Analytical Biochemistry 627 (2021) 114242

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Gene selection using hybrid dragonfly black hole algorithm: A case study on RNA-seq COVID-19 data

Elnaz Pashaei [a,*], Elham Pashaei [b]

[a] *Department of Software Engineering, Istanbul Aydin University, Istanbul, Turkey*
[b] *Department of Computer Engineering, Istanbul Gelisim University, Istanbul, Turkey*

ARTICLE INFO

Keywords:
Gene selection
Binary black hole algorithm
Binary dragonfly algorithm
Gene expression
Cancer classification

ABSTRACT

This paper introduces a new hybrid approach (DBH) for solving gene selection problem that incorporates the strengths of two existing metaheuristics: binary dragonfly algorithm (BDF) and binary black hole algorithm (BBHA). This hybridization aims to identify a limited and stable set of discriminative genes without sacrificing classification accuracy, whereas most current methods have encountered challenges in extracting disease-related information from a vast amount of redundant genes. The proposed approach first applies the minimum redundancy maximum relevancy (MRMR) filter method to reduce the dimensionality of feature space and then utilizes the suggested hybrid DBH algorithm to determine a smaller set of significant genes. The proposed approach was evaluated on eight benchmark gene expression datasets, and then, was compared against the latest state-of-art techniques to demonstrate algorithm efficiency. The comparative study shows that the proposed approach achieves a significant improvement as compared with existing methods in terms of classification accuracy and the number of selected genes. Moreover, the performance of the suggested method was examined on real RNA-Seq coronavirus-related gene expression data of asthmatic patients for selecting the most significant genes in order to improve the discriminative accuracy of angiotensin-converting enzyme 2 (ACE2). ACE2, as a coronavirus receptor, is a biomarker that helps to classify infected patients from uninfected in order to identify subgroups at risk for COVID-19. The result denotes that the suggested MRMR-DBH approach represents a very promising framework for finding a new combination of most discriminative genes with high classification accuracy.

1. Introduction

Microarray technology has become one of the most important analytical tools for medical researchers and biologists to track the activity of tens of thousands of genes in a particular organism simultaneously, known as gene expression profiling. High-throughput gene expression analysis has enabled significant advances in 1) identification of diagnostic or prognostic biomarkers, 2) monitoring therapeutic response, 3) understanding the pathogenesis of the diseases, and 4) classification of deadliest diseases such as cancer, Alzheimer's, diabetes, etc. As a result, microarray analysis plays a crucial role in clinical medicine for the discovery process and medical decision support. Classification of expression microarray data is one of the hot topics in bioinformatics. Accurate diagnosis and prognosis of patient diseases and improved clinical decision-making can be achieved by conducting microarray data classification. The classification of microarray data is a challenging task due to the curse of dimensionality problem. The high-dimensional microarray data include the expression of a very large number of genes (features) with a limited number of samples, in which the majority of the genes are irrelevant and redundant. It negatively influences the classification accuracy, and gene selection is required to surmount this challenge [1].

The predictive accuracy of classification in microarray data is strongly dependent on the gen selection (GS) approaches. The GS identifies the most relevant and reliable genes and eliminates insignificant genes from microarray datasets so that more robust, accurate, and reliable predictions can be made. In other words, the GS aims to determine the optimum number of genes with the highest classification accuracy. From a biological perspective, the main advantages of the GS include: 1) discovering the most significant and discriminative genes called biomarkers, 2) increasing the interpretability of microarray data, and 3) reducing clinical cost. So far, many GS techniques have been

[*] Corresponding author.
E-mail addresses: elnazpashaei@aydin.edu.tr (E. Pashaei), epashaei@gelisim.edu.tr (E. Pashaei).

https://doi.org/10.1016/j.ab.2021.114242
Received 4 December 2020; Received in revised form 12 April 2021; Accepted 2 May 2021
Available online 8 May 2021

proposed in the literature to determine informative genes and consequently improve sample estimation accuracy in biological data.

GS techniques are mainly classified into three groups which are filter, wrapper, and hybrid approaches. In filter methods, informative theory or statistical relationships such as dependency or correlation between genes are computed to assign a relevance score to each gene without invoking a classification model. Then, according to the scores, the genes are ranked and a subset with the highest-ranking genes is selected for further analysis [2]. Some of the most widely used filtering methods in gene selection are minimum redundancy maximum relevance (MRMR) [3], information gain (IG) [4], Relief [5], chi-square [6], etc. In the wrapper-based methods, the process is composed of two parts: 1) search techniques to select the optimal feature subset, and 2) evaluation, in which the performance of a classifier, as a fitness function, is considered to evaluate the goodness of selected gene subsets. With the increasing number of genes, the space for gene search grows exponentially, causing the stage of generating all possible gene subsets in wrapper approaches to become an impractical and time-consuming task. Therefore, the wrapper approaches utilize nature-inspired optimization algorithms (NIOAs) to direct the search process efficiently. The process begins with a population of the solutions, in which each individual represents a candidate gene subset. Each individual's fitness is measured by using a classifier, and individuals are updated according to the adopted NIOA's characteristics. To improve the result, the process is repeated until reaching the necessary number of iterations.

