Applied Soft Computing 62 (2018) 230–250

Contents lists available at ScienceDirect

Applied Soft Computing

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

Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach

Prativa Agarwalla [a,*], Sumitra Mukhopadhyay [b,*]

[a] Heritage Institute of Technology, Kolkata, India
[b] Institute of Radiophysics and Electronics, Kolkata, India

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form 13 August 2017
Accepted 11 October 2017
Available online 4 November 2017

Keywords:
DNA microarray
Biological pathway
Feature selection
Particle swarm optimization (PSO)
Artificial bee colony (ABC)

ABSTRACT

Background: Understanding of molecular mechanism, lying beneath the carcinogenic expression, is very essential for early and accurate detection of the disease. It predicts various types of irregularities and results in effective drug selection for the treatment. Pathway information plays an important role in mapping of genotype information to phenotype parameters. It helps to find co-regulated gene groups whose collective expression is strongly associated with the cancer development.
Method: In this paper, we have proposed a bi-stage hierarchical swarm based gene selection technique which combines two methods, proposed in this paper for the first time. First one is a multi-fitness discrete particle swarm optimization (MFDPSO) based feature selection procedure, having multiple fitness functions. This technique uses multi-filtering based gene selection procedure. On top of it, a new blended Laplacian artificial bee colony algorithm (BLABC) is proposed and it is used for automatic clustering of the selected genes obtained from the first procedure. We have performed 10 times 10-fold cross validation and compared our proposed method with various statistical and swarm based gene selection techniques for different popular cancer datasets.
Result: Experimental results show that the proposed method as a whole performs significantly well. The MFDPSO based system in combination with BLABC generates a good subset of pathway markers which provides more effective insight into the gene-disease association with high accuracy and reliability.

1. Introduction

The fundamental cause of cancer [1] involves abnormal growth of cells and the underlying factor is the certain changes at the genetic expression level. Gene controls the functioning of a living cell. When the normal expression profile of gene changes, it causes some abnormalities. However, a series of pathway activities related to biological process collectively leads to the growth of cancer. So it becomes very essential to analyze the molecular mechanism and genomic expression profile of cancer. The efficient mapping of genomic information to the phenotype parameter can help in better understanding in the progression and early detection of the disease. Moreover, analysis about the changes in gene expression helps in proper cancer identification and class prediction that is necessary for the drug selection in the course of treatment. The variation in the expression profile of genes can be visualized by the recent advancement of microarray technology [2] which is nothing but the expression level of thousands of genes in a single chip. To prepare the gene expression data, samples are collected from patients having different classes of disease and then through the hybridization procedure the changes in expression level are examined. The schematic diagram of generating microarray gene expression is shown in Fig. 1 where the dataset is formed for two classes of disease. Now, to identify differentially expressed genes for different classes and to study their effects on diseases, we need to investigate gene expression dataset. This leads to statistical and analytical challenges because of its huge dimension. Again, the availability of larger number of genes compared to the number of samples in the dataset can cause the overfitting of classification model and the genetic heterogeneity across patients weakens the discriminating power of individual gene [3]. Also the presence of noisy genes makes it very challenging to understand the nature of gene regulation in the cancer samples compared to the normal cell. The removal of uninformative genes not only reduces the processing time but also diminishes the interference of noisy or unwanted information leading to the incorrect classification of data. The procedure of relevant gene (feature) selection from the gene expression profile and the expected outcome is described by a schematic diagram, given

* Corresponding authors.
E-mail address: sumitra.mu@gmail.com (S. Mukhopadhyay).

https://doi.org/10.1016/j.asoc.2017.10.024
1568-4946/

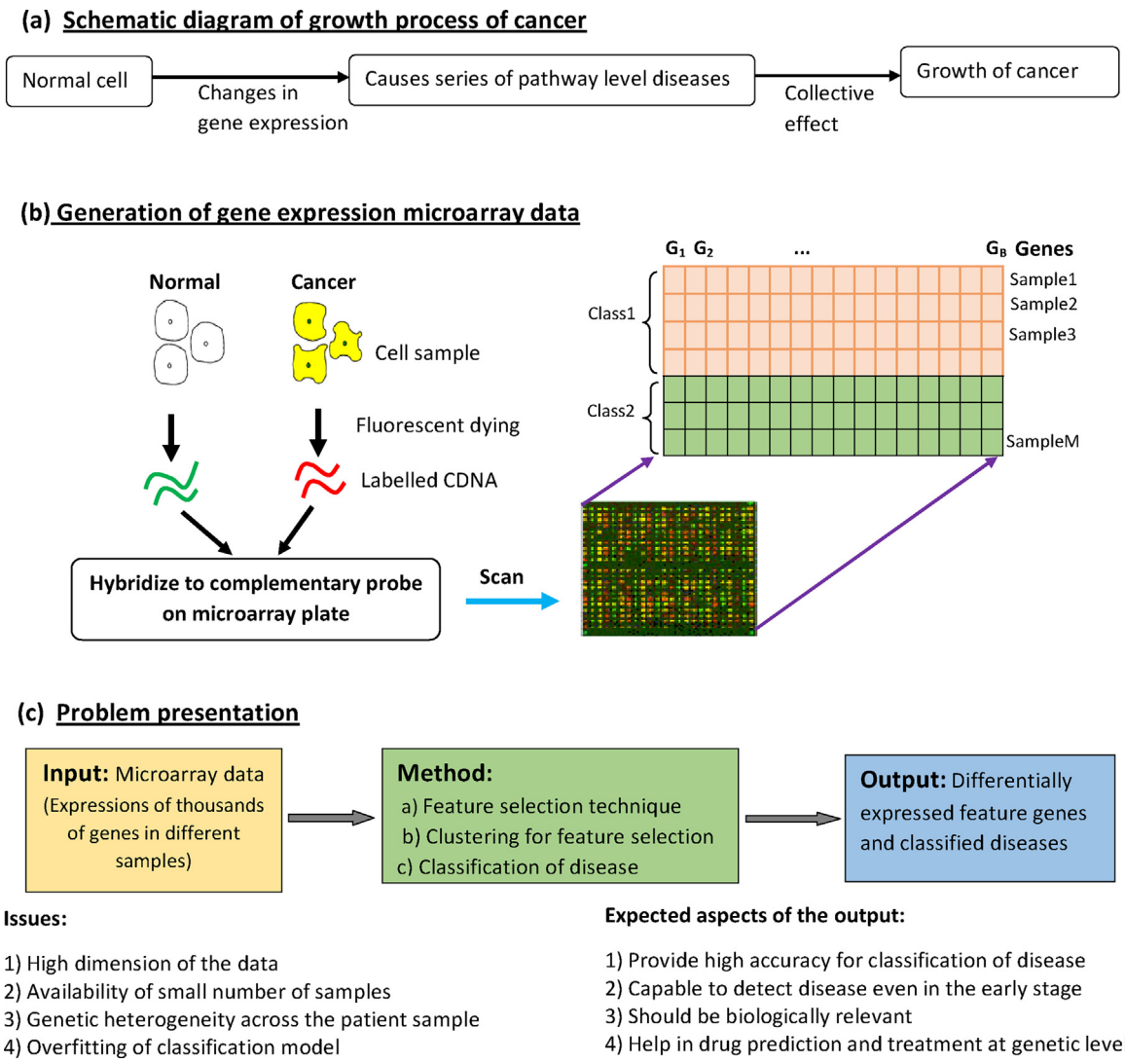

Fig. 1. General problem presentation of gene marker selection for cancer growth.

in Fig. 1. To solve the above stated issues, different methodologies are approached in different literatures [4–11] for the feature selection and classification of cancer, but the reproducibility of results is a challenging task as in most of the cases the resultant genes are varying. Also, the selected genes which are providing good classification result may not always be biologically relevant to cancer progression. So, the genes used for cancer prediction can lead to a false discovery of disease which could be vital for the patient. It would be better to enrich with additional biological knowledge rather concentrating only on high classification accuracy.

While exploring the features of significant non-redundant genes participating in a tumor progression, it has been observed that those genes are functioning in a co-regulated manner and work as a group. They confer a selective growth advantage during the development of certain cancer. So, those differentially expressed genes having a similar biological contribution to the progression of tumor are to be identified and their phenotypical changes at the pathway level [12,1] for all the patients are to be studied for the treatment of prior related pathway diseases. So shifting to pathway based study can help in better understanding of prior disease related information and how a disease occurs altogether. Several publicly available databases are developed for accessing the pathway information and detailed information about the interaction of genes and their regulatory pathways [13,14]. To use those pathway-based marker genes in disease classification, initially we need a way to infer the activity of a given biological pathway on the gene expression and then the differential genes (features) significantly associated with the disease outcomes are to be identified for the efficient classification of samples.

Rather, computing the statistical significance of each and individual genes, we propose an efficient swarm based stochastic method for the selection of pathway marker genes. The marker genes are selected from functional genomic expression data incorporating different parametric and non-parametric statistical techniques that are able to reflect the heterogeneity in the gene expression level. The proposed method consists of two basic stages: first a multi-fitness discrete particle swarm optimization (MFDPSO) technique in combination with multiple statistical filters is used for the initial selection of genes from biomedical databases. It reduces the dimensionality of the microarray data as well as the combined effect of different filters provides a better differential gene subset. Reduced feature set obtained in the first stage is used in the second stage where distinct gene expression levels in each class are identified using a newly proposed blended Laplacian artificial bee colony (BLABC) based automatic inter-class clustering technique. The resultant genes can be used for classification. It solves the problem of overfitting of the classifiers and reduces the false discovery rate of the disease identification.

Recently, many feature selection and classification techniques utilizing various bio-inspired algorithm have been introduced in order to extract a small subset of relevant genes from the microarray dataset [4–10,15–17]. Also, different approaches are developed by the researchers for identification of pathway marker genes [18,11] related to different diseases. The statistical approaches like mean, median [19] methods are applied for this purpose. Also, other different methodologies like principal component analysis (PCA) [20], log-likelihood ratio (LLR) [21], condition responsive genes (CORGs) [3], binary version of PSO (BPSO) [22] have also been applied by formulating the problem as a global optimization problem for the extraction of significant genes at pathway level. However, in many of the cases, the results are not satisfactory and they hardly identify a common unique feature set for the classification problem.

Our study presents a robust bi-stage hierarchical swarm based pathway marker selection method that combines the filter and wrapper based approaches in the first stage. It applies *t*-test [23], *F*-test [24], Wilcoxon Ranksum test [23] and correlation statistic [24] as the fitness metric of MFDPSO for detecting differentially expressed genes in microarray studies. Multi-filter technique, combined with a little modified version of basic PSO, shows more promising results than a single-filter technique, as the later may fail to notice some informative genes required for the development of biological insight in cancer progression. The MFDPSO itself works very effectively as it has multiple fitness functions for evaluation of each filter metric. It implies that each particle of the swarm has as many fitness values as there are numbers of pathways present in the dataset for each filtering technique. The entire proposed modification develops a rigorous selection procedure and the elimination of non-competitive and uninformative genes become obvious. The discrete value of PSO introduces the stochastic nature in the selection process of genes. So, as a whole, for each pathway activity, a robust gene subset gets selected. Next, a new, more effective set of informative gene is produced, while combining the results obtained from the previous step of four wrapper-filter methods. Here, for the fusion purpose, we have formulated the problem as a constrained optimization problem and then the best gene subset is selected for further processing. In the second stage, the class dependent pathway feature genes are selected using the BLABC automatic clustering algorithm for each class. We have modified the ABC algorithm and a new blended Laplacian ABC is proposed. The novel approach of BLABC for inter-class clustering methodology adds a great value when those features are used for classification. The developed BLABC based automatic clustering technique is very robust, and it is justified using some commonly used benchmark dataset available in [25]. For the performance validation of the proposed scheme, various real life cancer microarray datasets [26–28] are chosen and 10 times 10 fold cross-validation is done. The pathway markers are selected which are further verified against biological pathways to prove the biological significance of the selection. The experimental results are also compared with some popular techniques presented in different literatures to validate the superiority of the algorithm.

The rest of the paper is organized as follows: a brief review work is presented in Section 2 followed by the description of the proposed technique for pathway marker selection in Section 3. Section 4 presents the experimental result for the different cancer dataset. Next the biological relevance of the result is also given. Finally, in Section 5, we conclude our work .

2. Review work

Selection of features and classification of gene expression data is always an interesting research topic in the field of bioinformatics. The problem can be treated as an optimization problem where the aim is to select a non-redundant subset of features or attributes. Again, the problem can be formulated as a clustering problem where the centers of the clusters are the selected features of the problem. Those selected features help in the classification of the disease, as well as, do needful for gene discovery and drug prediction for the treatment. Several novel methods are proposed by the researchers to solve the issue of biomarker selection for disease. Different filtering techniques are utilized for the selection of genes such as *t*-test combined with self organizing map (SOM) [29], a maximum relevance minimum redundancy method (MRMR) [30] using mutual information, etc. However, the filtering approaches suffer from low comprehensibility. Researchers have come up with several classifiers dependent statistical learning processes like multi-criterion support vector machines (SVM) [31], SVM with fuzzy concept where fuzzy preference based rough set with semi-supervised SVMs [32] are applied for classification of data using feature genes. The filters and classifiers are also implemented together for the classification of disease [4]. Partial least square method [33] is also used to reduce the dimension of microarray dataset and a discriminate analysis method is applied in this field [34].

Different embedded techniques are proposed where variants of bio-inspired optimization algorithms along with the classifiers are used for microarray data classification, considering it as a feature selection optimization problem. Adaptive genetic algorithm (GA) embedded with k-nearest neighbour method is applied for the relevant gene selection and classification purpose [35]. In another work [36], GA is used with SVM where a new fitness function, defined by the combined results from SVM and GA is used to optimize the process. Again, a novel class dependent feature selection method for cancer biomarker discovery with different classifier is introduced in the literature [6]. A decision tree model empowered by PSO [5] is also proposed for the problem. Different hybrid algorithms embedded with filters and classifiers are proposed in the literature [16,8,9]. For an example, gene selection for cancer tumor detection using memetic algorithm [7] is introduced. More recently, in 2016, a novel multiple filter embedded hybrid approach is investigated for the purpose [8]. Hybrid ant bee algorithm [9] is also applied for microarray studies. Those embedded algorithms are less prone to local optima compared to deterministic methods as they directly interact with the classifier. But, they have a higher risk of overfitting. The problem of feature selection is again solved using clustering approaches where the features are selected based on different proximity measures [24]. For an example, simulated annealing based on fuzzy fitness function is introduced [10]. Application of different clustering techniques on microarray data are demonstrated in another literature [37]. However, the resultant genes are able to provide good accuracy for the classification of disease, but most of the genes suffer from the lack of biological relevancy. Again, for the identification of genes, a multi-objective PSO based framework is developed in the literature [38]. In a very recent work in 2017 [17], information gain is applied at the initial selection stage and then GA and Genetic Programming are used for the estimation of genes. A novel pipelining method is used in [15] for ranking of genes and the significant genes are selected for classification. Two hybrid wrapper-filter techniques are also proposed [16] where a combination of binary differential evolution (BDE) algorithm with a rank-based filter method is applied for microarray dataset.

