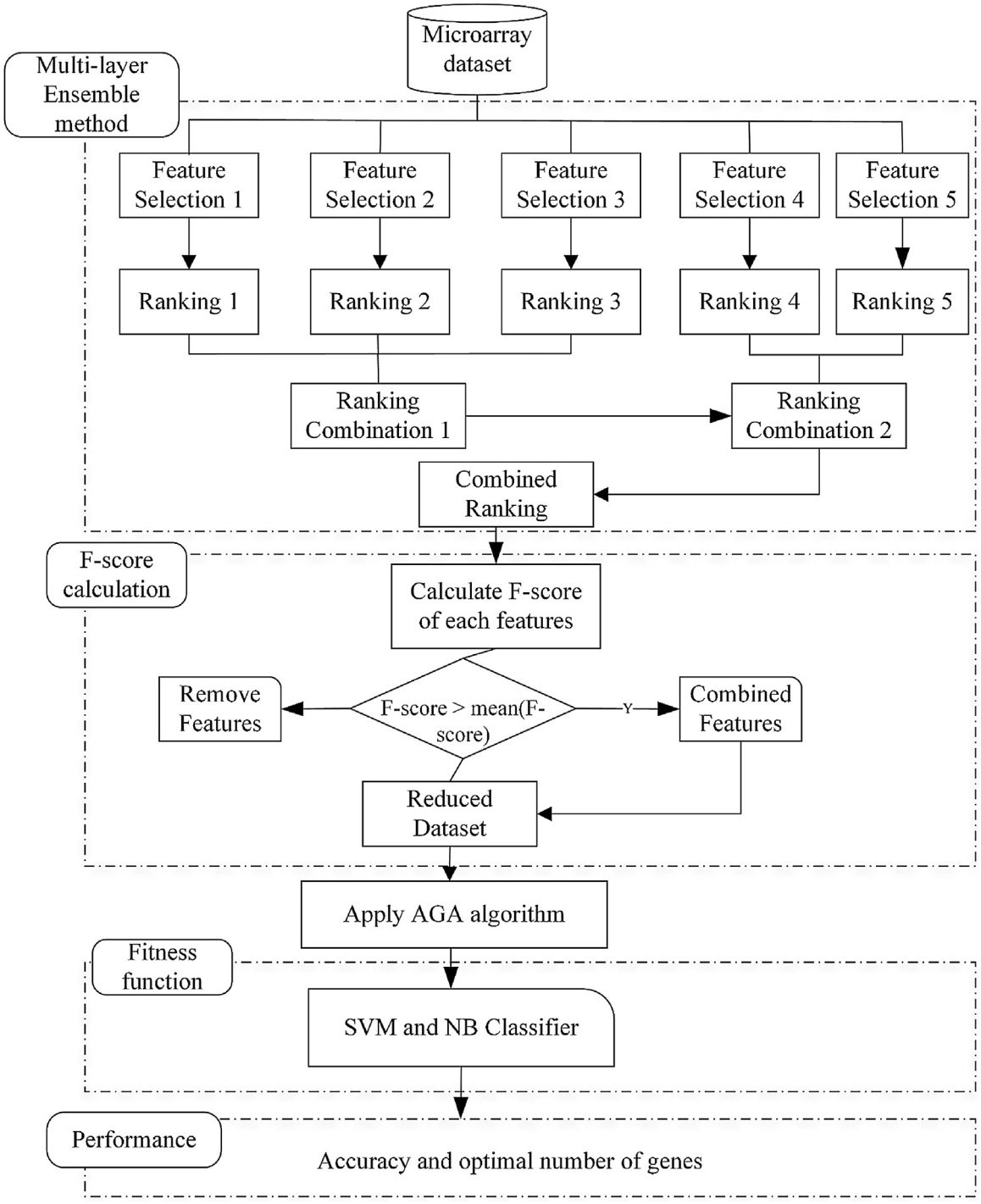

Fig. 2 – Overall process of proposed gene selection method.

5.1.1.2. *Layer-2 for gene selection*. In order to achieve diversity among the gene selection methods in layer-2 approach, we have used chi-square and information gain method. In layer-2, the output of layer-1 is integrated with further two different gene selection methods with the help of combination method known as a conglomerate. In layer-1 no statistical-based method used for gene ranking, therefore, we used chi-square method in layer-2 for significant features. Information gain (IG), also referred to as Mutual Information, helps us to measure the dependence between the two attributes. Intuitively, mutual information (MI) measures the information between X (attributes) and y (object class) share: it measures how much knowing one of these attributes reduces uncertainty about the other [55].

5.2. *Ranking combination*

Ensemble method is a fruitful strategy to improve the machine learning results. We utilize the ensemble method (see Pseudo-code 1) that accomplishes a final ranking of

features by combining outputs (rankings) from individual rankers using a combination method known as a conglomerate to make a remarkable final production. In the past literature, a few diverse combination methods have been presented, such as mean, minimum, maximum, etc. But, here we use the more sophisticated measures namely ranking vote.

5.3. *Ranking vote*

Ranking vote is a group decision-making framework and has found as useful as other more composite schemes [56]. For an input sample, the individual gene selection method generates an independent conclusion regarding the individuality of the sample. Then, the identity is assigned according to which the frequency of voters agrees (the output prediction is the one that receives more than one vote). In the context of gene selection, the input samples are the features ranking and the identity of each features ranking is an ensemble or not in combination. If none of the rank numbers get more than one of the votes, we may say that the ensemble method cannot make a stable rank for this instance then we pick-up the smallest rank value. Although this is a widely used technique, you may try the most voted rank as the final ranking according to Eq. (14).

$$\sum_{M}^{n=1} d_{n,j} = \text{argmax}_{j \in \{1,2,\ldots,L\}} \sum_{M}^{n=1} d_{n,j} \tag{14}$$

where M represents the number of gene selection method, and L has selected some attributes. For attribute j, the sum $\sum_{M}^{n=1} d_{n,j}$ tabulates the number of votes for j. Plurality chooses the attribute j which maximizes the sum.