Filter approaches are computationally efficient in selecting gene subsets relative to wrapper methods, though they show lower classification performance than wrapper approaches. To improve the discovery of disease biomarkers within microarray data, hybrid approaches have been introduced in the literature. The hybrid approaches combine a filter method with a wrapper method to take benefit from the high classification performance of the wrappers and the efficiency of the filters simultaneously. It first utilizes a filter approach to reduce the feature space dimension, and then a wrapper method to select the optimal feature subset. The majority of suggested methods in the literature for gene selection are based on hybrid approaches, which lead to choosing more informative genes with acceptable classification accuracy. In hybrid methods, various NIOA based wrapper approaches have been used for gene selection such as genetic algorithm (GA) [7], memetic algorithm (MA) [8], ant colony optimization (ACO) [9], bat algorithm (BA) [1,10,11], flower pollination algorithm (FPA) [12], artificial bee colony (ABC) [13], particle swarm optimization (PSO) [14], harmony search algorithm (HSA) [15], grasshopper optimization algorithm (GOA) [16], adaptive inertia weight teaching-learning-based optimization algorithm (ATLBO) [17], binary dragonfly (BDF) algorithm [18], etc.

Since the hybrid models view these two filter and wrapper stages as being separate, the accuracy of hybrid methods is still a key issue [19]. To enhance the performance of the hybrid methods, more advanced and sophisticated models have been developed. Recently, a combination of two NIOAs has been suggested to improve the efficiency and robustness of the gene selection approaches. Some of them are hybrid of TLBO and simulated annealing (SA) algorithm (TLBOSA) [20], hybrid of TLBO and gravitational search algorithm (GSA) [21], hybrid of BA and Hill climbing algorithm [3], etc.

Although a variety of NIOAs has been applied to the microarray data, there is still no guarantee to find the optimal subset of genes for classification problems due to the stochastic nature of these techniques [22]. Also, gene selection continues to be a difficult task due to the huge gene search space and complex gene interactions [1,19]. Therefore, further investigations are required to develop an efficient gene selection method [3].

The DF algorithm (DFA) [23] is a widely used swarm intelligence-based optimization approach that is inspired by dragonflies' static and dynamic flocking behavior in nature. Various extensions and improvements of DFA have been suggested for optimizing a large variety of different problems [24,25]. In the context of feature selection, binary DFA (BDF) [26] and its different variants [27,28] have been developed. Although DFA and its hybridized variants have demonstrated successful results in solving several complex optimization problems, they still have the disadvantage of trap in local optima. Yet, it is firmer than other NIOAs, and can easily be combined with other algorithms [24]. To reduce the probability of the algorithm falling into local optima, hybrid of BDF and binary black hole algorithm (BBHA) [29] is suggested in this study.

The BHA [30] is an efficient physical-based optimization approach that is inspired by the actual behavior of the black hole in space. The BHA has shown an excellent performance in finding a near-optimal solution as compared to other NIOAs in various application domains [31,32]. The BHA has many advantages, including free-controller parameter, simplicity, easy implementation, strength search capabilities, fast coverage speed, and local optima avoidance. However, the BHA lacks exploration capabilities on some complex datasets [33]. To address this drawback, the BHA has been modified and hybridized with other NIOAs to empower the algorithm's exploration and exploitation process. Different variants of BHA have been proposed for various kinds of continuous optimization problems [32,34,35]. For solving feature selection problems, binary BHA [36], and its variant [37] have been introduced. However, BBHA is not yet properly investigated for a kind of gene selection problem.

In this paper, a hybrid filter/wrapper gene selection model is proposed based on the MRMR approach and hybridized BDF with BBHA (DBH), for cancer classification. In the suggested method, MRMR is first operated as a filter method to select high-ranked genes. This is intended to fine-tune the search space for the wrapper approach. Then, the high-ranked genes are passed to the hybrid DBH algorithm, in which BDF acts as an internal operator to provide a strong initial population for BBHA. The main goal is to improve the mechanism of the initial population production stage in the BBHA to cover the feasible space properly. The hybridization in the wrapper stage boosts the efficiency of BBHA in terms of satisfactory local optima avoidance and balancing the exploration and exploitation to offer better prediction accuracy, fast convergence, and robustness compared to standard BBHA, standard BDF, and other state-of-art existing methods for biomarker discovery and cancer diagnosis. The proposed method is known as MRMR-DBH, and it uses the support vector machine (SVM) classifier to evaluate each candidate gene subset in DBH.

It's also worth noting that the efficiency of the standard BHA [36] and a combination of PSO and BHA [38] for gene selection has been examined before. In Ref. [36] a two-stage hybrid algorithm was introduced for feature selection of biological data, in which the chi-square method is employed to filter the genes at the first stage. Meanwhile, in the second stage, standard BHA is used to generate the gene subsets, and the random forest (RF) classifier is utilized for the evaluation purpose. In Ref. [38], random-forest-recursive feature elimination (RF-RFE) algorithm is used to filter out irrelevant and redundant genes, and a hybrid of BPSO and BBHA is employed to seek further informative genes. SPLSDA classification model is the evaluator for each candidate gene subset.

Even though the suggested MRMR-hybrid DBH-BBHA (DBH)-SVM method exploits the BBHA in its wrapper stage, its utilized filter approach, adopted classifier, and wrapper's collaborative structure are entirely different from previous BBHA based gene selection approaches. In Ref. [38], BBHA was used as a local search within PSO, while in our research, BDF was used as an internal initialization operator within the BBHA. To de best of our knowledge, integration of MRMR and BBHA with SVM classifier has not been investigated, and BBHA and BDF have collaborated for the first time to develop a hybrid wrapper model aimed to find the more robust and insightful genes in a complex search space.