In another set of work, some pathway activation techniques are applied for the outlining of gene expression data so that the selected genes can be treated as biological pathway markers for the development of the disease. Recently, a number of pathway activity inference methods have been proposed for gene selection at the pathway level. Pathway knowledge is compiled into database

and the gene expression values of member genes are summarized to extract the feature subset. The pathway markers are then utilized as the features for cancer classification and they also help in better biological interpretation of results. Some approaches use gene expression parametrically by representing pathway activity with a function. Some research papers estimate the probabilities of pathway activation based on the consistency of changes in gene expression [12]. Alternative approaches engage normal cells to activate pre-selected oncogenic pathways. Those activated pathways determine gene signatures which can distinguish tumor characteristics [13]. The method applied some statistical estimation in a deterministic way. Tomfohr et al. [20] used principal component analysis to estimate the activity of a given pathway. For the estimation of heterogeneous microarray data where the expression of a particular gene varies from patient to patient for the same class of disease, it is quite difficult to ensure the separability of two classes with simple statistical techniques. Lee et al. [3] proposed a method called the condition responsive genes (CORGs) where combined expression levels are used to discriminate the phenotypes of interest. In another work, pathway activities are estimated using the log-likelihood ratio (LLR) [21] method. But the selection of relevant genes again depends on some deterministic method which greatly affects the class prediction capability. So, the use of stochastic bio-inspired algorithm may be used for the analysis and detection of responsible genes. More recently, a multi-objective optimization model based on GA (MOGA) is employed to identify driver genes and driver pathways promoting cancer proliferation [18]. A BPSO [22] has been adopted to find the pathway regulated genes based on mean t-score. In this method, the genes are selected using only one statistical technique, and are used for the disease classification purpose. The features that are selected in this way may cause the overfitting of the classifier. So, if we use more than one filter and rather averaging their effect, if we model the system with a new multi-fitness concept in the pathway based domain, more relevant genes may be found. Again, an automatic clustering technique can eliminate the features which are carrying redundant information by generating good cluster centers.

In this paper, our aim is to identify gene pathways having significant associations with the clinical outcomes. The steps of the proposed scheme are stated below.

- The microarray dataset is first pre-processed and then KEGG [14] pathways are constructed using existing pathological information of genes.
- Then the pathway activity is estimated by applying a multi-fitness discrete PSO (MFDPSO) procedure based on multi-filtering technique to identify the relevant genes at pathway level.
- The gene sets obtained from multi-filtering methods are combined and subsequently top genes are selected after fusion.
- Next, the proposed blended Laplacian ABC (BLABC) algorithm selects the optimal feature gene subset.

The proposed method demonstrates classification accuracies that are better in comparison to the conventional feature selection methods and pathway based analysis. It also provides a strong biological interpretation about the association of the expression profile with a particular type of disease. In the next section, the proposed gene selection technique is discussed in details.

3. Proposed gene selection technique

3.1. General framework

Two stages of feature selection tasks are investigated in this paper. Both stages use bio-inspired optimization algorithm for its searching purpose. The general framework is summarized in Fig. 2. The experimental dataset is first pre-processed to eliminate redundancy and KEGG [14] pathways are constructed using existing pathological information of genes. After inferring pathway activity total dataset is divided into two separate subsets. One is used for training purpose and the remaining dataset is used for testing purpose. The training dataset is further used for feature reduction and selection task using bio-inspired optimization procedures. PSO [39] is a state-of-the-art high performing optimization technique that provides the optimal solution of a problem very efficiently. In order to find out the optimal feature subset in the first stage, we have integrated a multi-filter multi-fitness concept with the modified version of basic discrete PSO (MFDPSO). The multi-filter approach is realized with some popular statistical filtering techniques. Here, the optimization problem is formulated using evaluation scores of different statistical parametric and non parametric tests such as *t*-test [23], *F*-test [6], Wilcoxon Ranksum (WR) test [23] and correlation method [24]. Four unique feature subsets are subsequently formed using the MFDPSO based method. All the feature sets are then combined using a fusion method and an optimized gene subset is obtained. In the second stage, a modified ABC algorithm, termed as Blended Laplacian ABC algorithm (BLABC) is applied on the selected gene subset. The top differentially expressed genes (DEGs) from the combined subset are selected which are 10 times 10-fold cross validated using the testing samples. The detailed descriptions of each stage are described below.

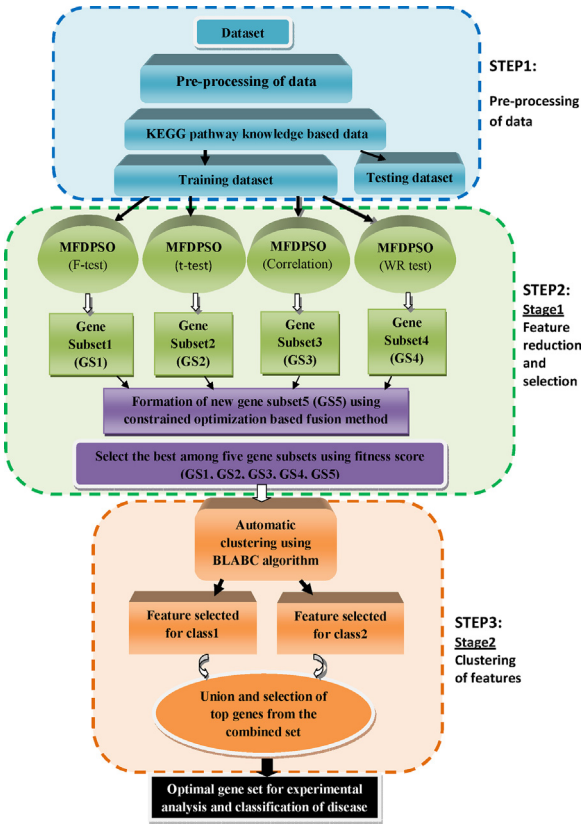

Fig. 2. General framework of the proposed bi-stage hierarchical pathway selection.

3.2. Preprocessing of data

3.2.1. Elimination of noisy genes

Two classes of raw microarray data as explained in Fig. 1 are collected from different types of cancer dataset [26,27,29,40,41,28]. They are processed before computation by eliminating the genes

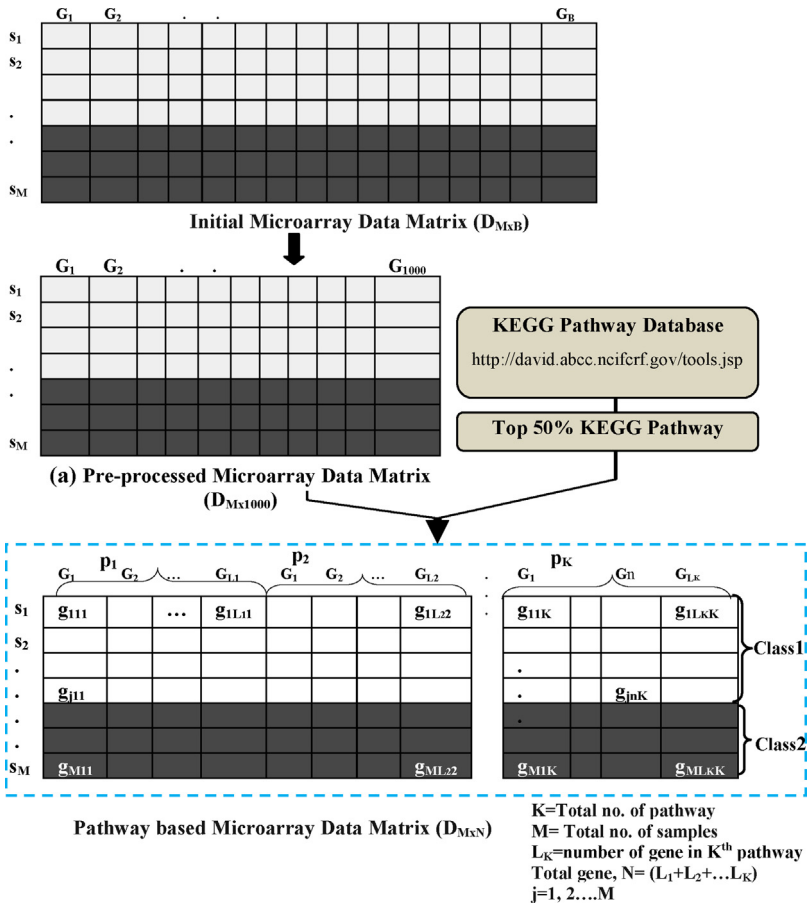

Fig. 3. Inferring pathway information in the microarray dataset.

having missing value in the dataset. Therefore, we obtain a two class microarray data matrix of size ($M \times B$), as shown in Fig. 3, where M is the sample number in the row and B represents the total number of genes present in the columns of the matrix. After that, to eliminate the noisy and redundant genes, the signal-to-noise ratio (SNR) is computed for each gene of the dataset, considering all the samples together. As a result, we obtain B number of SNR values, having one value for each gene. The expression of SNR for the nth gene is given in Eq. (1). m_n^1 and m_n^2 represent the mean of the expression of nth gene over the samples of the first class and second class respectively. σ_n^1 and σ_n^2 represent the standard deviation of the expression of nth gene for the first and second class respectively. After obtaining the SNR values, the genes are arranged in descending order of the values and top 1000 genes are selected for computation.

$$SNR_n = \frac{(m_n^1 - m_n^2)}{(\sigma_n^1 - \sigma_n^2)} \tag{1}$$

where, $n = 1, 2, \ldots, B$ and $B \gg 1000$.

3.2.2. Gathering pathway information

In the next step, we have searched for those 1000 genes' affymetrix IDs in the http://david.abcc.ncifcrf.gov/tools.jsp website and the corresponding KEGG pathway [14] information is collected. Top 50% KEGG pathways are considered for processing purpose. The expressions of the genes which are present in the top 50% pathway activities are collected from the pre-processed DNA microarray data matrix of size $D_{M\times1000}$ and are used for the formation of pathway based data matrix $D_{M\times N}$. The steps of formation of pathway based data matrix $D_{M\times N}$ from pre-processed microarray data matrix $D_{M\times1000}$ are shown in Fig. 3. Thus, for K number of pathways, a data matrix $D_{M\times N}$ is formed. N represents the total number of genes present, combining all the pathways, M is the sample number and $p_1, p_2, \ldots, p_K$ represent the K number of top 50% pathways where each pathway consists of few genes related to it.

3.2.3. Normalization process

In this step, the resultant dataset is normalized using min-max normalization [23] procedure. If for the kth pathway, the expression of the nth gene for the jth sample is represented by the variable g_{jnk}, then the min-max normalization formula for a data point g_{jnk} is described by Eq. (2). Here, $min(g_{nk})$ and $max(g_{nk})$ are the minimum and the maximum value of the nth gene over the samples for kth pathway.

$$g_{jnk}(normalized) = \frac{g_{jnk} - \min(g_{nk})}{\max(g_{nk}) - \min(g_{nk})} \tag{2}$$

where $j = 1, 2, \ldots, M$; $n = 1, 2, \ldots, N$; $k = 1, 2, \ldots, K$.

3.3. Methods for detecting differentially expressed pathway genes

For the two different classes, it is very effective to select the DEGs from pathway based data matrix and they need attention for medical diagnosis. To achieve our goal, we have employed proposed MFDPSO technique. It is very useful for searching the best DEG subset from the pathway based dataset using four different statistical methods. As, proper choice of filter increases the probability of discoveries, we have selected four efficient filtering methods as the fitness function of MFDPSO. The proposed scheme is explained below.

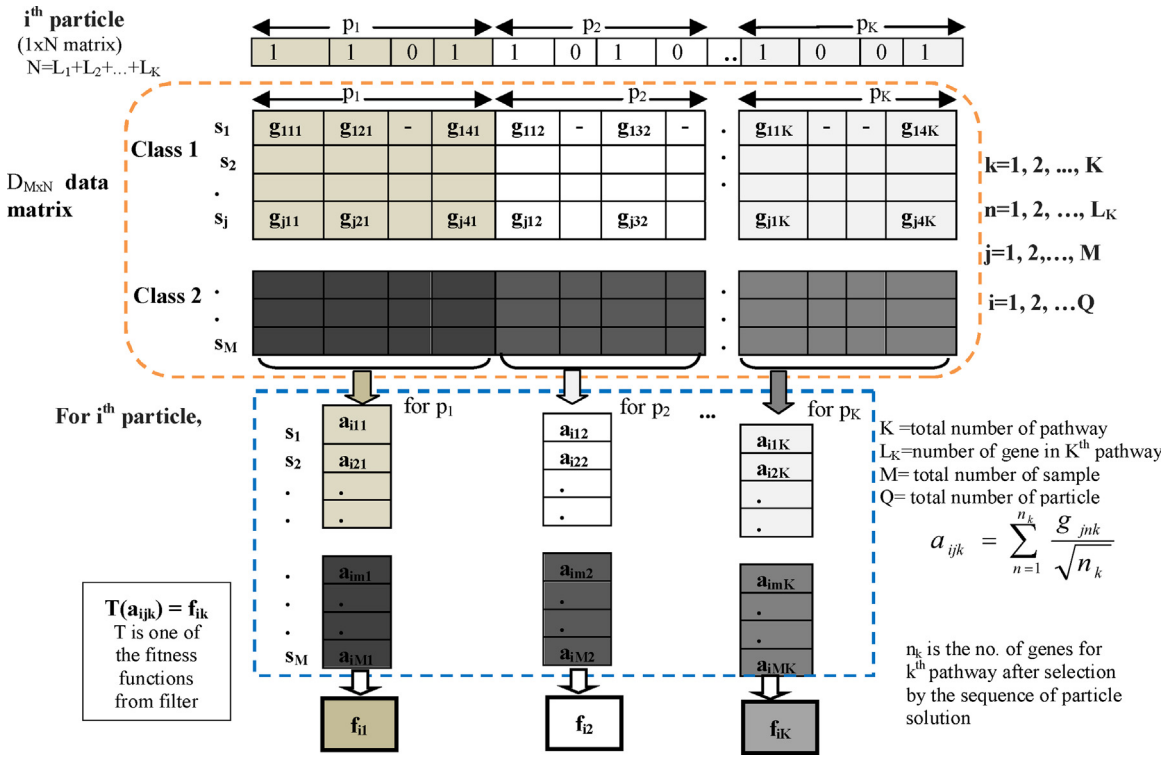

Fig. 4. Inferring pathway activity using MFDPSO technique.