5.4. *F-score method*

The F-score strategy is an essential and simple method which measures the recognizing a pair of classes with real values. The proposed method utilizes the F-score [57] strategy as a threshold which decides to select the most critical and essential features of gene datasets. In the F-score technique, estimations of each attribute in typically the datasets are computed merely as Eq. (15) and after this with a specific objective to discover the features from the complete datasets, the threshold rate (δ) is attained by controlling the mean value of F-score of entire features space. If the F-score value of any gene is lesser than the threshold value, then gene discards from gene ranking. Otherwise, gene is added to the gene ranking. For an input sample, individual attributes produce a unique value regarding the identity of each sample.

$$F(i) = \frac{\left(\overline{X}_i^+ - \overline{X}_i\right)^2 + \left(\overline{X}_i^- - \overline{X}_i\right)^2}{\frac{1}{(n_+ - 1)} \sum_{k=1}^{n_+} \left(X_{k,i}^+ - \overline{X}_i^+\right)^2 + \frac{1}{(n_- - 1)} \sum_{k=1}^{n_+} \left(X_{k,i}^- - \overline{X}_i^-\right)^2} \tag{15}$$

where $\overline{X}_i, \overline{X}_i^+$, and $\overline{X}_i^-$ represent the average of the ith gene of the full attribute, +ive (positive), and −ive (negative) data sets, respectively; $X_{k,i}^-$ is the ith gene of the kth −ive sample, and $X_{k,i}^+$ is the ith gene of the kth +ive sample. As shown in Eq. (15), the numerator depicts the inequality between the +ive and −ive sets, and the denominator defines the one within each of the two sets. The value of f-score is larger, it means this gene is more discriminative.

Pseudo-code 2: Proposed algorithm.

Filter method

Input: D - Xm × Fn
ε – fitness function as SVM and NB
δ – threshold of the number of features to be selected
D' – reduced dataset
M – the number of ranked methods
Output: Accuracy and the optimal number of genes
Initialization: $F' = \varnothing$ and $At = \{\}$

Begin
  for each $f_m \in M$
    Evaluate At for f_m according **to Pseudo-Code 1** //filter gene ranking method
  end for
  for each i = 1 to M − 2
    ens1 = ensemble (At(i), ***vote***) //ensemble of three gene ranking (Layer-1)
  end For

  for each i = M − 2 to M
  At = ensemble (At(i), ***vote***) //ensemble of two gene ranking
    ens2 = At*Uens*1 //ensemble output of first layer and output of two method (Layer-2)
  end for
    select $(f_n) \leftarrow f$ (ens2, δ) //select the optimal gene subsets by f-score value
    select the gene F′ (depend on f-score value) from f_n; $F' = F - f_n$
return D' (Xm × F′)

Wrapper Method AGA: Op = AGA (acc, opt)
Acc = maximum classification accuracy
Opt= optimal number of genes
t = 0; Initialize random chromosome Pi; |P|= maximum population size; Tmax = maximum iteration
for i = 1 to |P|
  Initialize Pi
end for
F(Pi)= Evaluate (Pi, ε) //Estimate the fitness for each chromosome using SVM and NB classifier
while (t <Tmax) **do**
  t = t + 1
    while (size of (p) does not meet |P|) **do**
      select p1 and p2 from P //Select the two chromosomes using tournament method
      [c1, c2] = crossover (p1, p2) using Eq. (11) //Apply adaptive crossover operator
      [m1, m2] = mutation (c1, c2) using Eq. (12) //Apply adaptive mutation operator
      L= combined (p, c1, c2, m1, m2) //combination operator
      F(L)= Evaluate (L, ε) //Evaluate fitness value using SVM and NB classifier
      P= Elitism (L) //Replace old chromosome to new chromosome according to high fitness value
    end while
end while
return Op //maximum accuracy and optimal number of genes
end

Above mentioned Pseudo-code 2 structure, the proposed methodology is used for better diagnosis of diseases with two stages process. The building of the new diagnostic system, much of the work is concentrated on a relevant gene subset

which gives excellent performance regarding accuracy and some optimal genes.

5.5. *Wrapper approach*

An adaptive genetic algorithm has widely applied to optimization problems such as gene selection and classification. Today, it is often used as a wrapper to refine gene selection of the fusion of the most informative genes in cancer prediction [58] and to discriminate a type of a tumor in the case of a new patient which helps to discover the early diagnosis. Finally, SVM and NB classifier are used for fitness function in the adaptive genetic algorithm (AGA) and find into the search space for obtaining best chromosome with small gene subsets. It also reduces the computational complexity needed to extend a subset of essential features. The overall process of a genetic algorithm for gene selection is described as follows:

- Generate the random population for AGA. The population size is dependent on defines space. The larger the size is, the more comfortable the AGA searches for the optimal solution and longer time will elapse. In this work, the population size p is set to 30. Each population consists a number of attributes selected in EGS.
- Adapt binary encoding scheme in chromosomes. After encoding, each chromosomes length should be some gene present in the reduced dataset.
- Calculate the fitness value concerning SVM and NB classification accuracy for each chromosome.
- Estimate all fitness value of $f_{\max}$, f_{avg}, and f'.
- Select a highly fitted chromosomes by selection method.
- Randomly paired the chromosomes in (11), according to the value of Pc in the formulation.
- Using the cross-over method to produce a new population.
- According to the value of *Pm* in the formulation (12), using the mutation method to produce a new population.
- Test, whether the current optimal fitness value meets the target or the termination criteria, are met. If yes, go to next; otherwise, go to previous.
- Output as an optimal subset of genes.

The key point of the new hybrid method is to explore the entire search space and find an optimal gene with maximum classification accuracy on the datasets. The chromosome has *n* important genes selected from the first stage of the proposed method. This chromosome contains the binary bits the value of 1 or 0; if the gene is selected that show the 1 otherwise 0. In this paper, the optimal subset of genes and maximum classification accuracy are selected by adaptive genetic algorithm utilizing SVM and NB fitness function. To avoid the over-fitting issue, the SVM and NB techniques are used tenfold cross-validation error estimation method.