The main contributions of this study in the context of gene selection and cancer classification can be briefly outlined as follows.

- Using the MRMR filtering method in conjunction with hybridized BBHA (DBH) to identify strongly discriminative genes with the help of the SVM classifier.
- Developing a new wrapper approach by merging BDF and BBHA to direct the quest for a more robust and informative gene subset considering the accuracy and feature selection stability.
- To examine whether the suggested MRMR-DBH will reach a gene subset with a limited number of genes and higher classification accuracy than the existing gene selection method.
- To explore the efficiency of MRMR-DBH in a real RNA- Seq coronavirus-related gene expression dataset of asthmatic patients, to identify a small appropriate gene subset to improve the discriminative power of the angiotensin-converting enzyme 2 (ACE2) biomarker.

The suggested algorithm's performance was evaluated on eight well-known benchmark gene expression datasets of different characteristics. First, MRMR performance on dissimilarity of selected top M genes was investigated as a preprocessing step, and compared to other well-regarded filtering methods. The filtering approaches used SVM to estimate the accuracy of the classification. Also, a performance comparison of the SVM with other classifiers was conducted when they have used M top genes. The results obtained using BDF-BBHA technique has been compared with standard BDF and BBHA techniques in the term of accuracy, the number of selected genes, and execution time. The result was also compared with numerous gene selection techniques. The comparative results indicate that the MRMR-DBH-SVM approach consistently outperformed current BBHA based and other established gene selection approaches in terms of classification accuracy, and the optimal number of retained genes.

The rest of the paper is structured as follows. Some theoretical background is described in Section 2. The details of the proposed DBH algorithm are presented in Section 3. Section 4 explains the details of the conducted experiments and obtained results. Finally, Section 5 concludes this study.

2. Preliminary concepts

This section offers a brief definition of the gene selection problem as well as a detailed overview of the MRMR filtering method. Then, the standard BDF and the BHA approaches, which are later integrated to develop a more robust optimization algorithm are described. Finally, SVM is demonstrated in detail.

2.1. Problem definition

The gene expression microarray data consists of a very large number of genes with a limited number of samples. The majority of genes (features) in microarray data are redundant, irrelevant, and noisy, causing classifier algorithms to perform poorly on microarray data in regard to classification accuracy and computational cost. Gene selection is a process of providing a small biologically important gene subset to achieve more accurate classification results. The total number of possible gene subsets for N genes is 2^N.The search for optimal gene subset is an NP-hard problem since the search space, which includes all possible subsets, exponentially grows as the number of genes increases [10,12].

2.2. MRMR

The MRMR filter feature selection method [39] was employed as a preprocessor before the wrapper approach, to filter out the irrelevant and redundant features from high-dimensional data sets. The MRMR's idea is to choose features that have a high correlation with the class label (maximum relevancy) but a low correlation between themselves (minimum redundancy). MRMR calculates the score of all features and selects the feature with the maximal score at each iteration in a greedy forward manner. MRMR utilizes mutual information to measure features' redundancy and relevance in order to rank them [40]. The mutual information between two variables X and Y is defined as:

$$I(Y;X) = H(Y) - H(Y|X) \tag{1}$$

$$H(Y) = -\sum_{y} p(y) log_2(p(y)) \tag{2}$$

$$H(Y|X) = \sum_{x} p\left(y\right)\left(-\sum_{y} p(y|x) log_2\left(p\left(y|x\right)\right)\right) \tag{3}$$

where $H(.)$ represents entropy, and p denotes (empirical) probability mass function.

Considering a data set with n instances of the m features $X_1, .., X_m$ and a class variable Y, the filter score for each feature X_k is defined as:

$$J_{MRMR}(X_K) = I(Y;X_k) - \frac{1}{|S|}\sum_{X_j \in S} I(X_k;X_j) \tag{4}$$

The term $I(Y;X_k)$ measures feature relevance to the target, based on the knowledge it has about Y. Let S stand for the list of features that have already been selected. The initial value of S is set to $S = \{X_k\}$ with $I(Y;X_k) = \underset{j \in \{1, \ldots, m\}}{max} I(Y;X_j)$. The term $\frac{1}{|S|}\sum_{X_j \in S} I(X_k;X_j)$ determines the feature's redundancy by assessing the mean information it shares with the features in S. At each iteration, the feature that maximizes the respective score is added to S.

2.3. Binary dragonfly algorithm

The DFA is a recently proposed metaheuristic approach inspired by the unique swarming behaviors of dragonflies. Dragonflies swarm for two reasons. Either they form small flocks and fly back and forth over a tiny region for hunting prey, which is called the static swarm. Or a large number of them create a single group for long-distance migration in one direction, termed as the dynamic swarm. Each NIOA should provide a satisfactory balance between exploration and exploitation to produce good results. Static and dynamic swarming behaviors of dragonflies play the role of exploration and exploitation in DFA, respectively. The position of each dragonfly (solution) in the swarm (population) is updated using a step vector (velocity), which consists of five primitive factors:1) separation, 2) alignment, 3) cohesion, 4) attraction, 5) distraction.