3.3.1. *The fitness functions*

Four different statistical methods are used separately for the calculation of fitness value of the pathways. They are *t*-test, *F*-test, WR test and correlation coefficient. The *t*-test [23] is a popular choice for detecting DEGs in microarray studies which compares the difference between two means in relation to the variation in the data. Correlation coefficient [24] is another statistic that is applied to measure how strong a relationship is made between two variables. Pearson's correlation [24] is a correlation coefficient commonly applied for the microarray data filtering. Whenever, the groups are not normally distributed, especially for small sample sizes, the result of the *t*-test may not be valid. WR test [23]is a very good alternative option as its operation is based on rank-transformed data. The *F*-statistic is also widely used for the feature subset selection from microarray dataset [6]. We have employed these four methods separately as the fitness function of the MFDPSO to accommodate the varied characteristics of the dataset. Thus, after inferring the pathway activity, the particles search for the optimal feature set depending upon the *t*-score, where the aim is to maximize the *t*-score for each and every pathway. Similarly, for the *F*-statistic fitness, we obtain another optimal gene subset by maximizing the metric. For correlation coefficient and WR test based filters, our target is to minimize the score, as the lower value of the score indicates good DEGs. Thus, a subset of informative genes is generated for each filtering method using MFDPSO technique.

3.3.2. *Multi-fitness discrete particle swarm optimization (MFDPSO) technique*

For the selection of an optimal feature subset from a high dimensional space, bio-inspired evolutionary algorithm may be employed. PSO [39], developed by Eberhart and Kennedy, is one of such techniques which is very popular due to its simplicity and ability to find out the optimal solution in limited time. In PSO, particles move in multi dimensional problem space with a velocity which alters with each generation. The velocity, v_i and position, x_i for *i*th particle at *t*th iteration are updated as

$$v_i(t+1) = w.v_i(t) + c_1.r_1(pbest_i - x_i) + c_2.r_2(gbest - x_i) \quad (3)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (4)$$

where w is the inertia weight, c_1 and c_2 are the cognitive and the social learning factor respectively. The particle's best position is represented as *pbest* and swarm's best position is represented as *gbest*. r_1, r_2 are two random numbers. In our work, we use a discrete version of PSO where the position of the particle is considered as a pattern of either 1 or 0. Most of the previous versions of discrete PSO [6,22] uses sigmoid function to obtain the position of the particles. If the velocity is high, due to the nature of sigmoid function, all the position will be oriented towards 1. Here in this paper, we use a simple step function to provide the discrete nature in particles position rather using the probability based on the velocity of the particle. In our work, position is updated in continuous domain so that it can acquire different ranges of values. Subsequently, step function is used for simple binary value assignment, as explained in Eq. (5). The selection depends on a threshold value c which is taken as 0.5 that is the midway between 0 and 1.

$$x_i(t) = \begin{cases} 1 & if\ x_i(t) >= c \\ 0 & if\ x_i(t) < c \end{cases} \quad (5)$$

The selection of gene from the microarray data is itself a voluminous problem. To address the large dimension of the pathway gene expression data for each of the filtering methods as stated above, we have modified the discrete PSO and integrated a new multi-fitness concept. In the previous works, [22,21,3] a single fitness value was assigned against the particle to estimate the improvement. Here, we use multiple fitness value where the fitness of a particle is defined by a fitness set. Fitness set is defined as a collection of fitness values where the number of such fitness values for each particle is equal to the number of pathway level present in the dataset. The particle in the problem consists of the selection instance of genes in each pathway as shown in Fig. 4. It is evident from the figure that rather searching for the best DEG set for *K*

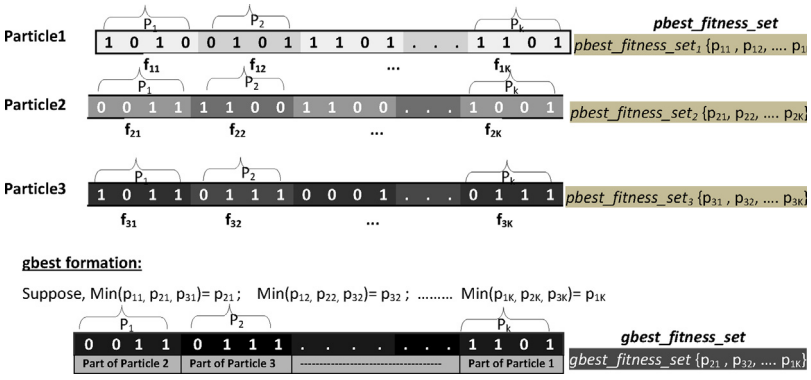

Fig. 5. Selection of best solution using MFDPSO technique.

number of pathways separately one by one, we may construct the problem in such a way that for each pathway a unique fitness is assigned and all the pathways present in the dataset are concurrently processed. This also helps to save the execution time for the analysis of big dataset. Genes corresponding to the value 1 of the particle solution are considered for the computation and then using those genes, the pathway activity of each sample for a certain pathway is calculated. The entire mechanism is described in Fig. 4 for the ith particle. The similar processing is applicable for all Q number of particles. Suppose, at a particular iteration, for kth pathway, n_k genes are selected among L_k genes, by particle solution stream, where L_k is the number of genes present in the kth pathway. Then, the pathway activity for ith particle, a_{ijk} for the jth sample is calculated using Eq. (6), combining all the n_k genes, where g_{jnk} represents the gene expression of the nth gene at jth sample for kth pathway. Thus, after inferring the pathway activity, the fitness of ith particle for the kth pathway, f_{ik}, combining all the M samples, is determined using one among the four filter based fitness functions which are described in the previous subsection. For an example, if the fitness function is t-score, then f_{ik} is calculated by doing t-test on column matrices consisting of a_{ijk} values for two classes.

To determine the ultimate fitness of a particle, we calculate the elements of fitness set, f_{ik} for individual pathways in the same way and as a whole the fitness set for ith particle, defined as, $fitness\_set_i = \{f_{i1}, f_{i2}, f_{i3}, \ldots, f_{iK}\}$ is constructed for K number of pathways, as shown in the last phase of Fig. 4. The scheme provides better fitness value to the swarm compared to other methodology where the use of the average result causes suffering in the computation for pathway activity. Now, at every iteration, member elements of a fitness set are changing due to the searching process. A particular string of solution corresponding to a pathway is updated if the present result is better than the previous one. Thus, we have personal best fitness set for each particle, termed as $pbest\_fitness\_set_i$ for the ith particle described as $\{p_{i1}, p_{i2} \cdots p_{ik} \cdots p_{iK}\}$. For Q number of particles, the overall personal best fitness set is termed as *pbest_fitness_set*, where $pbest\_fitness\_set = [pbest\_fitness\_set_1; pbest\_fitness\_set_2; \cdots pbest\_fitness\_set_Q]$ and it is shown in Fig. 5. For each pathway, based on the minimum value of the fitness member element of the *pbest_fitness_set* of the particles, the best particle of the swarm, *gbest* is formulated and its corresponding fitness is termed as *gbest_fitness_set*. The estimation of *gbest_fitness_set* is explained in Fig. 5 with an example. Let us consider, 3 particles in the problem with the *pbest_fitness_set* as shown in the figure. For the first pathway, the minimum value is calculated among the 3 particles, which in this example, is assumed to be p_{21}. It means that the second particle achieves the best result for the first pathway. Same calculation is performed for each pathway. Now, as shown in the figure, collecting the best solutions for K pathways covering all the particles, the best string solution *gbest* is formed. Thus, *gbest* represents the solution that is the pathway markers of a disease. The overall technique of MFDPSO is described in Algorithm 1. The parameters used for MFDPSO based computation are also given in Table 1.

Table 1
Parameters used for MFDPSO.

Parameters	Explanation	Value
N	Number of particle(s) in one swarm	20
$c1, c2$	Acceleration constants	1.49
w	Inertia	0.7
$r1, r2$	Random numbers	[0,1]
t	Iteration	200

$$a_{ijk} = \sum_{n=1}^{n_k} \frac{g_{jnk}}{\sqrt{n_k}}; \tag{6}$$

$j = 1, 2, \ldots, M$; n_k, no. of selected genes for kth pathway.

Algorithm 1. Description of MFDPSO

Input: Q : *Total population*; K : *Number of pathway level*;
Output: *The optimum solution* (*pathway marker gene set*);
Initialization
for $i = 1$ *to* Q **do**/* *Initialization of a particle* */
Randomly initialize the position(x_i) *and velocity*(v_i) *of ith particle*
Calculate the fitness_set of the particle fitness(x_i) = $\{f_{i1}, f_{i2} \ldots f_{ik} \ldots f_{iK}\}$
Calculate the pbest_fitness_set of the particle, $x_i = \{p_{i1}, p_{i2} \ldots p_{ik} \ldots p_{iK}\}$
end for
Calculate gbest and gbest_fitness_set
Termination check
if *the termination criterion holds* **stop**;**else** *go to next step*
Set $t = 1$/* t = *iteration counter* */
for $i = 1$ *to* Q **do**/* *Updation of a particle* */
Update the position and the velocity according to Eqs. (3) *and* (4)
Assign discrete value in position according to Eq. (5)
for $k = 1$ *to* K **do**/* *Updation of fitness_set for each pathway* */
Select genes corresponding to 1 *value of the particle stream*
for $j = 1$ *to* M **do**/* M *is total number of sample* */
Calculate a_{ijk} *using Eq.* (6) *for the ith particle*
end for
Calculate the element of fitness_set (f_{ik}) *of ith particle*
if f_{ik} *is better than previous* p_{ik}
Update pbest_fitness_set p_{ik};*Update* x_i *corresponding to kth pathway*
end if
end for/* *End of fitness updation for each pathway* */
end for/* *End of updation of a particle* */
for $k = 1$ *to* K **do**/* *Selection of gbest* */
gbest_fitness_set($1, k$) = $min\ f(p_{1k}, p_{2k}, \ldots, p_{Qk})$/* *Minimum of all* */
if *gbest_fitness_set*($1, k$) *is better than previous*
Update gbest ;
end if
end for
Set $t = t + 1$. *Go to step* 4
Solution is the gbest

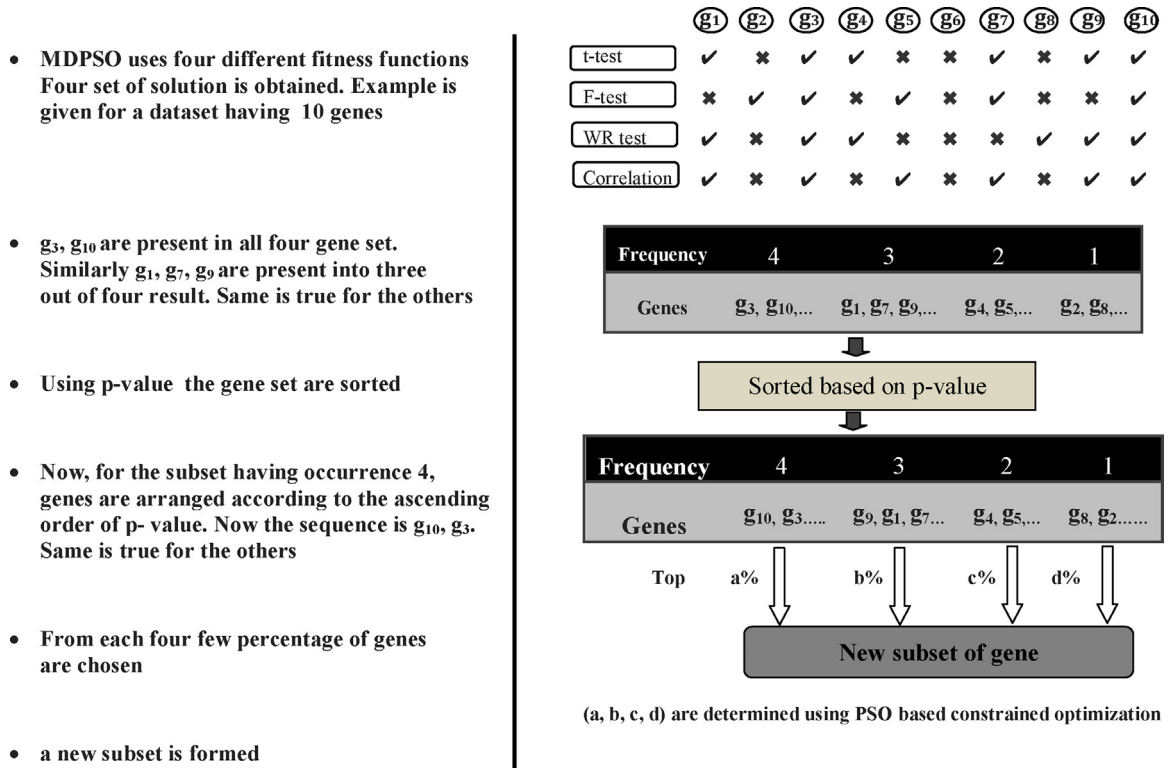

Fig. 6. Description of the generation of new feature set using proposed fusion method.

3.3.3. Fusion for new feature set generation

Using the above mentioned technique, from pathway associated microarray data matrix $D_{M\times N}$, we obtain four different subsets of genes for the four different fitness functions. As each fitness function has its own advantages and disadvantages, so each of the four subsets consists of some useful marker genes for the diagnosis of a particular disease. Now a novel percentage based gene fusion method is developed for the fusion of multiple informative subsets to improve the robustness of the gene selection process and we get a new optimal feature subset. The fusion technique is described in Fig. 6. First, using the four optimal gene subsets, the frequency of occurrence of a particular gene in those four subset is evaluated. Thus, we obtain N_1 number of genes which are selected only by one of the methods, N_2 number of common genes which are selected by any two methods, N_3 number of common genes which are selected by three methods out of the four methods and N_4 number of genes selected by each fitness method. The genes which are not selected by any of the methods are omitted. Now the gene set, corresponding to a particular frequency of occurrence, are sorted according to ascending order of their p value. After sorting and clustering all the genes based on their frequency of occurrence, top a% genes from the subset having frequency four are selected. Similarly, top b% genes having frequency three, top c% genes having frequency two and top d% genes having frequency one are selected. The above percentage based gene selection procedure is adopted based on the motivation as discussed below. Here, in the fusion method, we initially define a fused feature set from the four MFDPSO based feature sets. The number of genes present in the fused set must be less than the total collection of genes obtained from the first stage. Now from this combined gene set, if we select all the genes having frequency four or three then for a large experimental dataset, we may get a substantial number of selected genes from the category of gene set having frequency four or three only. In such a scenario, the genes of frequency two or one may not get the chance of being a member of final reduced feature set and subsequently we may fail to select some unique informative feature represented by genes of frequency category one or two. Therefore, to keep a balance between the gene selection procedure from all categories of frequency, the fusion problem is formulated as a percentage based fusion method and the percentage values are determined in a stochastic manner. The fitness function, F is designed based on p-value of the selected genes where our aim is to minimize the overall p-value of optimized selection of genes. The statement of the hypothesis and the decision rules are given in Eqs. (7) and (8). Here, m_n^1 and m_n^2 represent the mean of expression of nth gene over the samples of the first class and second class respectively. The null hypothesis used for p-value calculation is stated as follows:

The expressions of particular nth *gene in* 1st *and* 2nd *class of the disease are the samples from continuous distributions with equal means, against the alternative that they are not.*

$$Statement \begin{cases} H_0 : m_n^1 = m_n^2 \\ H_1 : m_n^1 \neq m_n^2 \end{cases} \tag{7}$$

$$Decision \begin{cases} H_0 \text{ is true if } p\text{-value} >= 0.05 \\ H_1 \text{ is true if } p\text{-value} < 0.05 \end{cases} \tag{8}$$

The values of a, b, c, d are evaluated using PSO based constrained optimization technique [42]. Here, PSO is randomly searching for a, b, c, d satisfying two constraints as described in Eqs. (10) and (11) and then the fitness value is calculated by Eq. (9). The most fitted particle provides the solution for a, b, c, d. In the first constraint, as the four parameters define selection of some percentage of genes, the summation of a, b, c, d is considered as 100% gene selection. In the second constraint, the maximum selection percentage of genes in one category is kept below a threshold value, T. The selection of threshold value will be always less than 50%. If T is set equal to 50%, then in some scenario, it may happen that the number of genes from one category may be equivalent to the number of total genes from all the other categories. Also, if $T > 50\%$, then any one of

the categories may alone dominate all the other categories of gene selection losing the effect of diversity in the system. Therefore, it is always wise to keep $T < 50\%$. In our experiment, we have kept $T = 40\%$, so that the effect of any category does not dominate other. After following the fusion method stated above, a new fused gene set is evolved.

$$Minimize \quad F = \sum_{f=1}^{4} \sum_{n=1}^{\frac{A(f)\cdot N_f}{100}} P_{value}(G_n) \tag{9}$$

$$Subject\ to: \quad \sum_{f=1}^{4} A(f) = 100\% \tag{10}$$

$$0 < A(f) <= T \ for\ all\ f \quad where,\ matrix\ A = [d, c, b, a] \tag{11}$$

3.3.4. Best pathway marker subset selection

We have now total five feature subsets, among which four are obtained from four MFDPSO based statistical techniques and the fifth one is the fused gene set. Now our job is to select the most effective and informative set among them. Each of the four subsets obtained from four MFDPSO based techniques may be better from one another with respect to some features. So, we cannot definitely say that any one of them is the best. Now, we have another optimal subset produced by the proposed fusion method that is targeted to be the best in most of the cases. But in any random selection scenario, if some of the important genes are missed during the fusion procedure then the resultant fused set may not appear as a winner. However, an efficient selection of the most informative gene plays an important role in the detection of a particular disease. So, instead of blindly considering the fused set as the best option, the entire selection is justified by a new fitness score for those five optimal subsets so that most relevant and worthy feature subset may be identified. The *fitness_s core* is defined in Eq. (12). The proposed fitness function balances the performances of all four statistical techniques as well as it restricts on the number of selected genes by each method. The fitness function *fitness_s core* is constructed as a minimization problem where *score* matrix is the set of average scores obtained using different filtering techniques. In our case, the number of elements in the *score* set is four. For WR test and correlation based fitness, smaller values imply good result, but for *t*-score and *F*-score, higher values indicate more differentially expressed genes. So, we have taken inverse of the values of *t*-score and *F*-score for this minimization problem. Again, the values of the different statistical score are in different scale, so we have used the normalized average score for each statistic. The same weightage value, *w* is given to all these four statistical score for the calculation of *fitness_s core*.

$$fitness\_score = z.(features\_selected) + \sum_{i=1}^{4} w.score(i) \tag{12}$$

$where, score = [t_{score}^{-1}, F_{score}^{-1}, WR_{score}, correlation_{value}]$.

Here, *features_selected* is the number of selected features present in a gene subset, obtained by a certain method. Our aim is to get a good fitness score using fewer features. So, it has an impact on the selection of the optimal gene set among the five. As the number of selected genes in each optimal subset is nearly hundred, so as to keep balance with other scores, the value of *z* is taken as 0.001. After calculating the *fitness_scores* of all five gene sets, the minimum one is selected to define the marker genes. The best fitted gene subset is used for the next stage of computation.

3.4. Class dependent feature selection

In the best pathway marker subset, most relevant genes for each individual pathway are present. Now, we focus on selection of the most informative genes which commonly represent the pathway level of a particular disease as well as those selected genes will be useful for the early detection of diseases. In the pathway information, there are some genes which are common to many pathways that is some genes are the fundamental cause of many pathway diseases. Again, as the functionally co-regulated genes are collectively associated with a pathway disease, most relevant and significant genes are required to be selected to represent those groups. The representative genes are sufficiently competent for the classification. For the purpose, the selected optimal gene subset of the previous stage is now further processed to decrease the number of feature set. Here, we have applied a clustering method on the selected informative marker gene subset to extract the representative genes at the pathway level and to determine the most effective feature set of that class. For the purpose, a novel BLABC algorithm is proposed and used for automatic clustering of the class dependent features, which is explained below.

3.4.1. Blended Laplacian artificial bee colony algorithm

Karaboga has described ABC algorithm based on the foraging behaviour of honey bees for numerical optimization problems [43]. In ABC algorithm three groups of bees: employee bees, onlookers and scouts are involved in the searching process. The number of employee bees is equal to the number of food sources around the hive and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution. An employee bee produces a modification of the position (solution), $x_i = [x_{i1}, x_{i2}, \ldots, x_{iD}]$ and depending on the nectar amount (fitness value) of the new source (new solution), $xnew_i = [xnew_{i1}, xnew_{i2}, \ldots, xnew_{iD}]$, it memorizes the one which has higher nectar amount. After all employee bees complete their search process, they share both the nectar information of the food sources and their position information with the onlooker bees on the dance area. An onlooker bee chooses a food source with a probability, P_i related to its nectar amount that is calculated as follows.

$$P_i = \frac{a * fit_i}{\max(fit)} + b \tag{13}$$

where a, b are constants and fit_i is the fitness value of the ith solution, which is proportional to the nectar amount of the food source in that position. In order to produce ith candidate's new food position, $xnew_i$ from the old position, x_i, in memory, the ABC uses the following expression (14) where x_{qd} is the randomly chosen neighboring point and $d = 1, 2, \ldots, D$ is randomly chosen dimension among D dimensions of the problem. R_{id} is a random number between $[-1, 1]$.

$$xnew_{id} = x_{id} + R_{id}(x_{id} - x_{qd}) \tag{14}$$

The food source which is abandoned by the bees is replaced with a new food source. In ABC, if a position does not improve for a predetermined number of iterations, then that food source is assumed to be abandoned. At this point, the scout discovers a new food source to be replaced with x_{scout}, provided that the new food has equal or better nectar than the old source. For the generation of scouts, we have implemented a blended Laplacian operator [44] which is described using a set of equations as given below. The justification of such modifications is given in subsequent section. First, two new random sources, sol_1 and sol_2 are generated using the best food source ($search_{best}$) discovered so far. Then, using a random coefficient termed as *beta* two new solutions y_1 and y_2 are formed. The new position of the scout, x_{scout}, is a combination of these two new solutions y_1 and y_2 having a weightage factor gamma. The process

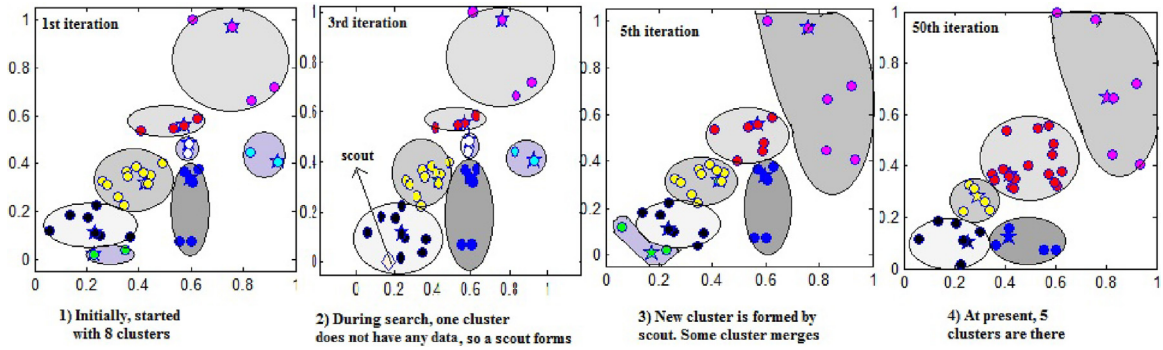

Fig. 7. 2D plot of a portion of DLBCL data using automatic BLABC clustering technique.

of generation of scout provides a good diverse solution to the swarm which in turn helps to enhance the efficiency of the algorithm.

Blended Laplacian operator used for scout generation in BLABC

$/*(search_{best}) = best\ food\ source\ iter = no.\ of\ iteration; D = dimension*/$
$sol_1 = search_{best} .* rand(1, D);$
$sol_2 = search_{best} .* rand(1, D);$
$gamma = 0.1 + (1 - 0.1)^{0.95^{iter}};$
$beta = 0.5 * log(rand(1, 1));$
$y_1 = sol_1 + beta * (sol_1 - sol_2);$
$y_2 = sol_2 + beta * (sol_1 - sol_2);$
$x_{scout} = gamma * y_1 + (1 - gamma) * y_2;$

3.4.2. Feature selection by automatic BLABC clustering

Clustering is the grouping of a set of objects where the objects in the same group (called cluster) are similar (in some sense or another) to each other than those in other groups (clusters). While exploring a small number of significant genes participating in a tumor progression, it is possible to view the problem as the clustering problem. The locations of the centroids of the clusters represent the most significant feature of a particular cancer. For the selection of feature subset of genes, we have to cluster L genes (g_1, g_2, ..., g_l, ..., g_L), where $g_l = (g_{1l}, g_{2l}, \ldots, g_{Ml}) \in R_M$ and M is the dimension of the data object (i.e., sample number). Clustering algorithm tries to find a partition $C = C_1, C_2, \ldots, C_H$ of H clusters, such that the similarity of the patterns in the same cluster is the maximum and patterns from different clusters differ as far as possible. The expressed genes, obtained in the first stage, are considered for clustering into functionally similar subgroups from which the relevant genes are selected for cancer classification. In our proposed clustering method, initially, the number of centroids is kept equal to the number of employee bees. Suppose, 30 feature genes are selected randomly from the dataset which are used for the initialization of the employee bees. Those represent the cluster centers and each data vector is associated with their nearest cluster center. Now, the number of employee bees will change in every iteration as few of the employee bees do not have any food to consume. Accordingly, the number of centroids changes automatically. The ultimate cluster centers symbolize the feature genes. The automatic clustering technique using BLABC algorithm is stated below.

- The input data vectors of microarray dataset act as food for the bees. First, randomly pick up data objects from the dataset for the initialization of the employee bees which represent the initial centroids of clusters.
- During each iteration, each employee bee picks its corresponding foods based on the Euclidean distance [24]. Thus, each data vector is assigned to some employee bee which is nearer to it.
- The fitness related to the position of a bee is determined by the number of foods or the data vectors covered by the bee.
- Depending on the nectar amount, the onlooker bees are involved in searching for the most fitted food source. They use the cluster centroids, discovered by the employee bees, for their searching.
- In each iteration, a new food source is discovered using Eq. (14) where neighboring point towards which the bee will move or not is determined by Eq. (13). When few data vectors are covered by a cluster center, then at each iteration a bee will move towards the neighboring point x_q, where x_q is the average of all the data vectors which is included in a particular cluster. If the fitness of the new source is better than the old one, then the bee memorizes the new source. In the course of searching, some clusters which are having lesser number of data vectors associated with them are needed to be merged. For that purpose, the bees which are having very limited foods associated with them are eliminated from the colony and the data vectors are consumed by the nearest candidate. In our case, we use 5% of total number of data vector as the limited food counter for the elimination of a bee from the colony. After elimination, the number of employee bee is updated.
- After each iteration, the assignment of the data vector to a cluster is updated. At the end of each iteration, if any bee of the existing colony is not consuming any food, then it is considered as the scout. It is re-allocated to a new source randomly using blended Laplacian operator and its fitness value is calculated. This helps in the generation of new clusters and a good amount of diversity can be introduced in the clustering process.
- A generation of the bee consists of few numbers of iterations. Now the data vector which is nearest to the cluster centroids of the generation are considered as the featured gene.
- A generation is considered strong if the cluster centers are well distributed over the entire set of data vectors and performing well for the classification problem. The classification result is used here as the performance indicator of the generation. The generation having the best ever result over all the generations is considered as the solution of the clustering. The genes, nearest to the cluster centers of the best generation are considered as pathway genes which are used for the experimental purpose.

Thus, we are developing a feature gene subset for the two classes separately. Next, the union is carried out of those two class dependent feature subsets. Here, we have to select the most differentially expressed genes which carry useful information for the disease classification. To evaluate the ranking of the selected genes, we have arranged the genes in the ascending order of their p-values and then top 50% genes of the merged set are fed into the classifier. The mechanism of automatic clustering using BLABC is explained in Fig. 7 where we have applied it on a portion of DLBCL dataset [40] and 2D plot of the dataset is shown. Initially, we start with eight clusters and then with iteration weaker clusters are captured by the nearest stronger clusters and new clusters are formed by the scout. At the end, we are getting five optimized clusters. The pseudo code

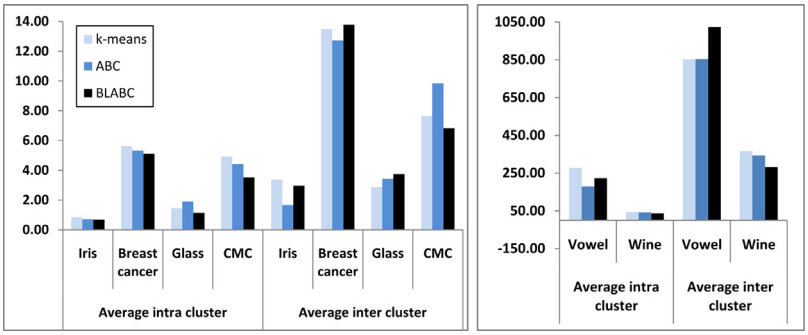

Fig. 8. Comparison of proposed BLABC with other clustering techniques for different benchmark dataset.

Table 2
Parameters used in BLABC algorithm.

Parameters	Explanation	Value
N	Number of bee(s) in one swarm	20
MCN	Maximum cycle number	200

of the proposed feature selection method using BLABC is given in the Algorithm 2. The parameters used for BLABC based computation is also given in Table 2.