6. Result and discussion

6.1. *Dataset description*

To estimate the effectiveness of the proposed method, we take into consideration different types of selection of characteristics in different sets of reference data such as diffuse large-B-cell lymphoma (DLBCL), small-blue-round-cell-tumor (SBRCT), and lung cancer. They are download from www. http://www.gems-system.org/. The datasets such as breast cancer, colon cancer, leukemia are download from [59]. Each dataset contains a varied set of attributes with the class of disease. A brief description of samples, genes, and classes of used datasets are summarized in Table 2.

Table 2 – Dataset description.

No.	Dataset	Instances	Genes	Classes
1.	Breast cancer	97	24,481	2
2.	Colon cancer	62	2000	2
3.	DLBCL	45	4026	2
4.	Leukemia	72	7129	2
5.	SBRCT	83	2308	4
6.	Lung cancer	203	12,600	5

6.2. *Parameter setting*

For the sake of clarity, Table 3 shows the suitable parameters of the applied algorithms. In this experimental study, we define value as threshold 100 to select the significant subset of genes by the EGS method for getting maximum performance.

6.3. *Performance measures*

We measure the classification performance with the help of two classifiers such as SVM and NB, which works as a fitness function in the AGA algorithm. The performance is evaluated by four parameters, i.e., accuracy, sensitivity, precision, and F-measure on six gene datasets. These performance measures are defined as:

Accuracy: To predict the percentage of correctly classified samples, it is formulated as:

$$\text{Accuracy} = \frac{T_N + T_p}{T_p + T_N + F_N + F_p}$$

Sensitivity: Percentage of positive instances that are predicted as positive. It is also called TPR or Recall. It is formulated as:

$$\text{Sensitivity (Sen)} = \text{Recall (Re)} = \frac{T_p}{T_p + F_N}$$

Precision: It is a percentage of positive predictions that are correct. This is also called PPV (positive predicted value). It is formulated as:

$$\text{Precision (Pre)} = \frac{T_p}{T_p + F_p}$$

Table 3 – Parameters setting.

S. no.	Parameter	Value
1.	Population Size	30
2.	Number of generations	100
3.	Chromosome length	100
4.	K2	0.05
5.	K5	0.03

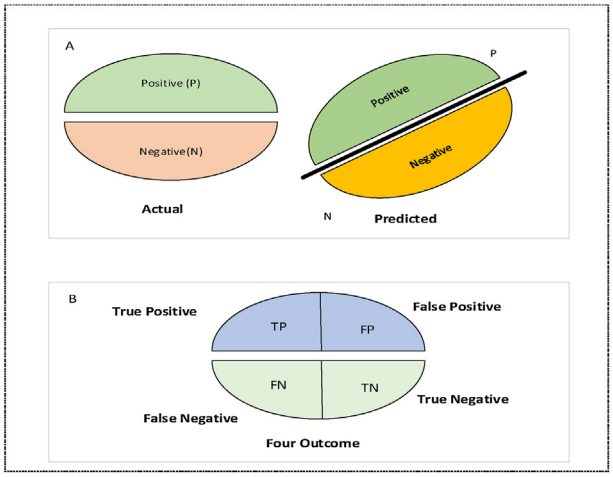

Fig. 3 – Actual and predicted labels generate four outcomes of the confusion matrix.

F-measure: It is a composite measure that favors algorithms with greater sensitivity and challenges those with greater specificity.

$$\text{F-measure (Fmes)} = \frac{2T_p}{2T_p + F_p + F_N}$$

Here T_p, T_N, F_p, and F_N are true positive, true negative, false positive and false negative in the independent datasets, respectively. Based on confusion matrix (see Fig. 3), we evaluated the performance of the proposed method and rival gene selection.

6.4. *Experimental results*

In the literature, several gene selection methods have been applied to the diagnosis and classification of gene datasets. Despite, there is no agreement by which gene selection technique can produce the significant gene subsets for accurate prediction. A particular gene selection strategy may be superior to others for some specific gene datasets, or other gene selection strategy can perform better for some other datasets. This motivates us to invent a hybrid gene selection method which is a help to identify the meaningful genes and classify a disease correctly with less computational time.

In order to show the superiority of proposed genes election method, is applied on gene datasets including as breast cancer, colon cancer, diffuse large-B-cell lymphoma (DLBCL), leukemia, small-blue-round-cell tumor (SBRCT), and lung cancer using the ten-fold and approach. In the first set of experiment, we have computed classification performance by using two classifiers such as SVM and NB on six gene data sets. The experimental results on gene datasets are shown in Table 4 regarding classification performance; accuracy, sensitivity, precision, and f-measure. As we can see, the accuracy of the classification is not very interesting, especially for the lung data set. Furthermore, we have found that the SVM classifier provides better classification rate in all data sets except the lung cancer dataset. More precisely, the performance on two classifiers is compared. This section illustrates the achievements of proposed method against five most common gene selection methods from microarray datasets. To evaluate the classification performance of individual filter method as reported in Table 5. The proposed method achieves the highest accuracy as 94.06% in SRBCT using SVM classifier concerning another existing method. In addition to, the results reported in Table 5, shows that the average accuracy of all filters and proposed method with highest accuracy in all gene datasets as 93.23%, 89.91%, 92.89%, 93.19%, 84.97%, 94.06%.

Table 6 shows that proposed method achieves the highest accuracy as 94.09% in SRBCT data using NB classification method as compared to other existing filter methods. The maximum classification accuracy for the proposed method is reported in Table 6 at bold value. In addition to this, the results reported in Table 6, shows that the average accuracy of the proposed method is the highest accuracy over the other classification method. The maximum accuracies of the existing gene selection range from 94.01%, 90.19%, 91.73%, 93.87%, 88.34%.

Table 4 – Percentage of average performance using SVM and NB classifiers on six gene datasets.

Dataset	NB				SVM			
	Acc	Sen	Pre	Fmes	Acc	Sen	Pre	Fmes
Breast	91.03	86.19	86.65	86.41	91.46	89.57	88.11	89.23
Colon	91.28	92.47	98.97	97.91	92.99	91.58	95.74	94.98
DLBCL	92.29	90.18	92.71	92.84	96.57	96.64	96.02	95.88
SBRCT	89.02	85.54	86.33	85.77	90.07	88.47	87.08	85.34
Lung	94.87	93.91	93.15	93.13	95.22	94.46	95.89	93.24
Leukemia	94.26	87.79	89.32	87.73	94.34	92.47	94.19	93.75

Note: Acc, accuracy; Sen, sensitivity; Pre, precision; Fmes, F-measure.