To avoid the collision of a candidate dragonfly (individual) with its adjacent dragonflies in the search space, the separation factor is used. The alignment factor is utilized to match the velocity of the individual with its neighboring dragonflies. The cohesion factor represents the tendency of the individual towards the center of the mass of the neighborhood. Attraction and distraction, as survival factors, reflect the dragonflies' attraction towards food and their escape from enemies. Each factor is modeled mathematically as follows:

$$S_i = -\sum_{j=1}^{N} X - X_j \quad j = 1, 2, 3, \ldots, N \tag{5}$$

$$A_i = \frac{\sum_{j=1}^{N} V_j}{N} \tag{6}$$

$$C_i = \frac{\sum_{j=1}^{N} X_j}{N} - X \tag{7}$$

$$F_i = X^{+} - X \tag{8}$$

$$E_i = X^{-} + X \tag{9}$$

where S_i, A_i, C_i, F_i, and E_i represent the separation, alignment, cohesion, attraction, and distraction motion for i^{th} dragonfly, respectively. X

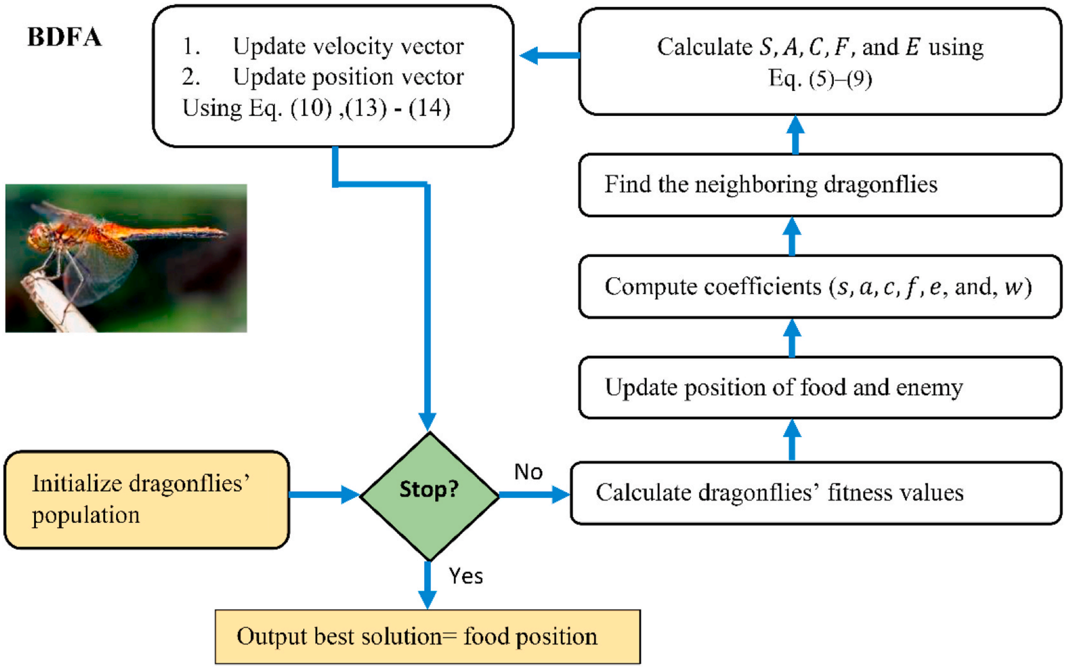

BDFA
1. Update velocity vector
2. Update position vector
Using Eq. (10) ,(13) - (14)
Calculate S, A, C, F, and E using Eq. (5)–(9)
Find the neighboring dragonflies
Compute coefficients (s, a, c, f, e, and, w)
Update position of food and enemy
Initialize dragonflies' population
Stop?
No
Calculate dragonflies' fitness values
Yes
Output best solution= food position

Fig. 1. Flowchart of the BDF algorithm.

```
n ← swarm size (population size)
Max_Iter ← maximum number of iterations
for $i =$ ($1$ $to$ $n$) do
    $X_i$ ← initialize the position of $i^{th}$ dragonfly
    $\Delta X_i$ ← initialize the velocity of $i^{th}$ dragonfly
end for
while ($t <$ Max_Iter) do
    Calculate the fitness value of all dragonflies
    Update positions of food (best solution) and enemy (worst solution)
    Calculate the inertia weight ($w$) and factor weights ($s$, $a$, $c$, $f$, and $e$)
    for $i =$ ($1$ $to$ $n$) do
        Find the neighboring dragonflies around $i^{th}$ dragonfly
        Calculate $S_i$, $A_i$, $C_i$, $F_i$, and $E_i$ according to equations (5) to (9))
        for $k =$ ($1$ $to$ $d$) do // $d$=dimension of each solution
            Update dragonfly velocity( $\Delta X_{t+1}$) according to equation (10)
            Calculate the probability of changing position ( $T(\Delta X)$ ) using equation (13)
            Update $i^{th}$ dragonfly position according to equation (14)
        end for
    end for
    $t = t + 1$
end while
Output: the best solution = Food position
```