Algorithm 2. BLABC based automatic clustering

Input : *Total population* (Q);*Number of generations* (G);*Dataset*(D_{MXL}) ;
Ouput : *The best generation* (*class dependent feature set*) ;
1. **for** $g = 1 : G$/* *generation counter* */
2. *Initialization*
3. *Randomly initialize the position of the employee bees* x_i *from data matrix,* (x_i *is a MX*1 *matrix, where M is the total number of samples*)
4. **for** $i = 1$ *to* Q/* *initialization of bees* */
5. *Custer centre* $C_i = (x_i)$
6. *Calculate Euclidean distance of all the genes from* C_i
7. *Fitness of the cluster centre = No. of data points associated*
8. **endfor**/* *end of initialization* */
9. *Termination check*
10. **if** *the termination criterion holds* **stop**;**else** *go to step* 11
11. *Set* $t = 1$/* t = *iteration counter* */
12. **for** $i = 1$ *to* Q/* *Updation of employee bee* */
13. *Generate neighboring point = mean center of all the data associated with* C_i
14. *Produce new solutions,* $xnew_i$ *for the employee bees by using Eq.* (14)
15. *Evaluate fitness of* $xnew_i$
16. *Apply the greedy selection process and update the solution* x_i
17. *Calculate the probability values* P_i *for the solutions* x_i *by Eq.* (13)
18. **endfor**/* *End of updation of employee bee* */
19. **for** $i = 1$ *to* Q/* *Generation of of onlooker bee* */
20. *Produce new solution* $xnew_i$ *for the onlookers from* x_i *using Eq.* (14)
21. *Depending on* P_i *and evaluate the fitness of the onlooker bee*
22. *Apply the greedy selection process and update*
23. **endfor**/* *End of updation of onlooker bee* */
24. *Determine the abandoned solution*/* *Identification of scout bee* */
25. *Generate new center for scout using blended laplacian operator*
26. *Evaluate the fitness of new scout bee*
27. *Set* $t = t + 1$.
28. *go to step* 9
29. *Store the cluster centre position* C_i *in an archive*
30. **end for**
31 *Select the best generation*

3.5. Verification of BLABC as clustering algorithm

In this section, we have established the importance of BLABC as clustering algorithm and validated the result for two cases through comparison with other techniques. Initially, we chose the widely used UCI machine learning datasets [25] which are freely available to be used for testing and validation. Some of the common datasets which are used in the study are listed in Table 3.

We select two important performance comparison measures and they are inter-cluster distance (separation of clusters), and intra-cluster distance. Low intra-cluster distance is better than high intra-cluster distance and vice-versa is true for inter-cluster distance. The comparative result with k-means [45] and ABC clustering technique is given using a bar chart in Fig. 8 and in Table 4. From the result, it can be stated that BLABC is useful to obtain compact cluster as it has gained the best intra-cluster distances for five out of six datasets. The result establishes its effectiveness to solve the clustering problems.

Table 3
Description of commonly used benchmark clustering datasets [25].

Name of dataset	No. of classes	No. of features	No. of observations
Iris	3	4	150
Breast cancer	2	9	699
Vowel	6	3	871
Wine	3	13	178
CMC	3	9	1473
Glass	6	9	75

As seen above, the use of blended operator for the generation of scout provides better quality of clusters than compared to the scout generation mechanism used in the general ABC algorithm. The proposed technique is further justified by applying it on few more additional datasets which are used in next experimental verification section. Here, DLBCL [40], Leukemia [29] and Colon cancer [26] datasets are used. The significance of the algorithm is established by comparing the average intra-cluster distances of the clusters with the general ABC algorithm used for the automatic clustering approach. In Fig. 9, bar charts explain the average intra-cluster distance for the same number of clusters generated by BLABC and ABC based automatic clustering. For BLABC, relatively smaller intra-cluster distances claim its superiority of generating compact clusters by enriching it with a well-located centroid.

4. Experimental results on cancer dataset

4.1. Dataset

Six microarray cancer datasets from public domain, are used for the experimental purpose. They are as follows: the biomedical databases of Leukemia [29], Colon cancer [26], Gastric cancer [28], DLBCL [40], Child_ALL [41], and, Prostate cancer [27]. A brief details of these microarray datasets are given below.

Leukemia [29]: Leukemia is one type of blood cancer, which occurs due to the abnormal production of white blood cells. Here, we use two different classes of Leukemia named acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL) type having 25 samples and 47 samples respectively. The dataset consists of 72 microarray experiments with 5147 gene expression levels. Tissue samples were collected from bone marrow (62 cases) and peripheral blood samples (10 cases).

Table 4
Comparison of average intra-cluster and inter-cluster distance for data clustering.

Dataset	k-means		ABC		BLABC	
	Avg intra-cluster distance	Avg inter-cluster distance	Avg intra-cluster distance	Avg inter-cluster distance	Avg intra-cluster distance	Avg inter-cluster distance
Iris	0.85	**3.37**	0.72	1.66	**0.68**	2.96
Breast cancer	5.63	13.48	5.32	12.73	**5.12**	**13.79**
Vowel	276.75	852.83	**179.27**	854.74	223.17	**1023.90**
Wine	43.46	**366.30**	40.81	343.84	**37.36**	282.30
CMC	4.92	7.64	4.41	**9.84**	**3.52**	6.83
Glass	1.44	2.87	1.91	3.43	**1.14**	**3.74**

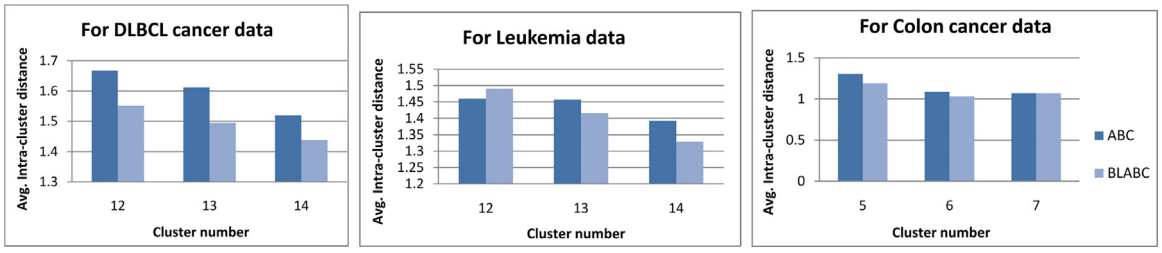

Fig. 9. Comparison of BLABC and ABC clustering technique for different cancer dataset.

Colon cancer [26]: The Colon cancer dataset contains in total 62 cell samples, among which 40 biopsies are taken from healthy parts of the colon and 22 biopsies samples are taken from tumors of the same patients. The dataset includes expressions of 6000 genes obtained from cancerous and the normal cell.

Gastric Cancer (GSE2685) [28]: Gastric cancer occurs due to the growth of cancerous cells in the lining of the stomach. This experimental dataset is constructed for 4522 genes from the total 30 number of tissue samples. The dataset is produced by combining the diffuse and intestinal advanced Gastric tumor samples into one class (22 samples) and noncancerous samples into another class (8 samples).

DLBCL [40]: DLBCL is a fast growing lymphoma that develops from the B-cells. Two types of sample are used and they are diffuse large B-cell lymphoma (DLBCL) type and follicular lymphoma (FL) type. The dataset has total 7070 genes and 58 samples of DLBCL type and 19 samples of FL type.

Prostate [27]: The dataset is collected from the normal and the cancerous tissues of Prostate. It consists of gene expression measurements of 50 non-cancerous prostate tissues and 52 prostate tumors. The expression matrix is formed of 12533 numbers of genes.

Child_ALL(GSE412) [41]: The dataset Childhood ALL is related to the cancer of childhood, which includes 110 samples of childhood acute lymphoblastic leukemia. Among them, 50 examples are collected before and 60 examples are collected after therapy. The samples are having expression level of 8280 genes.

4.2. Experimental background

The collected microarray dataset is first pre-processed and then top 50% KEGG pathway information is inferred to obtain a pathway based dataset. Next, the resultant dataset is normalized and the proposed methodology is applied to form an optimal subset of pathway level genes. For the experimental purpose, the dataset is partitioned into training and test set to ensure that both the sets contain at least one sample from each class. Here, following the general convention of 10-fold cross validation, we partition the whole dataset into ten partitions in a random fashion. Nine collectively form the training set and the remaining one partition is used for the testing purpose. Likewise, we design all the experiments for 10 fold cross-validation to evaluate the performance of our methodology. The experiment is repeated for 10 times and the average result is calculated to infer the efficiency of the scheme. In the result section, first we analyze and discuss the performance of MFDPSO and BLABC for the process of marker selection. Next to evaluate the classification model, five different popular classifiers are considered such as support vector machine (SVM), C4.5 classifier, decision tree (DT), K-nearest neighbour (KNN) classifier and Naive Bayes (NB) classifier which are widely used in different literatures [31,30,5,35]. Again, the result is compared with other existing methods reported in different research articles and also with other techniques used for pathway selection processes, such as BPSO [22], Mean, Median [19], LLR [21], CORG [3] methods, etc. From the experimental evaluation, pathway markers are identified which are further validated by means of their biological significance. The entire experiment is carried out in Intel Core i3, 1.87 GHz processor.

The evaluation of the result is conducted in terms of fitness value and classification results. The predictions for testing samples in a class are compared with the original clustering result to figure out the number of false positives (fp), true negatives (tn), false negatives (fn) and true positives (tp). The evaluation of classification is reported in terms of sensitivity, specificity, accuracy, *F*-score and AUC (area under curve). Sensitivity describes how a disease is properly identified in the positive class and the term specificity indicates the efficient detection of sample which belongs to negative class. Precision and accuracy are also indicators of the good classification of the disease. Another parameter, AUC represents the accuracy of the classification where the value 1 of AUC indicates the perfect result of the classification.

4.3. Result and discussion

4.3.1. Analysis of experiments

MFDPSO involves four different filtering functions as the fitness objectives. The normalized optimum fitness values obtained by MFDPSO using different filtering based optimum function are plotted with the number of iterations in Fig 10 . Since, the problem is constructed as the minimization problem for all the methodologies of fitness computation, with the increase in iteration number the fitness value of MFDPSO decreases. Overall, it is to be noted that within 100 iterations MFDPSO is getting into stable region and the normalized scores obtained by the *t*-test, *F*-score, WR test and correlation based filtering techniques can be estimated from the

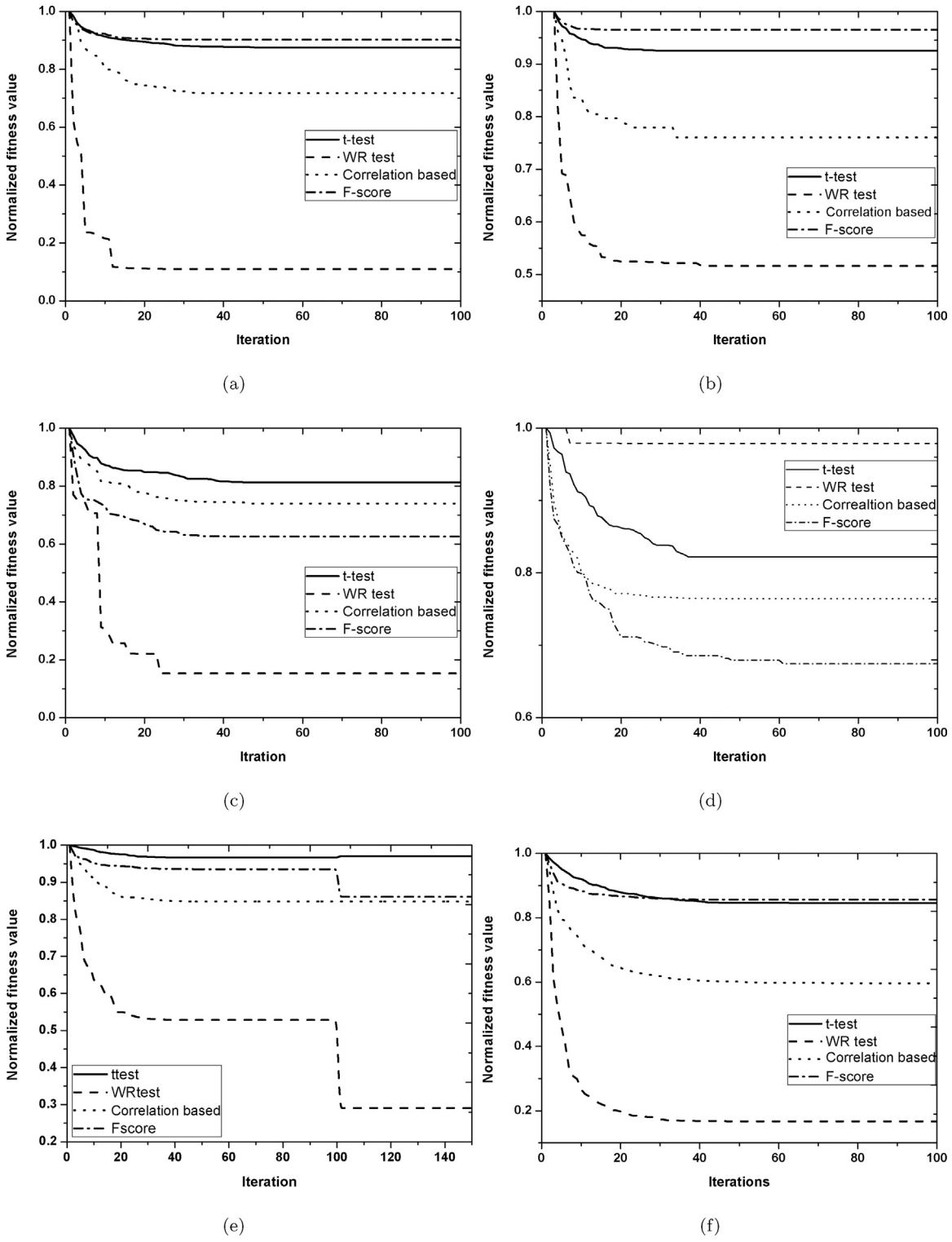

Fig. 10. Normalized fitness value vs iteration plot for (a) leukemia, (b) colon, (c) prostate, (d) gastric, (e) Child_ALL, (f) DLBCL cancer dataset.

graphs. For the Leukemia, Prostate and DLBCL cancer dataset, the normalized value of the WR test is minimized significantly at higher number of iterations compared to other datasets. This signifies the efficient working of WR test based fitness function compared to other techniques. But for Gastric cancer, the working of the WR-test is not very promising. *F*-stat works better to obtain differentially expressed genes for the Gastric microarray data as well as for the Prostate cancer. For DLBCL and Colon data, the performance of the correlation based filter is also very effective. We can see from the graphs, filters work differently for different cancer due to the varied information content in the values of the data. So, using one technique, we may not justify all the datasets and that motivates us to apply different measures for extracting the information. Thus, the four techniques are working independently and produce four different gene subsets. As stated above, the fitness scores of those generated four gene subsets are calculated using Eq. (12) and a new

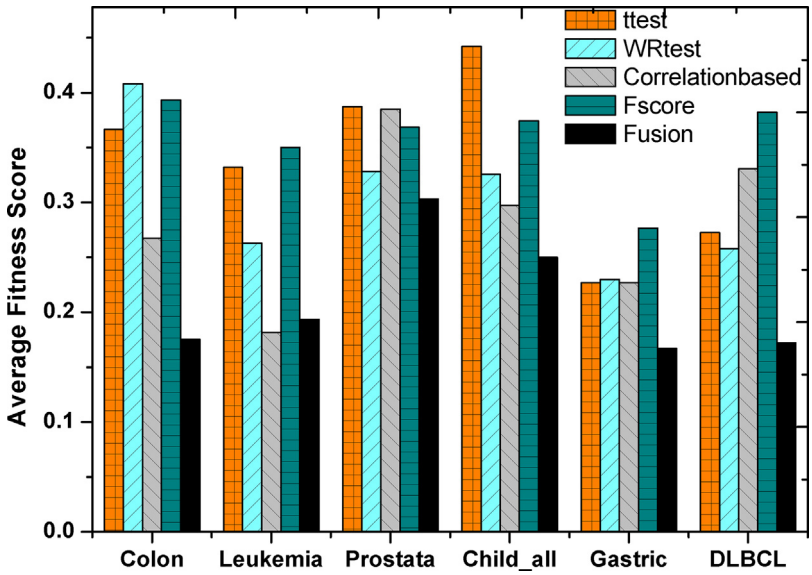

Fig. 11. Average fitness score of gene subsets obtained after MFDPSO and fusion method for different cancer datasets.

gene subset is produced by the fusion method, described in Section 3.3.4. The average fitness score of all these five informative gene subsets for 30 independent runs of the MFDPSO is shown in Fig. 11. The experimental results are shown for all the six cancer datasets in the figure which clearly shows that, except for the Leukemia, our proposed fusion based method is able to produce better informative gene subset for all the cases. For the Leukemia data, the gene subset generated by the correlation based technique is carrying more information. The minimum fitted gene subset which carries the significant information about the disease is used for the next step of the process.