Table 5 – Classification accuracy (%) with top 25 genes using SVM classifier.

Datasets	SVM					
	CMIM	JMI	mRMR	DISR	Relief-F	Proposed
Breast	80.25	71.36	83.54	77.39	74.15	80.98
Colon	81.39	72.98	80.36	80.96	71.25	82.89
DLBCL	93.23	77.03	90.87	88.31	81.06	91.96
SBRCT	93.03	89.31	92.89	93.19	84.09	**94.06**
Lung	89.01	88.37	90.13	88.94	84.97	90.38
Leukemia	88.06	89.91	92.39	87.94	78.34	91.26

Table 6 – Classification accuracy (%) with top 25 genes using NB classifier.

Datasets	NB					
	CMIM	JMI	mRMR	DISR	Relief-F	Proposed
Breast	78.98	69.98	84.42	78.98	70.98	83.39
Colon	82.96	74.07	82.93	80.58	73.09	83.54
DLBCL	88.03	80.06	91.09	89.39	88.34	93.37
SBRCT	94.01	90.19	91.73	93.87	88.29	**94.09**
Lung	91.03	89.39	90.89	89.67	87.09	93.31
Leukemia	90.36	84.1	90.81	88.31	79.36	93.37

Table 7 – Experimental results for each run using the proposed method with SVM on breast, colon, and DLBCL datasets.

Runs	Breast		Colon		DLBCL	
	Acc	#feat	Acc	#feat	Acc	#feat
1	89.35	15	98.25	15	99.15	15
2	84.03	19	96.27	16	**99.64**	**11**
3	90.14	14	**99.85**	**24**	99.54	19
4	85.95	11	97.11	11	99.05	28
5	**91.47**	**13**	98.03	16	98.15	21
6	90.65	17	98.89	22	99.07	27
7	88.78	16	98.9	19	99.04	14
8	91.34	21	97.84	17	98.26	10
9	86.74	24	98.78	13	99.34	18
10	87.95	22	98.86	9	98.95	21
Avg	88.64	17.2	98.9	16.2	99.01	18.4

Note: Acc, accuracy; #feat, optimal gene subsets.

Table 8 – Experimental results for each run using the proposed method with SVM on SBRCT, lung, and leukemia datasets.

Runs	SBRCT		Lung		Leukemia	
	Acc	#feat	Acc	#feat	Acc	#feat
1	98.54	8	99.32	11	98.35	15
2	99.01	15	98.62	15	99.26	14
3	97.88	11	99.86	17	98.58	11
4	**99.78**	**13**	98.34	20	99.21	7
5	98.75	19	99.89	10	98.86	12
6	99.26	9	98.32	14	99.45	17
7	99.24	18	99.01	13	98.05	19
8	98.15	16	97.85	8	97.11	13
9	97.65	7	99.26	18	98.46	15
10	99.65	15	**99.85**	**15**	**99.52**	**10**
Average	98.79	13.4	99.03	14.1	98.72	13.3

Note: Acc, accuracy; #feat, optimal gene subsets.

Table 7 shows that proposed method achieves the highest accuracy as 99.85% on colon with 24 genes. The maximum classification accuracy for the proposed method is reported in Table 7 at bold value. In addition to, the results reported in Table 7, shows each run classification accuracy by the proposed method. In breast cancer dataset, maximum accuracy is 91.47% with 13 genes in the fifth run. The average classification performance regarding accuracy is achieved in this dataset as 88.64% with 17.2 genes.

Based on the average classification accuracy in Table 8, results that fashioned by proposed were almost consistent on each dataset. Interestingly, all runs have achieved 98% accuracy with less than 20 selected genes on the SRBCT dataset.

The ultimate goal of the researcher is to measure the predictive classification accuracy and also should be devoted to efforts in the estimation of an optimal number of genes. As reported in Table 8, proposed method demonstrates the effectiveness of the technique in all runs.

6.5. *Evaluation of proposed method with past literature*

The performance of hybrid approach for classification is evaluated in two phases. In the first phase, we use EGS to select the best gene data classified from tumors and cancer dataset. In the second phase as AGA to calculate the accuracy of the classification using subsets of selected characteristics of the gene datasets. As seen in Table 9, the proposed method shows the optimal number of genes and the corresponding accuracy in the given datasets with small genes. In Table 9, first column shows the works reported in the past literature and compared with proposed method, the second, the third, the fourth, the fifth, the sixth, and the seventh column show the percentages of accuracy (in percentage) and the number of genes (in brackets) obtained for each series of gene data. As shown in Table 9, our proposed method is very economical in terms of accuracy with small subgroups of genes in all biomedical data sets.

The selection of the necessary gene is based on the variety of gene methods which are selected from the works recently used for the classification and prediction of tumors in the bioinformatics domain. The symbol "*" means that no information is available. We have performed a literature study of about 150 research papers and the selected method of gene selection that tends to produce high-level accuracy and prediction. Also, the literature review on gene selection methods using the inductive learning technique, as shown in Table 10, has consistently provided maximum accuracy.

Table 9 – Comparison of the proposed method with other methods in each data set.

Method	Colon cancer	SRBCT	Breast	Lung	DLBCL	Leukemia
8-S PMSO [60]	94.2 (20)	*	*	100 (20)	98.7 (20)	98.1 (20)
LDA-GA [29]	98.80 (7)	*	*	99.00 (3)	99.0 (3)	*
MIMAGA [52]	80.4	86.36	82.47	92.00	*	94.09
GANN [61]	*	59.3 (19)	*	*	*	47.4 (21)
GBC [12]	98.38 (10)	100 (6)	*	100 (4)	*	*
DRF0 [62]	90 (10)	*	*	98.66 (17)	93.33 (11)	91.18 (13)
BBA [63]	77.08	*	57.10	*	77.22	90.35
PSO-dICA [64]	94.73 (20)	*	*	97.95 (25)	94.73 (30)	97 (72)
BDF [65]	97.46	*	86.22	99.14	89.44	95.81
BBHA-RF [66]	91.41	*	87.77	*	*	98.61
BDE-X Rankf [28]	75.0 (3)	*	*	98.0 (3)	92.9 (3)	82.4 (7)
Proposed with SVM	**98.90 (16)**	98.79 (13)	**88.64 (17)**	**99.03 (14)**	**99.01 (18)**	**98.72 (13)**

Table 10 – Comparison of the proposed framework with state-of-art filter techniques.