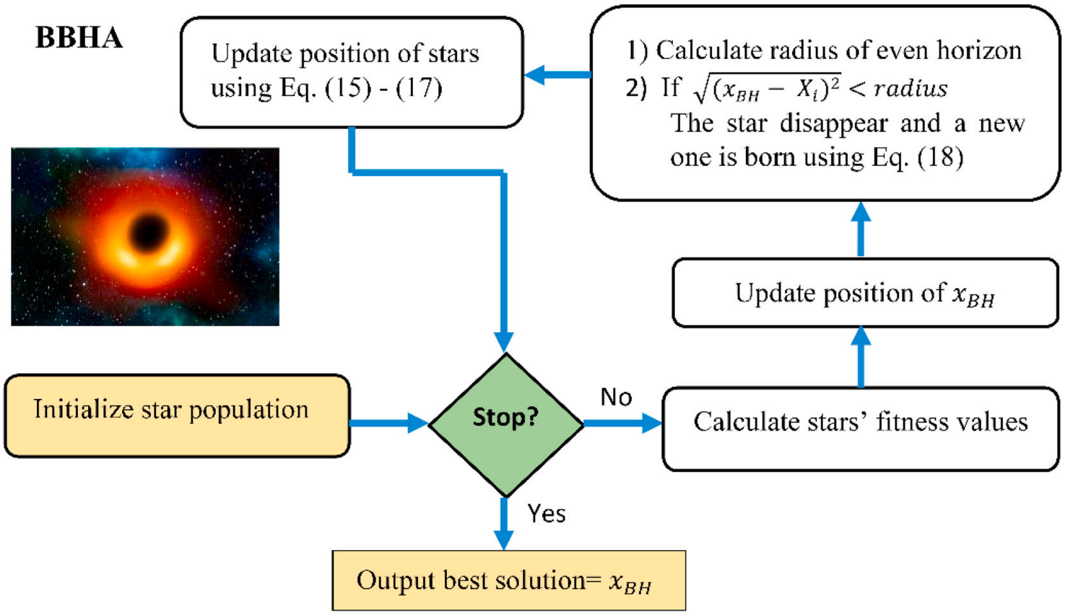

BBHA
Update position of stars using Eq. (15) - (17)
1) Calculate radius of even horizon
2) If $\sqrt{(x_{BH} - X_i)^2} < radius$ The star disappear and a new one is born using Eq. (18)
Update position of x_{BH}
Initialize star population
Stop?
No
Calculate stars' fitness values
Yes
Output best solution= x_{BH}

Fig. 2. Flowchart of the BBHA

shows the current position of the i^{th} dragonfly. X_j and V_j indicate position and velocity of neighboring dragonflies around X, whereas N represents the total number of adjacent dragonflies. The position of the food and enemy are denoted as X^+ and X^- in equations (8) and (9), respectively. It worth mentioning that all the dragonflies in the population are assumed as a group in X neighboring at binary (discrete) search spaces, while Euclidean distance is computed between all the dragonflies for determining the neighborhood of each dragonfly in continuous search space.

Then, the step vector showing the direction of dragonflies' motions is determined as follow:

$$\Delta X_{t+1} = (sS_i + aA_i + cC_i + fF_i + eE_i) + w\Delta X_t \tag{10}$$

where s, a, c, f, e indicate the separation weight, alignment weight, cohesion weight, food factor, and enemy factor for i^{th} dragonfly, respectively. And, w and t are inertia weight and iteration counter.

In the continuous search space, the new position of the i^{th} dragonfly is determined by adding the step vector to its previous position if at least one dragonfly is located in its neighborhood. Otherwise, the Levy flight operator is utilized to update the position of the dragonfly.

$$X_{t+1} = X_t + \Delta X_{t+1} \tag{11}$$

$$X_{t+1} = X_t + Levy^{*}X_t \tag{12}$$

However, in the binary search space, the new position of the i^{th} dragonfly is determined by employing the following transfer function to define the probability of changing position.

$$T(\Delta X) = \left| \frac{\Delta X}{\sqrt{\Delta X^2 + 1}} \right| \tag{13}$$

$$X_{t+1} = \begin{cases} \neg X_t & r < T(\Delta X_{t+1}) \\ X_t & r \geq T(\Delta X_{t+1}) \end{cases} \tag{14}$$

The transfer function takes the step vector as inputs and generates output between 0 and 1. Then the $T(\Delta X)$ result is used to convert the position vector variables to either 0 or 1 depending on equation (10), where r indicates a random number in [0, 1]. For more information refer to Refs. [25,26]. Pseudocode and flowchart of the BDF algorithm have been presented in Algorithm 1 and Fig. 1, respectively.

2.4. Binary black hole algorithm

The idea of BHA [30] is taken from the phenomenon of the black hole in space. A black hole (BH) is a region of space containing a large volume of mass, in which the entire mass lies in its middle. There is a heavy gravitational pull around the BH, called the event horizon, which ensures that no nearby object can escape from it. A binary version of BHA (BBHA) was introduced in Ref. [36] to make the algorithm applicable for discrete problems since BHA was originally designed for continuous problems.

The BBHA begins with random and binary initialization of candidate solutions. Each solution represents the position of a star in the search space. Then, an objective function is utilized to determine the goodness of each star to find the best solution among all the candidates. The prediction accuracy of a specific classifier is utilized to evaluate stars by computing their fitness value. Among all the solutions in the population, the star with the highest fitness value is chosen as the best solution and introduced as BH. The BH starts to attract stars, and consequently, all of the stars begin to move around the BH since the best place in search space belongs to BH. Stars update their positions, and their movements toward BH are carried out through the following equations:

$$X_i(t+1) = X_i(t) + rand\,[X_{BH} - X_i(t)] \quad i = 1,\ 2, \ldots,\ N \tag{15}$$

$$S(X_i(t+1)) = abs(tanh(X_i(t+1))) \tag{16}$$

$$X_i(t+1) = \begin{cases} 1, & if\ S(X_i(t+1)) > 0.6 \\ 0, & otherwise \end{cases} \tag{17}$$

where rand stands for a random value in the range [0,1]. The $X_i(t)$ and $X_i(t+1)$ represent the position of ith star in iteration t and $t+1$. The X_{BH}