In the next stage, automatic BLABC clustering algorithm is used for clustering of class dependent features and from the selected features, top differentially expressed genes (DEGs) are used for the classification. In order to analyze the performance of our methodology, we have observed the effect of feature size and also the effect of the selected pathway markers on the classification of the disease. The results of the classification analysis are given into subsequent section.

4.3.2. Classification results

In the first section of the result analysis, it is observed that the selection of top DEGs among the combined gene set is very vital as the feature size has huge impact on classification by different classifiers. As the number of features increases, it may cause overfitting. Again for some dataset, higher number of features produces good classification accuracy. We here design the experiment for different percentages of top selected genes using five different classifiers, like SVM, C4.5, DT, KNN, and NB classifiers [46]. The accuracy of the classification for 10 times 10-fold cross validation is plotted for different percentage of selected gene sets and different datasets in Figs. 12 and 13. For Child_ALL dataset, classification with lower number of gene set produces more significant results when SVM, DT, KNN, NB classifiers are used. The highest accuracy is obtained by KNN classifier which is equal to 87.6%. For DLBCL, use of higher number of feature genes produces better results when classification is performed using SVM, KNN and NB classifiers. Here, 100% accuracy is achieved by KNN classifier. For Gastric cancer, lesser number of feature size is able to generate 100% accuracy with SVM whereas, other classifiers obtain 100% accuracy for a relatively higher number of features. For Prostate cancer, 92.8% accuracy is reached using 60 percent of top features. For the Leukemia, SVM achieves 99% accuracy using lesser number of features. C4.5 and DT classifiers obtain the highest accuracy of 91% and 93.4% respectively using top 50–60% of the total feature set. For Colon cancer, we have achieved 79.6% accuracy using top 50% of the total class dependent feature set. C4.5 classifier works well on higher number of feature set, whereas the performance of KNN and NB classifier improves for lesser number of features.

In the second section of the analysis, we observe the effect of selected marker on classification on the disease. From the class dependent combined feature set, top 50% pathway genes are selected for the classification. Here, we have designed the experiment for 10 times 10-fold cross validation. The proposed pathway biomarker selection methodology is implemented using five different well known classifiers (SVM, C4.5, DT, KNN, NB) in the classification stage. Average results of classification for the above mentioned dataset are given in terms of sensitivity, specificity, accuracy, *F*-score, precision and AUC in Table 5. For Child_ALL dataset, with top 50% pathway feature genes, classification with SVM classifier is providing highest accuracy of 83.15% compared to other classifiers. But, KNN is able to achieve better specificity and precision than the SVM. The same is true for the DLBCL dataset, where the highest accuracy is observed using SVM classifier having value 93.52% and KNN is providing good specificity. In the experiment with Gastric cancer dataset, classification result using KNN classifier is the best when compared with others, but we get 100% specificity using SVM and NB classifiers. SVM classifies the dataset of Prostate cancer more efficiently with an accuracy of 90.71%. For Leukemia dataset, on an average KNN classifier is the best performer giving an accuracy of 95.37% and for Colon cancer the highest accuracy of 79.68% is observed with SVM classifier.

4.3.3. Comparative results

To estimate the effectiveness of the proposed MFDPSO-BLABC based method, experiments are conducted and results are compared with other approaches, used earlier for pathway marker selection. A comparative study is reported in Table 6 where BPSO [22], Mean, Median [19], LLR [21], and CORG [3] methods are used to compare the performance. For all the results in the table during the classification stage, SVM classifier is used for 10 times 10 fold cross validation. The average and the standard deviation of the results are reported. For Child_ALL dataset the proposed methodology is working most efficiently, achieving a good result of the

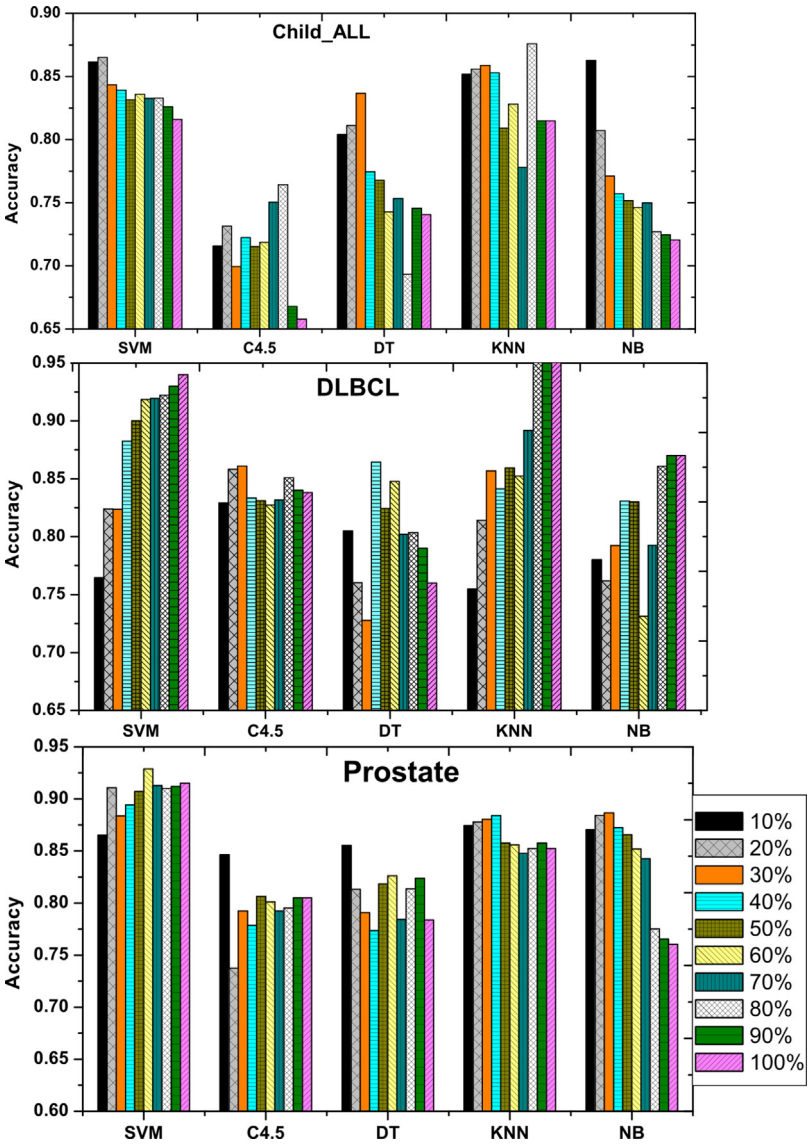

Fig. 12. Accuracy of classification for different feature size using five different classifiers for (i) Child_ALL; (ii) DLBCL; (iii) prostate cancer dataset.

classification compared to all other methods. For DLBCL, it outperforms in terms of accuracy and *F*-score of the classification analysis. Also for the Gastric cancer, the value of the specificity is 1. For Prostate and Colon cancer, the MFDPSO-BLABC obtains the highest score in all evaluation measure and proves its efficiency compared to all other pathway approaches. For Leukemia cancer, the accuracy of the classification is the highest for MFDPSO-BLABC among all the methodologies used in the comparative study. Hence, the effectiveness of the proposed methodology is justified.

The order of complexity of any algorithm describes its efficiency. The order of complexity in terms of *O* notation, for the process of initialization, evaluation and update of the algorithm along with other pathway based approaches are shown in Table 7 where *N*, *cof* and *iter* represent the population size, corresponding

Table 5
Result of classification using different classifiers for different cancer datasets.

Dataset	Methodology	Sensitivity	Specificity	Accuracy	*F*-score	Precision	AUC
Child_ALL	Proposed+SVM	**0.8933**	0.7847	**0.8315**	**0.8448**	0.8153	**0.8390**
	Proposed+C4.5	0.7150	0.7651	0.7153	0.7414	0.8075	0.7401
	Proposed+DT	0.7850	0.7801	0.7678	0.7879	0.8152	0.7826
	Proposed+KNN	0.8433	**0.7895**	0.8091	0.8251	**0.8225**	0.8164
	Proposed+NB	0.7633	0.7844	0.7516	0.7740	0.7746	–
DLBCL	Proposed+SVM	**0.9433**	0.8262	**0.9001**	**0.9352**	0.9403	**0.8852**
	Proposed+C4.5	0.8963	0.6500	0.8261	0.8841	0.8837	0.7746
	Proposed+DT	0.8727	0.6167	0.8043	0.8742	0.8975	0.7567
	Proposed+KNN	0.8877	**0.8542**	0.8594	0.9110	**0.9534**	0.8771
	Proposed+NB	0.8620	0.8500	0.8300	0.8950	0.8283	–

Table 5 (*Continued*)

Dataset	Methodology	Sensitivity	Specificity	Accuracy	*F*-score	Precision	AUC
Gastric	Proposed+SVM	0.6700	**1.0000**	0.8920	0.8773	0.9467	0.8350
	Proposed+C4.5	0.9000	0.9450	0.9270	0.9181	0.9297	0.9225
	Proposed+DT	0.8500	0.9710	0.9320	0.9173	0.9567	0.9105
	Proposed+KNN	**1.0000**	0.9960	**0.9971**	**0.9960**	**0.9933**	**0.9980**
	Proposed+NB	0.9000	**1.0000**	0.9723	0.9810	0.9500	–
Prostrate	Proposed+SVM	**0.9200**	**0.9078**	**0.9071**	**0.9052**	**0.9043**	**0.9139**
	Proposed+C4.5	0.8080	0.8338	0.8064	0.8016	0.8276	0.8209
	Proposed+DT	0.8160	0.8466	0.8183	0.8145	0.8400	0.8313
	Proposed+KNN	0.8640	0.8721	0.8576	0.8541	0.8671	0.8680
	Proposed+NB	0.8660	0.8852	0.8655	0.8601	0.8756	–
Leukemia	Proposed+SVM	0.9565	0.9688	0.9449	0.9595	0.9727	**0.9685**
	Proposed+C4.5	0.9090	**0.9833**	0.8471	0.8922	0.8916	0.9091
	Proposed+DT	0.8815	0.8167	0.8386	0.8861	0.9048	0.9403
	Proposed+KNN	**0.9630**	0.9542	**0.9537**	**0.9664**	**0.9763**	0.9586
	Proposed+NB	0.9445	0.9542	0.9411	0.9540	0.9493	–
Colon	Proposed+SVM	**0.7883**	0.7931	**0.7968**	0.7205	**0.8043**	0.6571
	Proposed+C4.5	0.7035	0.6116	0.6475	0.6475	0.7859	0.6337
	Proposed+DT	0.6964	**0.7932**	0.7110	0.7110	0.7451	0.6753
	Proposed+KNN	0.6428	0.7218	0.6574	0.6574	0.7769	**0.6916**
	Proposed+NB	0.5964	0.6383	0.4873	**0.7390**	0.5691	–

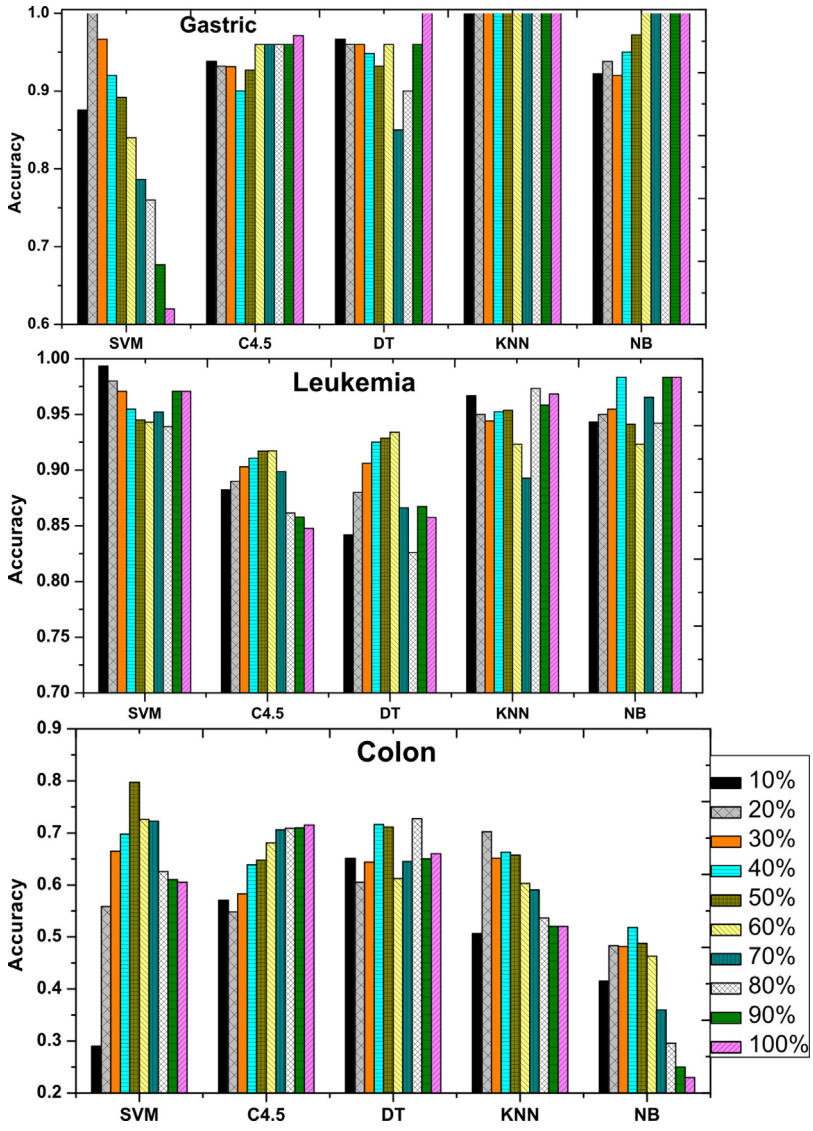

Fig. 13. Accuracy of classification for different feature size using five different classifiers for (i) gastric; (ii) leukemia; (iii) colon cancer dataset.