Gene selection	Learning classifier	Paper referred	Accuracy
Relief-F	K-nearest neighbor	Zainudin et al. [67]	85.20%
	Naïve Bayes	Huang et al. [23]	80.2%
	Support vector machine	Sasikala et al. [68]	76.69%
JMI	Random forest	Pashaei and Aydin [66]	81.17%
	Naïve Bayes	Hall [69]	83.78%
mRMR	Support vector machine	Moradi et al. [70]	83.38%
	Linear SVM	Ferreira et al. [71]	80.4%
Information gain	Decision tree	Huang et al. [23]	89.5%
	Random forest	Sahin and Subasi [72]	99.2%
Chi-square	Artificial neural network and decision tree	Houby [73]	69% and 73.33%
	Nearest neighbor, RIPPER, SVM, Naive Bayes, and C4.5	Jin et al. [24]	55%, 70%, 71.5%, 72%, and 73%

However, the number of selected genes adapted in our study is only five, but the theoretical, comparative assessment is performed on eleven articles published in the literature.

In order to demonstrate the efficiency of the proposed method, after applying the SVM classifier on the gene dataset regarding average accuracy and number of generations which are seen in Fig. 4. In Fig. 4, the maximum classification accuracy is achieved by a proposed method in the DLBCL dataset, and minimum accuracy is achieved in breast cancer dataset.

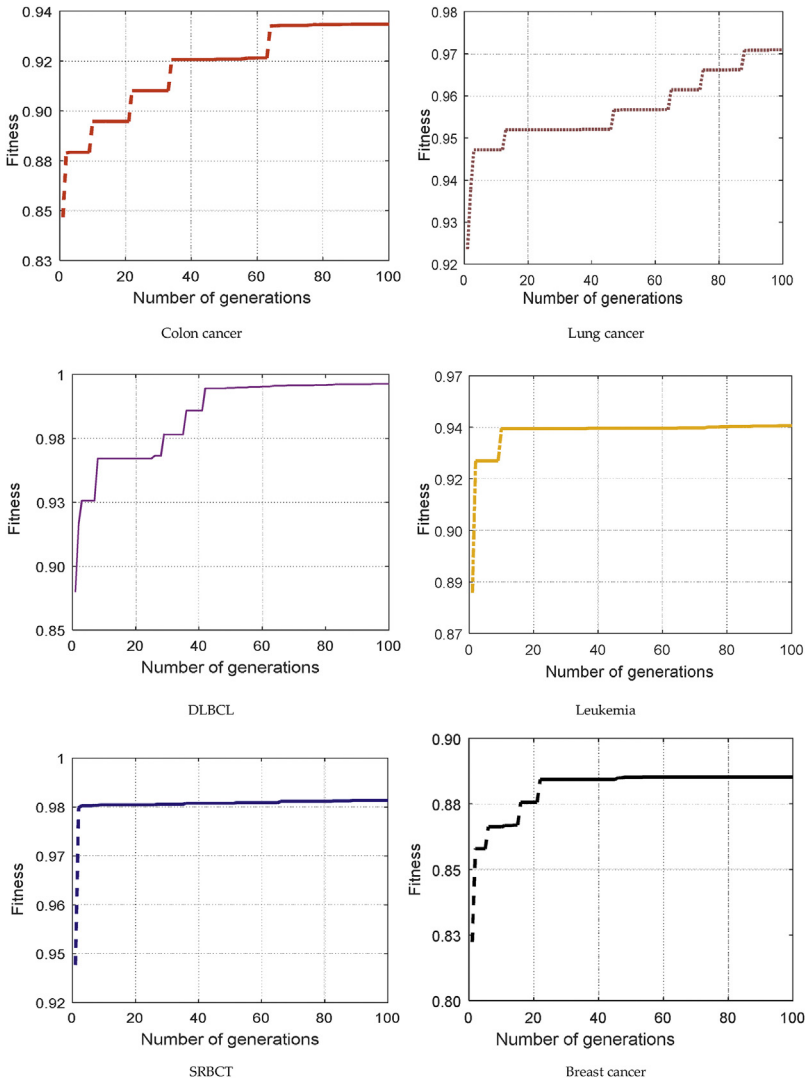

Fig. 4 – The average of fitness values and the number of generations for the proposed method with SVM classifier.

Table 11 – The best number of a gene selected by the proposed method.

Dataset	Acc	#feat	Name of genes
Breast cancer	88.64	17	AB033006, Contig29373_RC, Contig17079_RC, Contig26077_RC, NM_000910, NM_000988, NM_002306, AL035297, D42039, Contig37062, AF260665, Y00978, Contig9259, M77498, AB007877, Contig20635_RC, NM_013982
Colon cancer	98.9	16	X63432, L38696, T49397, J04026, H24754, T57780, X54163, T47562, H59599, H69834, U22897, L23823, U24077, M97676, R51644, X77548
SRBCT	98.79	13	gene16, gene46, gene81, gene113, gene211, gene306, gene508, gene724, gene828, gene1017, gene1286, gene1783, gene2228
Lung cancer	99.03	14	31323_r_at, 31431_at, 33062_at, 38489_at, 39272_g_at, 33306_at, 33846_at, 34376_at, 37586_at, 39850_at, 40191_s_at, 1332_f_at, 1182_at, 1245_i_at
DLBCL	99.01	18	GENE3067X, GENE3951X, GENE3510X, GENE3786X, GENE3667X, GENE2563X, GENE2235X, GENE2203X, GENE2859X, GENE2756X, GENE1329X, GENE3104X, GENE3371X, GENE691X, GENE770X, GENE2596X, GENE1575X, GENE1824X
Leukemia	98.72	13	AFFX-HUMISGF3A/M97935_5_at, AB006190_at, D14661_at, D38449_at, D79994_at, HG2573-HT2669_at, HG919-HT919_at, M13450_at, M26061_at, M64554_rna1_at, U70323_at, Y00486_rna1_at, U82979_at

Note: Acc, accuracy; #feat, number of optimal genes.