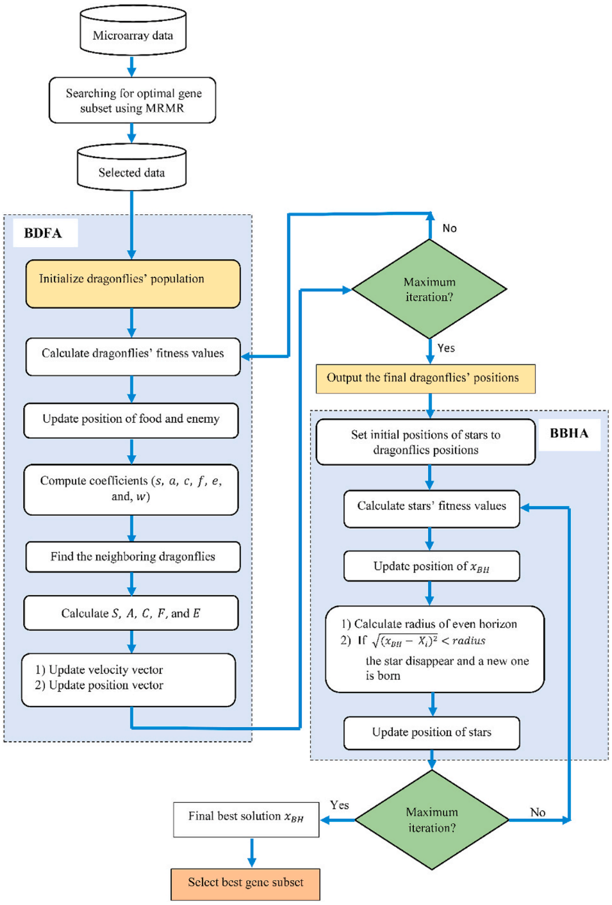

Microarray data
Searching for optimal gene subset using MRMR
Selected data
BDFA
Initialize dragonflies' population
Calculate dragonflies' fitness values
Update position of food and enemy
Compute coefficients (s, a, c, f, e, and, w)
Find the neighboring dragonflies
Calculate S, A, C, F, and E
1) Update velocity vector
2) Update position vector
Maximum iteration?
No
Yes
Output the final dragonflies' positions
BBHA
Set initial positions of stars to dragonflies positions
Calculate stars' fitness values
Update position of x_{BH}
1) Calculate radius of even horizon
2) If $\sqrt{(x_{BH} - X_i)^2} < radius$
the star disappear and a new one is born
Update position of stars
Maximum iteration?
Yes
No
Final best solution x_{BH}
Select best gene subset

Fig. 3. Flowchart of the suggested MRMR-DBH algorithm.

Table 1
Datasets characteristics.

Dataset	#genes	#samples	#classes
Breast Cancer	24481	97 (46, 51)	2
Colon Cancer	2000	62 (40, 22)	2
DLBCL	4026	47 (24, 23)	2
Prostate Tumor	10509	102 (50, 52)	2
MLL	12533	72 (24, 20, 28)	3
CNS	7129	60 (21, 39)	2
Ovarian Cancer	15154	253 (91, 162)	2
ALL-AML	7129	72 (25, 47)	2
GSE149273 (COVID-19)	11765	90 (30, 30, 30)	3

shows the position of the BH (best solution) in search space, and *N*demonstrates numbers of stars in the population with *d*dimension.

After accomplishing stars' movements toward the BH in the first iteration, the fitness values for new stars are calculated. If any of the stars have a better position than the BH considering the fitness values, the BH's position is changed and it takes the position of the star. The algorithm proceeds and stars again begin to travel around this new BH. After the predefined number of iteration, the algorithm is stopped and BH is reported as the best solution for the optimization problem.

It's worth mentioning that if stars cross the event horizon when moving around BH, the BH immediately swallows them and they disappear from search space. It enables the random formation of new stars in the search space. The event horizon radius (R) is calculated as follow:

$$R = \frac{f_{BH}}{\sum_{i=1}^{popsize} f_i} \tag{18}$$

where f_{BH} is the black hole fitness value, f_i is the ith star's fitness value, and . is the total number of stars in the population. The star vanishes from the search space when its distance from the BH is less than the radius of the event horizon. For more details on the steps of the original algorithm refer to Ref. [36]. Algorithm 2 shows the pseudocode of the BBHA and Fig. 2 demonstrates its flowchart.

Due to BBHA's simple structure, durability, high local optima avoidance, and the parameter-less nature of the BHA, it has been successfully applied in solving numerous optimization and engineering problems. However, in some datasets, the BBHA fails to demonstrate its excellent results, so a hybrid version of BBHA has been suggested to enhance its global search capability utilizing the BDF algorithm. It is worth mentioning that the initial population for both BDA and BBH is generated using uniformly distributed random numbers.

2.5. Support vector machine

SVM has demonstrated its usefulness in various biological classification tasks, such as gene expression microarrays. The SVM's outstanding performance in microarray data can be explained by a variety of factors, including their robustness to high-dimensional data, their ability to avoid overfitting by utilizing powerful regularization principles, and their efficiency in learning complex classification functions [41]. The basic idea behind SVM classifiers is to find a maximal margin separating line (or hyperplane) between data of two classes. If the data is not linearly separable at origin, kernel functions are utilized to implicitly map the data to a higher-dimensional space, where a separating hyperplane can be found.