Table 6
Comparison of mean (standard deviation) of classification result with other pathway based techniques.

Dataset	Methods	Sensitivity	Specificity	Accuracy	*F*-score
Child_ALL	Proposed	**0.89** (1.7E−02)	0.78 (1.8E−02)	**0.83** (1.5E−02)	**0.84** (1.5E−02)
	BPSO	0.76 (1.1E−02)	0.73 (2.7E−02)	0.75 (1.3E−02)	0.73 (1.4E−02)
	Mean	0.72 (1.9E−02)	0.79 (2.4E−02)	0.74 (1.8E−02)	0.77 (1.5E−02)
	Median	0.73 (1.8E−02)	**0.85** (1.7E−02)	0.76 (1.3E−02)	0.78 (1.2E−02)
	LLR	0.88 (1.9E−02)	0.75 (4.0E−02)	0.80 (1.6E−02)	0.78 (1.5E−02)
	CORG	0.68 (7.2E−03)	0.82 (3.4E−02)	0.72 (1.5E−02)	0.75 (2.4E−02)
DLBCL	Proposed	0.94 (1.4E−02)	0.83 (1.3E−02)	**0.90** (2.0E−03)	**0.94** (1.1E−02)
	BPSO	0.37 (2.3E−02)	**0.99** (1.1E−03)	0.83 (1.8E−02)	0.48 (1.7E−02)
	Mean	0.98 (0.0E−00)	0.73 (0.0E−00)	0.88 (1.1E−03)	0.92 (1.1E−03)
	Median	**1.00** (0.0E−00)	0.53 (3.9E−02)	0.82 (1.6E−02)	0.87 (1.1E−02)
	LLR	0.98 (1.3E−02)	0.69 (6.6E−02)	0.87 (1.7E−02)	0.90 (1.1E−02)
	CORG	0.98 (1.4E−02)	0.55 (3.9E−02)	0.82 (1.6E−02)	0.87 (1.2E−02)
Prostate	Proposed	**0.92** (2.4E−02)	**0.91** (7.3E−03)	**0.91** (1.2E−02)	**0.91** (1.3E−02)
	BPSO	0.88 (2.9E−02)	0.90 (1.8E−02)	0.89 (2.9E−02)	0.89 (3.0E−02)
	Mean	0.88 (4.4E−02)	0.88 (9.6E−03)	0.87 (2.9E−02)	0.87 (2.6E−02)
	Median	0.90 (4.7E−02)	0.88 (1.9E−02)	0.90 (3.1E−02)	0.90 (3.1E−02)
	LLR	0.82 (1.6E−02)	0.87 (1.2E−02)	0.84 (1.2E−02)	0.83 (1.3E−02)
	CORG	0.74 (6.5E−02)	0.90 (1.9E−02)	0.76 (5.1E−02)	0.76 (4.3E−02)
Leukemia	Proposed	0.96 (1.5E−02)	**0.97** (1.1E−02)	**0.95** (1.2E−02)	**0.96** (1.2E−02)
	BPSO	0.90 (2.1E−02)	0.95 (1.1E−02)	0.94 (1.6E−02)	0.94 (1.9E−02)
	Mean	**1.00** (0.0E−00)	0.50 (0.0E−00)	0.74 (5.2E−05)	0.79 (8.4E−04)
	Median	**1.00** (0.0E−00)	0.50 (0.0E−00)	0.74 (2.1E−05)	0.79 (3.6E−04)
	LLR	**1.00** (0.0E−00)	0.56 (2.1E−02)	0.78 (1.3E−02)	0.83 (1.1E−02)
	CORG	**1.00** (0.0E−00)	0.50 (0.0E−00)	0.74(2.7E−05)	0.79 (4.7E−04)
Gastric	Proposed	0.67 (1.5E−02)	**1.00** (0.0E−00)	0.84 (3.3E−03)	0.88 (4.9E−03)
	BPSO	**0.95** (0.0E−00)	1.00 (0.0E−00)	**0.97** (6.1E−04)	**0.98** (2.1E−03)
	Mean	0.20 (0.0E−00)	1.00 (0.0E−00)	0.63 (4.5E−04)	0.64 (1.4E−03)
	Median	0.20 (0.0E−00)	1.00 (0.0E−00)	0.63 (5.6E−03)	0.64 (0.0E−00)
	LLR	0.20 (0.0E−00)	1.00 (0.0E−00)	0.73 (5.3E−04)	0.56 (1.9E−03)
	CORG	0.40 (0.0E−00)	1.00 (0.0E−00)	0.63 (5.1E−04)	0.61 (1.8E−03)
Colon	Proposed	**0.79** (2.4E−02)	**0.79** (1.7E−02)	**0.80** (2.3E−02)	**0.72** (1.3E−02)
	BPSO	0.76 (2.6E−02)	0.72 (1.7E−02)	0.74 (2.3E−02)	**0.72** (1.6E−02)
	Mean	0.75 (6.2E−02)	0.35 (2.1E−02)	0.60 (2.6E−02)	0.70 (1.7E−02)
	Median	0.73 (4.5E−02)	0.49 (2.8E−02)	0.66 (2.8E−02)	0.69 (3.4E−02)
	LLR	0.75 (6.2E−02)	0.52 (5.7E−02)	0.70 (4.6E−02)	0.71 (3.3E−02)
	CORG	0.63 (4.4E−02)	0.53 (3.4E−02)	0.69 (3.3E−02)	0.67 (2.3E−02)

Table 7
Comparison of MFDPSO-BLABC with respect to order of complexity.

Methods	Initialization	Evaluate	Update	Overall
MFDPSO-BLABC	$O(ND)$	$Cof * N * iter * k$	$O(ND * iter)$	$O(D * FEs + Cof * FEs * k)$
MFDPSO	$O(ND)$	$Cof * N * iter * k$	$O(ND * iter)$	$O(D * FEs + Cof * FEs * k)$
BLABC	$O(ND)$	$Cof * N * iter$	$O(ND * iter)$	$O(D * FEs + Cof * FEs)$
BPSO [22]	$O(ND)$	$Cof * N * iter * k$	$O(ND * iter)$	$O(D * FEs + Cof * FEs * k)$
Mean [19]	–	Cof	–	$O(Cof)$
Median [19]	–	Cof	–	$O(Cof)$
LLR [21]	–	$Cof * k$	–	$O(Cof * k)$
CORG [3]	–	$Cof * k$	–	$O(Cof * k)$

D = dimension of the problem, k = number of pathways, $N * iter$ = FEs.

Table 8
Comparison of accuracy of classification with other results reported in the literatures.

[Ref] (year)	Methods	Leukemia	Colon	Gastric	DLBCL	Prostate	Child_ALL
[34] (2004)	Uncorrelated discriminant	0.97	0.85	–	–	–	–
[33] (2007)	Partial least square method	0.97	0.83		0.93	–	–
[47] (2007)	ensemble neural network	0.96	0.87	–	0.93	–	–
[48] (2011)	Recursive clustering	0.71	0.8	–	0.82	–	–
[49] (2014)	Kernelized fuzzy rough set/SVM	0.97	–	–	0.93	–	–
[38] (2014)	Multiobjective PSO/SVM	–	–	–	0.93	0.92	0.81
[22] (2015)	Binary PSO/SVM	–	–	0.96	0.82	0.89	0.74
[8] (2016)	GA/Tabu search/ SVM	0.99	0.90	–	0.93	–	–
[16] (2016)	Binary DE/SVM	0.82	0.75	–	0.93	–	–
[15] (2016)	Pipelining method	0.95	0.68	–	–	–	–
[17] (2017)	Information gain/SGA	0.97	0.85	–	0.94	–	–
Proposed	MFDPSO-BLABC/SVM	0.99	0.79	0.97	0.94	0.93	0.86

Table 9
Biological significance of the selected genes.

Dataset	Diseases	Related genes (Pubmed citation number)	
Child_ALL	Hemic & Lymphatic diseases		CXCR2(11), SMAD3(6), PLK1(17), ITGA6(1), IKBKB(10), CREBBP(23), CHUK(4), CDK1(4)
	Cardiovascular diseases		CXCR2(10), ITGA6(1), SMAD3(28), PLK1(2), IKBKB(9), CREBBP(4), CHUK(7), CDK1(4)
	Neoplasm	Acute Leukemia	ITGA6(1), SMAD3(2), PLK1(4), CREBBP(34), CDK1(4)
		Neoplasm Metastasis	CXCR2(17), ITGA6(9), SMAD3(13), CHUK(2), CDK1(4)
		Carcinogenesin	CXCR2(5), ITGA6(1), SMAD3(20), PLK1(15), CREBBO(6), CHUK(3)
		Myeloid Leukemia	PLK1(7), CREBBP(15), CDK1(8), SMAD3(1), CXCR2(1), CHUK(1), ITGA6(1)
	Tumor progression		PLK1(9), CREBBP(4), CHUK(3), SMAD3(1), ITGA6(1)
DLBCL	Hemic & Lymphatic diseases	Anemia	FASLG(2), AKT1(2), HLA-DQB1(7), TP53(32)
		Lymphoma	FASLG(17), AKT1(16), HLA-DQB1(15), TP53(271), CRKL(1), HRAS(3), MAKP3(5)
		T-cell Lymphoma	FASLG(2), AKT1(9), HLA-DQB1(2), TP53(54), HRAS(1), HDAC1(1)
		B-cell lymphoma	FASLG(1), AKT1(11), TP53(54), MAPK3(2)
	Cardiovascular diseases		FASLG(7), AKT1(35), HLA-DQB1(35), TP53(181), CRKL(3), HRAS(14), MAKP3(21), HDAC1(7)
	Neoplasm	Leukemia	FASLG(9), AKT1(24), HLA-DQB1(6), TP53(121), CRKL(2), HRAS(4), MAKP3(3), HDAC1(3)
		Acute Leukemia	FASLG(2), AKT1(34), HLA-DQB1(2), TP53(190), CRKL(1), HRAS(1), MAKP3(3), HDAC1(4)
		Neoplasm Metastasis	FASLG(5), AKT1(65), MAPK3(20), CRKL(3), HRAS(22), HDAC1(3)
		Carcinogenesin	FASLG(5), AKT1(116), HLA-DQB1(2), TP53(1233), CRKL(2), HRAS(64), MAKP3(7), HDAC1(6)
		Myeloid Leukemia	FASLG(4), AKT1(2), HLA-DQB1(4), TP53(21), CRKL(3), HRAS(5), HDAC1(1)
	Tumor growth & progression		FASLG(4), AKT1(21), HLA-DQB1(1), TP53(356), CRKL(3), HRAS(5), HDAC1(3), MAPK3(11)
Leukemia	Hemic & Lymphatic diseases	Anemia	CDN6(3), MSH6(2), ITGB2(1)
		Lymphoma	CCND3(7), LYN(4), RAG1(7), IL7R(1), MSH(5), SPI1(1)ITGB2(6)
		T-cell Lymphoma	CCND3(3), LYN(1), RAG1(1), IL7R(7), MSH4), ITGB2(8)
		DLBCL	CCND3(4), CD36(1), RAG1(1), IL7R(1), MSH(4)
		B-cell lymphoma	CCND3(5), LYN(1), RAG1(7), SPI1(2), CTSG(1), ITGB2(1)
	Cardiovascular diseases		CCND3(9), LYN(1), CTSG(7), RAG1(5), IL7R(3), CD36(151), ITGB2(15)
	Neoplasm	Leukemia	CCND3(2), LYN(1), RAG1(3), IL7R(5), MSH(61), SPI1(7), CD36(2), ITGB2(3), CTSG(3)
		Acute Leukemia	CCND3(1), LYN(1), RAG1(5), IL7R(3), SPI1(26), CD36(2), ITGB2(8), CTSG(8)
		Neoplasm Metastasis	CTSD(15), CD36(3)
		Carcinogenesis	CCND3(6), RAG1(2), CTSD(4), CD36(3), MSH(6)
	Tumor growth & progression		CTSD(4), MSH6(3), SPI1(1)
Colon	Digestive System disease	Cholestasis	VCAM1(1), RARA(1), EGR1(2)
		Colitis	VCAM1(4), RHOA(1)
	Neoplasm	Colorectal Cancer	RHOA(1), VCAM1(5), MAPK3(11), EGR1(1)
		Colorectal Carcinoma	RHOA(1), VCAM1(5), MAPK3(18), EGR1(14), RARA(1)
		Malignant tumor of Colon	RHOA(6), VCAM1(1), MAPK3(5), EGR1(6), RARA(1)
		Carcinogenesis	RHOA(2), VCAM1(5), MAPK3(5), EGR1(11), RARA(9)
		Neoplasm Metastasis	RHOA(8), VCAM1(21), MAPK3(20), EGR1(14)
	Tumor progression		VCAM1(7), MAPK3(11), EGR1(14), rara(2)
Prostate	Neoplasm	Prostatic Neoplasm	IGF1(14), CCND2(2), KLK3(75), FLNA(1), PXN(1), RAF1(1), IGF2(4), INS(2)
		Malignant neoplasm of Prostate	IGF1(67), CCND2(6), KLK3(781), FLNA(2), MAPK9(3), PXN(6), RAF1(7), IGF2(25), INS(31), PTENP1(1)
		Prostate carcinoma	IGF1(69), CCND2(5), KLK3(776), FLNA(2), MAPK9(4), PXN(6), RAF1(8), IGF2(25), INS(30), PTENP1(1)
		Carcinogenesis	IGF1(43), CCND2(15), KLK3(5), FLNA(2), RAF1(24), IGF2(25), IGF2(67), PTENP1(2), MAPKAPK5(1)
	Prostatic disease		IGF1(1), KLK3(11)
	tumor progression		IGF1(8), CCND2(3), KLK3(14), PXN(3), RAF1(9), IGF2(17), PTENP1(1)
Gastric	Stomach neoplasm		ARPC1B(1), ADH1A(2), SPP1(4), CDK4(1), ALDH2(2), ADH7(1)
	Malignant neoplasm of stomach		ARPC1B(1), ADH1A(1), CYP2C9(1), SPP1(20), ALDH2(15), ADH7(1)
	Stomach Carcinoma		ARPC1B(1), ADH1A(2), SPP1(21), NOS2(2), CKS1B(1), ALDH2(14), ADH7(1)
	Gastrointestinal stomal tumors		HSP90AB1(1), CYP2C9(1), SPP1(2), CDK4(3)
	Gastritis		SPP1(3), CYP2C9(2), ALDH2(3)

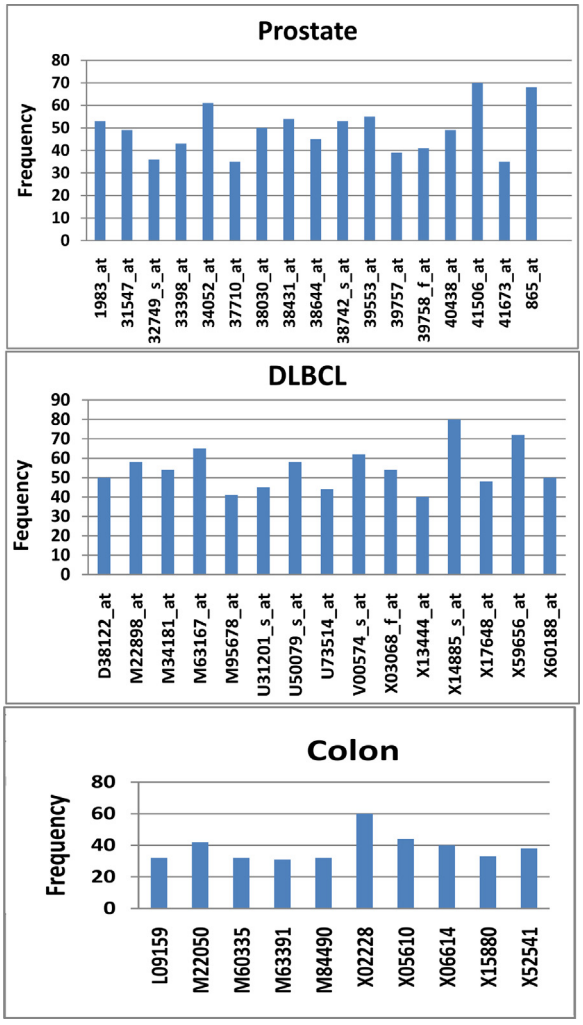

Fig. 14. Frequency plots for selected pathway genes for (i) prostate; (ii) DLBCL; (iii) colon cancer dataset.