6.6. *Biological interpretation*

From the biological point of view, only a tiny number of relevant genes are relevant to microarray datasets which used in the diagnostic of cancer. The proposed method aims to identify a significant subset of small genes with the high classification accuracy. It is crucial to analyses these genes with others reported in the literature to find a biological meaning for each set of microarray data. In this subsection, we analyze the subset of selected final genes (see Table 11) from our proposed model corresponding to the most significant genes obtained in each data set.

As can be seen in Table 12, the proposed method using four classifiers with the help of selected optimal gene subsets has achieved the maximum accuracy (acc) as 98.58% in Colon Cancer using SVM and minimum accuracy as 88.54% in breast cancer using NB. In Table 13, we have reported the average execution time (in seconds) for 100 iterations each of the iBPSO (↑ w)-NB [74] for six microarray datasets. It can be summarized that our method in Colon dataset takes lesser time as compare to iBPSO to find the optimal gene subsets.

Table 12 – Performance analysis of proposed method with selected subsets of genes using four classifier.

Classifier	Datasets					
	Breast	Colon	DLBCL	SBRCT	Lung	Leukemia
SVM	89.62	**98.58**	98.32	95.68	98.35	98.01
NB	88.54	97.63	97.54	93.05	97.87	97.35
k-NN	91.01	97.89	96.07	94.09	98.01	95.03
DT	90.37	96.57	97.89	95.37	97.56	95.01

Table 13 – Comparison of average execution time for iBPSO (↑ w)-NB for six microarray datasets.

Data	iBPSO (↑ w)-NB	Proposed
Breast	231.80	165.98
Colon	39.27	31.51
DLBCL	–	186.94
SBRCT	302.81	210.63
Lung	311.21	354.96
Leukemia	–	335.86

6.7. *Statistical results*

In general, the Friedman test is a non-parametric statistical method which is applied for ranking the performance of the algorithms. The main aim of the Friedman test is to find whether any significant difference exists between the results of different algorithms. This is based on the null hypothesis that there is no variation in the performance of all algorithms [75]. The finest algorithm becomes the lowest rank while the worst performing algorithm gets the highest rank as we can see in Fig. 5. The average rank obtained by each algorithm on all datasets is calculated for determining the Friedman test shown in Table 14.

Hochberg's procedure rejects those hypotheses that have an unadjusted p-value as 0.005556. Li's procedure rejects those hypotheses that have an unadjusted p-value as 0.046397 as shown in Table 15.

7. Conclusions

Due to the insufficient knowledge of the intrinsic data population of the high-dimensional dataset, researchers have shown the difficulty in extracting the biomarker genes for further data analysis such as classification. To resolve this problem, researchers have applied the effective gene selection approaches on large-scale feature sets. Although the particular gene selection method is capable of identifying biomarker genes from the high dimensional datasets, but it has few drawbacks. To overcome these problem, this paper presented a two-stage gene selection method using ensemble learning; and wrapper algorithm which quickly identifies the important genes thereby classify a disease correctly at the less computational time. In the first stage, an ensemble of three gene selection methods and then, an ensemble of two gene selection methods and output of first layer strategy

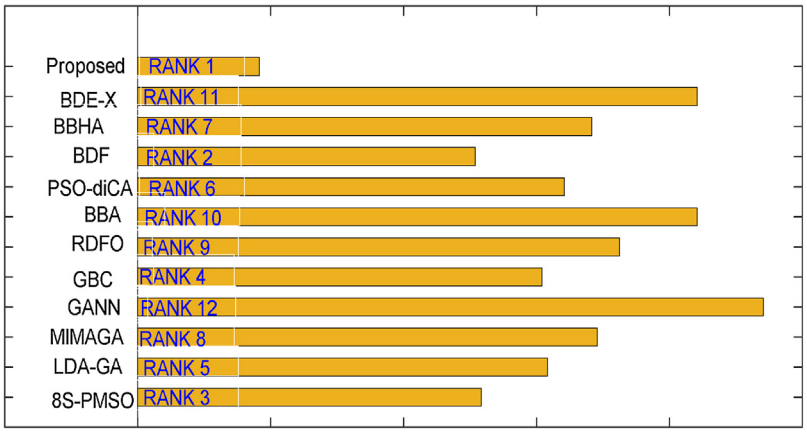

Fig. 5 – Ranking of existing algorithms and proposed method by Friedman test.

Table 14 – Average rankings of the algorithms (Friedman).

Algorithm	Ranking
8-S PMSO	5.1667
LDA-GA	6.1667
MIMAGA	6.9167
GANN	9.4167
GBC	6.0833
DRF0	7.25
BBA	8.4167
PSO-dICA	6.4167
BDF	5.0833
BBHA-RF	6.8333
BDE-X Rankf	8.4167
Proposed	1.8333

Table 15 – Post hoc comparison table for $p = 0.05$ (Friedman).

i	Algorithm	$z = (R_0 - R_i)/SE$	p	Holm Hochberg	Li
11	GANN	3.642915	0.00027	0.004545	0.046397
10	BDE-X Rankf	3.162531	0.001564	0.005	0.046397
9	BBA	3.162531	0.001564	0.005556	0.046397
8	DRF0	2.602082	0.009266	0.00625	0.046397
7	MIMAGA	2.441954	0.014608	0.007143	0.046397
6	BBHA-RF	2.401922	0.016309	0.008333	0.046397
5	PSO-dICA	2.201762	0.027682	0.01	0.046397
4	LDA-GA	2.081666	0.037373	0.0125	0.046397
3	GBC	2.041634	0.041188	0.016667	0.046397
2	8-S PMSO	1.601282	0.109315	0.025	0.046397
1	BDF	1.561249	0.118465	0.05	0.05

is combined. In the second stage, wrapper approach as AGA combines the learning algorithms as an evaluation function to reach with optimal gene subsets. In order to show the supremacy of the proposed method, we have used six gene datasets. The experimental results show that proposed method provides additional support to a significant reduction of cardinality and outperforms the state-of-arts gene selection method in terms of classification accuracy and an optimal number of genes. In testing of the proposed method, SVM behaves as the best fitness function when the diagnosis of tumor or cancer disease is performed.