We adopted the SVM classifier as an objective function in the proposed method to evaluate the candidate gene subset solution. Package "e1071" as an interface of "libSVM" in R was used for SVM implementation with linear kernel. Recall that the linear kernel of SVM can be

```
$n$ ← number of stars (population size)
Max_Iter ← maximum number of iterations
for $i = (1\ to\ n)$ do
    $X_i$ ← initialize the position of $i^{th}$ star
end for
while ($t <$ Max_Iter) do
    for $i = (1\ to\ n)$ do
        calculate the $i^{th}$ star's fitness value
        Update positions of the black hole $X_{BH}$ according to fitness value
        and the number of selected features
        Calculate the radius of the event horizon ($R$) using equation (18)
        if $\sqrt{(x_{BH} - X_i)^2} < R$ then
            $i^{th}$ star is swallowed and a new star ($X_{new}$) is generated
        end if
    end for
    for $i = (1\ to\ n)$ do
        for $k = (1\ to\ d)$ do // $d$=dimension of each solution
            Update position of the $i^{th}$ star according to equation (15-17)
        end for
    end for
    $t = t + 1$
end while
Output: the best solution = black hole position
```

Table 2
LOOCV classification accuracy (%) of the filter approaches on the different number of top-ranked genes using SVM classifier.

Dataset	Filter Methods					Dataset	Filter Methods				
	#Top	MRMR	IG	Relief	Chi-square		#Top	MRMR	IG	Relief	Chi-square
Breast cancer	25	74.22	77.31	80.41	75.25	**DLBCL**	25	97.87	97.87	97.87	97.87
	50	81.44	76.28	77.31	76.28		50	100	100	100	100
	78	75.25	74.22	71.13	76.28		75	100	100	100	100
							100	100	100	100	100
							125	100	100	100	100
							150	100	100	100	100
Colon Cancer	25	80.64	83.87	77.41	80.64	**Prostate Tumor**	25	87.25	92.11	90.19	96.07
	50	87.09	79.03	79.03	77.41		50	89.21	94.11	93.13	89.21
	75	79.03	77.41	79.03	75.80		75	92.15	90.19	93.13	91.17
	100	85.48	79.03	80.64	79.03		100	94.11	88.23	93.13	86.27
	125	83.87	87.09	77.41	85.48		125	92.15	91.17	93.13	90.19
	150	80.64	85.48	80.64	85.48		150	90.19	92.15	94.11	89.21
MLL	25	91.66	91.66	93.05	91.66	**Ovarian Cancer**	25	98.02	98.02	100	98.41
	50	95.83	90.27	93.05	91.66		50	100	100	100	100
	75	94.44	94.44	94.44	91.66		75	100	100	100	100
	100	94.44	95.83	94.44	94.44		100	100	100	100	100
	120	94.44	95.83	93.05	95.83		120	100	100	100	100
CNS	25	60	70	75	70	**ALL-AML**	25	95.83	97.22	94.44	95.83
	50	63.33	75	68.33	71.66		50	95.83	98.61	95.83	98.61
	75	70	73.33	71.66	71.66		75	97.22	98.61	97.22	97.22
	100	63.33	70	65	68.33		100	98.61	97.22	97.22	97.22
	125	71.66	66.66	65	65		125	98.61	95.83	98.61	97.22
	150	71.66	68.33	66.66	68.33		150	98.61	97.22	98.61	97.22

defined as:

$$K(x_i,\ x_j) = x_i^T x_j \tag{19}$$

where x_i and x_j are samples with gene expression values. The value of the SVM penalty parameter C was set to one, and feature scaling was applied to data.

3. Proposed method for gene selection

This section provides the methodology followed to implement the proposed MRMR-DBH gene selection algorithm. It contains two main modules which are the filter feature selection approach based on MRMR, and the wrapper feature selection approach based on novel hybrid BDF/BBHA (DBH) algorithm. The filter method is used in the first stage of the proposed model to identify a relevant feature set, and the wrapper model is used in the second stage to find the final feature subset. A detailed description of each stage of the proposed method is given in the following sub-section. The flowchart of the suggested approach is shown in Fig. 3.

3.1. Stage I: filter approach

In this stage, the MRMR feature selection approach was employed to select the initial subset features by analyzing the relevance and redundancy of the genes. MRMR ranking method aims to reduce the original dataset's high dimensionality instead of the exhaustive search over the feature subsets and feed the wrapper method with the most discriminative genes. As a result, it reduces computational load and improves classification accuracy. Performing filtering process is bound to give rank scores for all genes, and *n* top-ranked genes, which are highly correlated with class and uncorrelated with each other, are chosen. Then, the selected *n* top genes will be passed to the subsequent stage to further select a small set of informative genes.

3.2. Stage II: Wrapper approach

In this stage, the wrapper method is used to select a subset of the most informative genes from the set of top-ranking genes collected by the MRMR filtering method. A hybridized version of BHA with DFA was developed as an efficient search method to perform in the wrapper model. The DFA is prone to trapping in local optima, but it is more stable than other NIOAs and can be combined with other algorithms easily. BBHA, on the other hand, has outperformed other NIOAs in terms of finding a near-optimal solution, but it lacks exploration capabilities on some complex datasets. Therefore, the hybridization of DFA as an internal initialization operator with the BBHA can produce an efficient search method by covering the feasible space properly. The proposed DBH wrapper feature selection method uses the most popular SVM fitness function iteratively to evaluate the quality of feature subsets. Our proposed method's ultimate aim is to increase classification accuracy while reducing the number of selected genes.