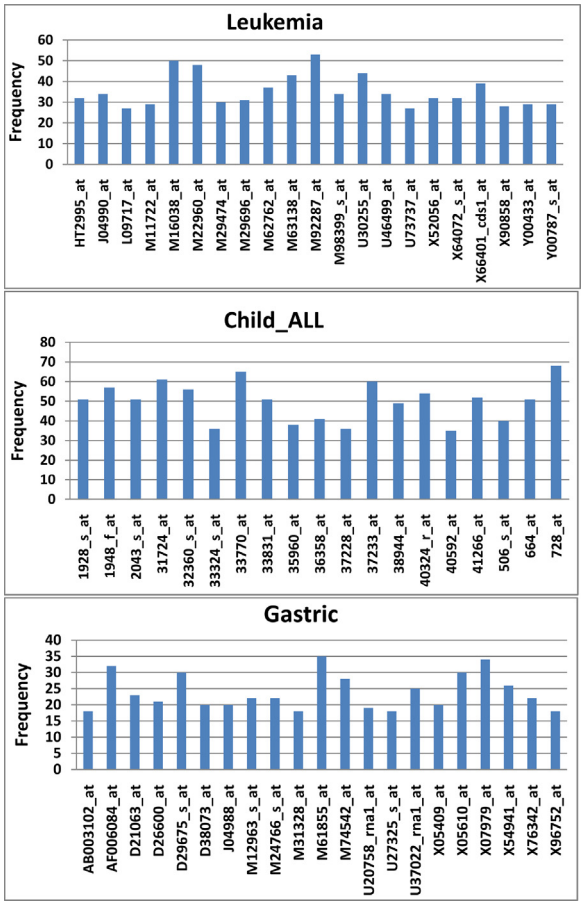

Fig. 15. Frequency plots for selected pathway genes for (i) leukemia; (ii) Child_ALL; (iii) gastric cancer dataset.

cost function and total number of iterations respectively. First, the computational complexity of the overall methodology MFDPSO-BLABC is given and then the complexity of MFDPSO and BLABC are given as both can also be used independently for the feature reduction purpose. The overall complexity of MFDPSO is same as BPSO based method [22]. BLABC also has complexity same as the basic ABC algorithm. Mean, median, LLR, CORG are simple deterministic methods, whereas BPSO and MFDPSO-BLABC are the population based heuristic searches. It is evident that the complexity of heuristic methodologies is higher than the basic approaches, but they are very efficient to produce promising results. Also, it is observed that though the MFDPSO-BLABC and other heuristic approaches are computationally comparable, the MFDPSO-BLABC algorithm produces more promising result.

The results are further compared with other results reported in different literatures on gene selection methodology and a comparative study is shown in Table 8. In Section 4.3.1, we observe that variation in the number of feature size produces varying accuracy from one dataset to another dataset. The result used for comparison is the result obtained using the SVM classifier. The classification is performed with the optimal feature set size for which we obtain the highest accuracy. Most of the methodologies used here for comparison are non-pathway based gene selection processes for cancer classification except [22]. From the result, we can conclude that our proposed work is able to perform more accurately for all the dataset except for the Colon dataset. But, the overall performance of the proposed technique indicates its promising aspects compared to other methodologies.

4.3.4. Cancer pathway markers discovery

The experiments performed with the developed classification model show that the resultant pathway gene feature set generated by the MFDPSO-BLABC method differs in every execution. This is obvious because at every execution the initialization values of MFDPSO and BLABC are changing.

Now our goal is to find out the most frequent pathway marker set which is helpful in pathway diagnosis of cancer and also maximizes the classification evaluation. In our experiment, 10 times 10 fold cross validation are conducted on the dataset, and 100 different sets of features are produced for each of the diseases. Now the frequency of occurrence of each gene within these feature set is recorded. The frequency plots for few top genes are shown in Figs. 14 and 15 for the different cancer dataset. The highly frequent genes are considered as pathway markers which is able to differentiate cancer classes.

4.4. Biological significance

The biological contribution of those identified pathway marker genes to the pathway of a cancer are illustrated in this section. Top 20% of the highly occurring genes are chosen and their involvement in biological activities and cancer progression are observed. The gene-disease association is presented in Table 9. For each dataset, few related pathway diseases are reported and for each disease the related pathway markers are identified which are present in

our selected feature list. We search the disease-gene association database in a public website http://www.disgenet.org/ and have reported the number of Pubmed citation present against the gene for this particular disease which in turn supports and validates the significance of our work. For an example, in our experimental result, we have obtained pathway genes like CXCR2, SMAD3, PLK1, ITGA6, IKBKB, CREBBP, CHUK, CDK as the marker gene for Child_ALL dataset. Now the relation between those genes and disease related to the Child_ALL are given in the table and corresponding Pubmed citations are noted as evidence in support. For all six datasets, the biological contributions are reported.

5. Conclusion

In this paper, a novel multi-fitness discrete PSO followed by blended Laplacian ABC algorithm is proposed and developed for the identification of the pathway gene markers. The markers are differentially expressed to identify different classes of cancer. Automatic clustering using BLABC selects some prominent class dependent pathway genes which are found to perform very well in disease classification. Our experimental setup has proven to be very efficient and effective to find out relevant and informative subset of pathway markers with high discriminating power. Further achievement of higher classification accuracies on most of the datasets compared to other relevant works, is validated in terms of robustness in the performance. Few pathway genes are selected from the frequency analysis of multiple runs and are validated as the real pathway markers via biological interpretation. In our future work, we will design a new multi-objective procedure that can work effectively in the selection process of gene as well as can be applied in varied feature selection problems in different research field.

References

[1] J.B. Overdevest, D. Theodorescu, J.K. Lee, Utilizing the molecular gateway: the path to personalized cancer management, Clin. Chem. 55 (2009) 684–697.
[2] L. Zhang, J. Kuljis, X. Liu, Information visualization for dna microarray data analysis: a critical review, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 38 (2008) 42–54.
[3] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, D. Lee, Inferring pathway activity toward precise disease classification, PLoS Comput. Biol. 4 (2008) e1000217.
[4] D. Yang, R.S. Parrish, G.N. Brock, Empirical evaluation of consistency and accuracy of methods to detect differentially expressed genes based on microarray data, Comput. Biol. Med. 46 (2014) 1–10.
[5] K.-H. Chen, K.-J. Wang, M.-L. Tsai, K.-M. Wang, A.M. Adrian, W.-C. Cheng, T.-S. Yang, N.-C. Teng, K.-P. Tan, K.-S. Chang, Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm, BMC Bioinform. 15 (2014) 1.
[6] W. Zhou, J.A. Dickerson, A novel class dependent feature selection method for cancer biomarker discovery, Comput. Biol. Med. 47 (2014) 66–75.
[7] A. Zibakhsh, M.S. Abadeh, Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function, Eng. Appl. Artif. Intell. 26 (2013) 1274–1281.
[8] E. Bonilla-Huerta, A. Hernández-Montiel, R. Morales-Caporal, M. Arjona-López, Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data, IEEE/ACM Trans. Comput. Biol. Bioinform. 13 (2016) 12–26.
[9] P. GaneshKumar, C. Rani, D. Devaraj, T. Victoire, Hybrid ant bee algorithm for fuzzy expert system based sample classification, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (2014) 347–360.
[10] S. Saha, A. Ekbal, K. Gupta, S. Bandyopadhyay, Gene expression data clustering using a multiobjective symmetry based clustering technique, Comput. Biol. Med. 43 (2013) 1965–1977.
[11] D.L. Masica, R. Karchin, Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival, Cancer Res. 71 (2011) 4550–4561.
[12] S. Efroni, C.F. Schaefer, K.H. Buetow, Identification of key processes underlying cancer phenotypes using biologic pathway analysis, PLoS ONE 2 (2007) e425.
[13] A.H. Bild, G. Yao, J.T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J.M. Lancaster, A. Berchuck, et al., Oncogenic pathway signatures in human cancers as a guide to targeted therapies, Nature 439 (2006) 353–357.
[14] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.
[15] R. Dash, B.B. Misra, Pipelining the ranking techniques for microarray data classification: a case study, Appl. Soft Comput. 48 (2016) 298–316.
[16] J. Apolloni, G. Leguizamón, E. Alba, Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments, Appl. Soft Comput. 38 (2016) 922–932.
[17] H. Salem, G. Attiya, N. El-Fishawy, Classification of human cancer diseases by gene expression profiles, Appl. Soft Comput. 50 (2017) 124–134.
[18] C.-H. Zheng, W. Yang, Y.-W. Chong, J.-F. Xia, Identification of mutated driver pathways in cancer using a multi-objective optimization model, Comput. Biol. Med. 72 (2016) 22–29.
[19] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E.J. Topol, et al., Towards precise classification of cancers based on robust gene functional expression profiles, BMC Bioinform. 6 (2005) 1.
[20] M. Shuangge, K.M. R, Identification of differential gene pathways with principal component analysis, Bioinformatics 25 (2009) 882–889.
[21] J. Su, B.J. Yoon, E.R. Dougherty, Accurate and reliable cancer classification based on probabilistic inference of pathway activity, PLoS ONE 4 (2009) 148–161.
[22] M. Monalisa, M. Jyotirmay, M. Anirban, A PSO-based approach for pathway marker identification from gene expression data, IEEE Trans. Nanobiosci. 14 (2015) 591–597.
[23] S. Bandyopadhyay, S. Mallik, A. Mukhopadhyay, A survey and comparative study of statistical tests for identifying differential expression from microarray data, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (2014) 95–115.
[24] P.A. Jaskowiak, R.J. Campello, I.G. Costa, Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2013) 845–857.
[25] C. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, vol. 55, University of California, Department of Information and Computer Science, Irvine, CA, 1998 http://www.ics.uci.edu/~mlearn/mlrepository.html.
[26] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. 96 (1999) 6745–6750.
[27] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, et al., Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.
[28] Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J.-M. Chong, M. Fukayama, T. Kodama, H. Aburatani, Global gene expression analysis of gastric cancer by oligonucleotide microarrays, Cancer Res. 62 (2002) 233–240.
[29] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[30] J.M. Arevalillo, H. Navarro, Exploring correlations in gene expression microarray data for maximum predictive-minimum redundancy biomarker selection and classification, Comput. Biol. Med. 43 (2013) 1437–1443.
[31] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics 21 (2005) 631–643.
[32] U. Maulik, D. Chakraborty, Fuzzy preference based feature selection and semisupervised SVM for cancer classification, IEEE Trans. Nanobiosci. 13 (2014) 152–160.
[33] G.-Z. Li, X.-Q. Zeng, J.Y. Yang, M.Q. Yang, Partial least squares based dimension reduction with gene selection for tumor classification, in: 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering, IEEE, 2007, pp. 1439–1444.
[34] J. Ye, T. Li, T. Xiong, R. Janardan, Using uncorrelated discriminant analysis for tissue classification with gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinform. 1 (2004) 181–190.
[35] C.-P. Lee, W.-S. Lin, Y.-M. Chen, B.-J. Kuo, Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method, Expert Syst. Appl. 38 (2011) 4661–4667.
[36] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, S. Mavroudi, YamiPred: a novel evolutionary method for predicting pre-miRNAs and selecting relevant features, IEEE/ACM Trans. Comput. Biol. Bioinform. 12 (2015) 1183–1192.
[37] G. Kerr, H.J. Ruskin, M. Crane, P. Doolan, Techniques for clustering gene expression data, Comput. Biol. Med. 38 (2008) 283–293.
[38] A. Mukhopadhyay, M. Mandal, Identifying non-redundant gene markers from microarray data: a multiobjective variable length PSO-based approach, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (2014) 1170–1183.
[39] R.C. Eberhart, J. Kennedy, et al., A new optimizer using particle swarm theory, in: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, vol. 1, New York, NY, 1995, pp. 39–43.
[40] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, et al., Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (2002) 68–74.
[41] M.H. Cheok, W. Yang, C.-H. Pui, J.R. Downing, C. Cheng, C.W. Naeve, M.V. Relling, W.E. Evans, Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells, Nature genetics 34 (2003) 85–90.
[42] K.E. Parsopoulos, M.N. Vrahatis, et al., Particle swarm optimization method for constrained optimization problems, Intell. Technol. Theory Appl.: New Trends Intell. Technol. 76 (2002) 214–220.

[43] D. Karaboga, B. Basturk, Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems, in: International Fuzzy Systems Association World Congress, Springer, 2007, pp. 789–798.
[44] V. Garg, K. Deep, Performance of Laplacian biogeography based optimization algorithm on CEC 2014 continuous optimization benchmarks and camera calibration problem, Swarm Evol. Comput. 27 (2016) 132–144.
[45] A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recogn. Lett. 31 (2010) 651–666.
[46] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.
[47] S.B. Cho, H.-H. Won, Cancer classification using ensemble of neural networks with multiple significant gene subsets, Appl. Intell. 26 (2007) 243–250.
[48] L.-K. Luo, D.-F. Huang, L.-J. Ye, Q.-F. Zhou, G.-F. Shao, H. Peng, Improving the computational efficiency of recursive cluster elimination for gene selection, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (2011) 122–129.
[49] D. Chakraborty, U. Maulik, Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning, IEEE J. Transl. Eng. Health Med. 2 (2014) 1–11.