REFERENCES

[1] Ang JC, Mirzal A, Haron H, Nuzly H, Hamed A. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinform 2016;13(5):971–89.
[2] Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Appl Soft Comput J 2018;62:441–53.
[3] Nakariyakul S. High-dimensional hybrid feature selection using interaction information-guided search. Knowl Based Syst 2018;145:59–66.
[4] Hancer E, Xue B, Zhang M. Differential evolution for filter feature selection based on information theory and feature ranking. Knowl Based Syst 2018;140:103–19.
[5] Bonilla-Huerta E, Hernández-Montiel M, Morales-Caporal R, Arjona-López M. Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data. IEEE/ACM Trans Comput Biol Bioinform 2016;13(1):12–26.
[6] Mohamad MS, Omatu S, Deris S, Yoshioka M. A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data. IEEE Trans Inf Technol Biomed 2011;15(6):813–22.
[7] Liu H, Li D. Predicting novel salivary biomarkers for the detection of pancreatic cancer using biological feature-based classification. Pathol Res Pract 2016.
[8] Hancer E, Xue B, Karaboga D, Zhang M. A binary ABC algorithm based on advanced similarity scheme for feature selection. Appl Soft Comput J 2015;36:334–48.
[9] Kumar S, Kumar P, Kumar A, Swarnkar T. Elitism based multi-objective differential evolution for feature selection: a filter approach with an efficient redundancy measure. J King Saud Univ – Comput Inf Sci 2017.
[10] Wang A, An N, Yang J, Chen G, Li L, Alterovitz G. Wrapper-based gene selection with Markov blanket. Comput Biol Med 2017;81:11–23.
[11] Aziz R, Verma CK, Srivastava N. A novel approach for dimension reduction of microarray. Comput Biol Chem 2017;71:161–9.

[12] Alshamlan HM, Badr GH, Alohali YA. Genetic Bee Colony (GBC) algorithm: a new gene selection method for microarray cancer classification. Comput Biol Chem 2015;56:49–60.
[13] Alshamlan H, Badr G, Alohali Y. mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling, vol. 2015. 2015.
[14] Aziz N, Verma R, Jha CK, Srivastava M. Artificial neural network classification of microarray data using new hybrid gene selection method. Int J Data Min Bioinform 2017;17 (1):42–65.
[15] Lee S, Xu Z, Li T, Yang Y. A novel bagging C4. 5 algorithm based on wrapper feature selection for supporting wise clinical decision making. J Biomed Inform 2017.
[16] Wan Y, Wang M, Ye Z, Lai X. A feature selection method based on modified binary coded ant colony optimization algorithm. Appl Soft Comput J 2016;49:248–58.
[17] Das AK, Goswami S, Chakrabarti A, Chakraborty B. A new hybrid feature selection approach using feature association map for supervised and unsupervised classification. Expert Syst Appl 2017;88:81–94.
[18] Paul D, Su R, Romain M, Sébastien V, Pierre V, Isabelle G. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Comput Med Imaging Graph 2016.
[19] Goldberg JHH, David E. Genetic algorithms and machine learning. Mach Learn 1988;3(2):95–9.
[20] Zheng H, Zhang Y, Liu J, Wei H, Zhao J, Liao R. A novel model based on wavelet LS-SVM integrated improved PSO algorithm for forecasting of dissolved gas contents in power transformers. Electr Power Syst Res 2018;155:196–205.
[21] Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, Alonso-Betanzos A. Ensemble feature selection: homogeneous and heterogeneous approaches. Knowl Based Syst 2017;118:124–39.
[22] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput 2005;3(2):185–205.
[23] Huang Y, McCullagh PJ, Black ND. An optimization of ReliefF for classification in large datasets. Data Knowl Eng 2009;68(11):1348–56.
[24] Jin X, Xu A, Bie R, Guo P. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene. International Workshop on Data Mining for Biomedical Applications. Berlin: Springer; 2006. p. 106–15.
[25] Mode H. Joint & conditional entropy, mutual information. Part I. Joint and conditional entropy; 2014.
[26] Sadri A, Ren Y, Salim FD. Information gain-based metric for recognizing transitions in human activities. Pervasive Mob Comput 2017;38:92–109.
[27] Leung Y, Hung Y. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. IEEE/ACM Trans Comput Biol Bioinform 2010;7(1):108–17.
[28] Apolloni J, Leguizamón G, Alba E. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 2016;38:922–32.
[29] Bonilla-Huerta E. Hybrid filter-wrapper with a specialized random multi-parent crossover operator for gene selection and classification problems. International Conference on Intelligent Computing. Berlin, Heidelberg: Springer; 2011. p. 453–61.
[30] Yin H, Member S, Jha NK. A health decision support system for disease diagnosis based on wearable medical sensors and machine learning ensembles. IEEE Trans Multi-Scale Comput Syst 2017;3(4):228–41.
[31] Silwattananusarn T, Kanarkard W, Tuamsuk K. Enhanced classification accuracy for cardiotocogram data with ensemble feature selection and classifier ensemble. J Comput Commun 2016;20–35.
[32] Al-Rajab M, Lu J, Xu Q. Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. Comput Methods Programs Biomed 2017;146:11–24.
[33] Ghorai S. Multicategory cancer classification from gene expression data by multiclass NPPC ensemble; 2010;41–6.
[34] Rachman AA. Cancer classification using fuzzy C-means with feature selection. 12th Int. Conf. Math. Stat. Their Appl.. 2016. pp. 31–4.
[35] Zhang J, Chung HS, Member S, Lo W. Clustering-based adaptive crossover and mutation probabilities for genetic algorithms. IEEE Trans Evol Comput 2007;11(3):326–35.
[36] Wei X, Shao W, Zhang C, Li J, Wang B. Improved self-adaptive genetic algorithm with quantum scheme for electromagnetic optimisation. Microwaves Antennas Propag 2014;8(12):965–72.
[37] Alirezazadeh P, Fathi A, Abdali-Mohammadi F. A genetic algorithm-based feature selection for kinship verification. IEEE Signal Process Lett 2015;22(12):2459–63.
[38] Pes B, Dessì N, Angioni M. Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. Inf Fusion 2017;35:132–47.
[39] Laura Emmanuella LEA, De Paula Canuto AM. Filter-based optimization techniques for selection of feature subsets in ensemble systems. Expert Syst Appl 2014;41(4):1622–31.
[40] Ebrahimpour MK, Eftekhari M. Ensemble of feature selection methods: a hesitant fuzzy sets approach. Appl Soft Comput J 2017;50:300–12.
[41] Moradi P, Gholampour M. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Appl Soft Comput J 2016;43:117–30.
[42] Srinivas M, Patnaik LM. Adaptive probabilities of crossover and mu tation in genetic algorithms. IEEE Trans Syst Man Cybern 1994;24(4):656–67.
[43] Arauzo-Azofra A, Benitez J, Castro J. A feature set measure based on relief. Proc. Fifth Int. Conf. Recent Adv. Soft Comput.. 2004. pp. 104–9.
[44] Mundra PA, Rajapakse JC. SVM-RFE with MRMR filter for gene selection. IEEE Trans Nanobiosci 2010;9(1):31–7.
[45] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.
[46] Thakur M. A new genetic algorithm for global optimization of multimodal continuous functions. J Comput Sci 2013;5 (2):298–311.
[47] Kundakcı N, Kulak O. Hybrid genetic algorithms for minimizing makespan in dynamic job shop scheduling problem. Comput Ind Eng 2016;96:31–51.
[48] Cai Z, Zhu W. Feature selection for multi-label classification using neighborhood preservation. IEEE/CAA J Autom Sin 2018;5(1):320–30.
[49] Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst 2008;14 (1):1–37.
[50] Bron E, Smits M, Van Swieten J, Niessen W, Klein S. Feature selection based on SVM significance maps for classification of dementia. Int Work Mach Learn Med Imaging 2014;19 (5):272–9.
[51] Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinform 2006;7(1).
[52] Lu H, Chen J, Yan K, Jin Q, Xue Y, Gao Z. A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 2017.

[53] Li Y, Yang Y, Li G, Xu M, Huang W. A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and mRMR feature selection. Mech Syst Signal Process 2017;91:295–312.
[54] Bashir S, Qamar U, Khan FH. IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. J Biomed Inform 2016;59:185–200.
[55] Lai C, Yeh W, Chang C. Gene selection using information gain and improved simplified swarm optimization. Neurocomputing 2016;218:331–8.
[56] Rankawat SA, Dubey R. Robust heart rate estimation from multimodal physiological signals using beat signal quality index based majority voting fusion method. Biomed Signal Process Control 2017;33:201–12.
[57] Chen Y, Lin C. Combining SVMs with various feature selection strategies. Featur Extr 2006;(1):315–24.
[58] Soufan O, Kleftogiannis D, Kalnis P, Bajic VB. DWFS: a wrapper feature selection tool based on a parallel genetic algorithm. PLoS ONE 2015;10(2):e0117988.
[59] Van't Veer GJ, Dai LJ, Van De Vijver H, He MJ, Hart YD, Mao AA, et al. Gene expression profiling predicts clinical outcome of breast cancer. Lett Nat 2002;415(345):530–6.
[60] Alba JGE. Parallel multi-swarm optimizer for gene selection in DNA microarrays. Appl Intell 2012;37(2):255–66.
[61] Tong DL, Mintram R. Genetic Algorithm-Neural Network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. Int J Mach Learn Cybern 2010;1(1–4):75–87.
[62] Bolón-Canedo A, Sánchez-Maroño V, Alonso-Betanzos N. Distributed feature selection: an application to microarray data classification. Appl Soft Comput 2015;30:136–50.
[63] Chen Y, Li Y, Wang G, Zheng Y, Xu Q, Fan J. A novel bacterial foraging optimization algorithm for feature selection. Expert Syst Appl 2017;83:1–17.
[64] Mollaee M, Moattar MH. A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybern Biomed Eng 2016;36(3):1–9.
[65] Medjahed SA, Saadi TA, Benyettou A, Ouali M. Kernel-based learning and feature selection analysis for cancer diagnosis. Appl Soft Comput J 2017;51:39–48.
[66] Pashaei E, Aydin N. Binary black hole algorithm for feature selection and classification on biological data. Appl Soft Comput J 2017;56:94–106.
[67] Zainudin M, Sulaiman M, Mustapha N, Perumal T, Nazri A, Mohamed R, et al. Feature selection optimization using hybrid Relief-f with self-adaptive differential evolution. Int J Intell Eng Syst 2017;10(3):21–9.
[68] Sasikala S, Balamurugan SA, Geetha S. A novel memetic algorithm for discovering knowledge in binary and multi class predictions based on support vector machine. Appl Soft Comput J 2016;49:407–22.
[69] Hall M. Correlation-based feature selection for machine learning; 1999.
[70] Moradi P, Rostami M. Integration of graph clustering with ant colony optimization for feature selection. Knowl Based Syst 2015;84:144–61.
[71] Ferreira AJ, Figueiredo MAT. Efficient feature selection filters for high-dimensional data. Pattern Recognit Lett 2012;33(13):1794–804.
[72] Sahin H, Subasi A. Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques. Appl Soft Comput J 2015;33:231–8.
[73] El Houby EMF. A framework for prediction of response to HCV therapy using different data mining techniques. Adv Bioinform 2014;11.
[74] Jain I, Kumar V, Jain R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl Soft Comput J 2018;62:203–15.
[75] Sree Ranjini KS, Murugan S. Memory based hybrid dragonfly algorithm for numerical optimization problems. Expert Syst Appl 2017;83:63–78.