3.2.1. Solution representation

One of the main challenges confronting the NIOAs when designing them for the problem at hand is solution representation. The solution in this study is a one-dimensional vector with *N* elements, where *N* represents the number of genes in the original dataset. A "1" or "0" value is assigned to each element in the vector. If the value is "1," it means the corresponding gene is selected; otherwise, it is "0."

3.2.2. Fitness function

In the wrapper model, the feature subsets in the search space are evaluated using a fitness function. The fitness function is one the most critical part of the suggested DBH wrapper feature selection approach since a poor judgment of the fitness function can lead to insufficiencies in the model. SVM classifier is used to calculate the fitness values of the candidate solutions in the population. Each solution's fitness is given by a measure of the classification accuracy. Accuracy is defined as the number correct prediction made by SVM using gene subset. The equation for the fitness function is as follows:

$$fintess = \frac{number\ f\ instance\ correctly\ classified}{total\ number\ of\ instance} \tag{20}$$

The model's main goal is to improve classification accuracy with a smaller number of genes. It's worth noting that if the classification accuracy of two subset features is equal, the subset with the fewer genes is chosen.

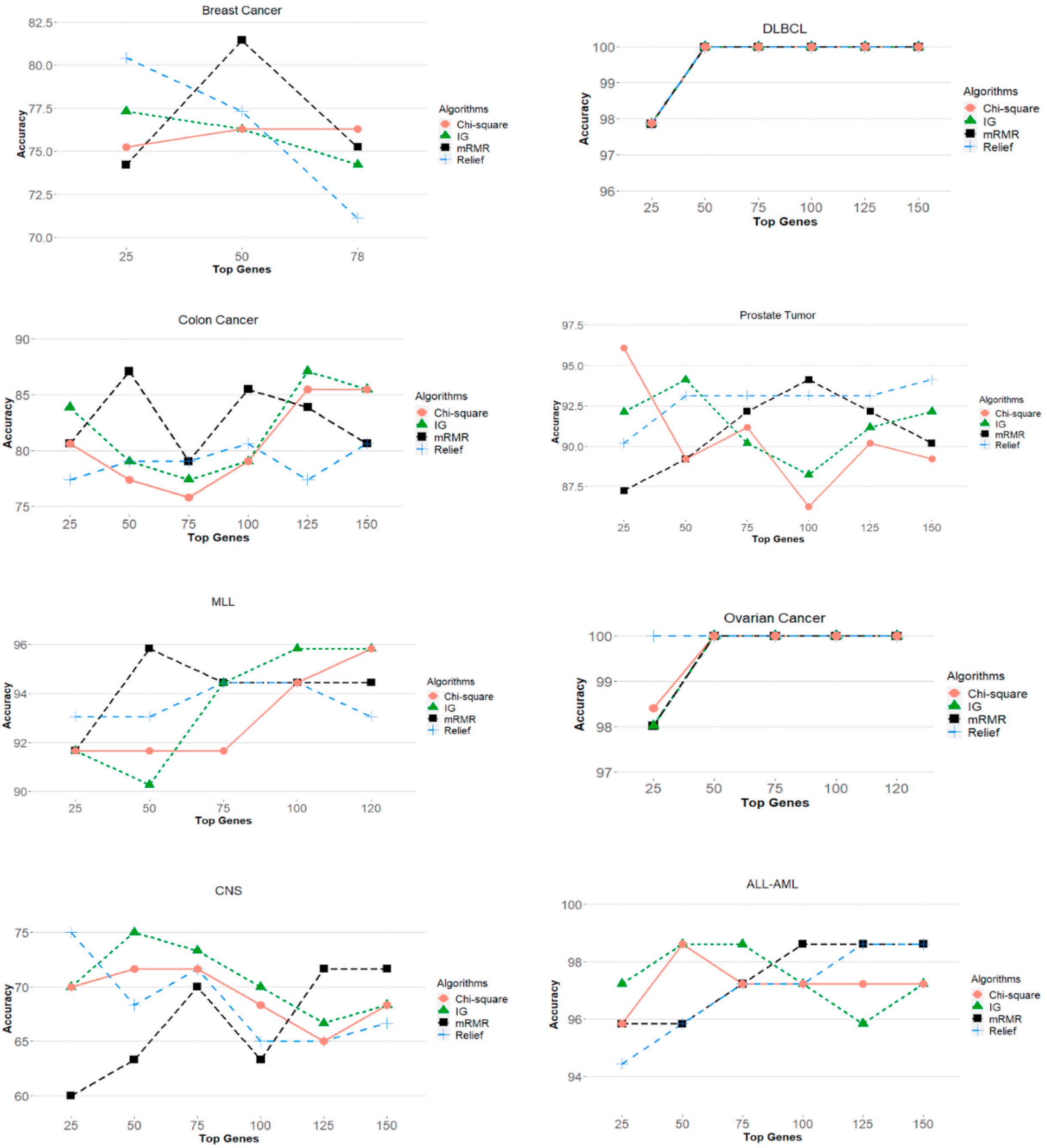
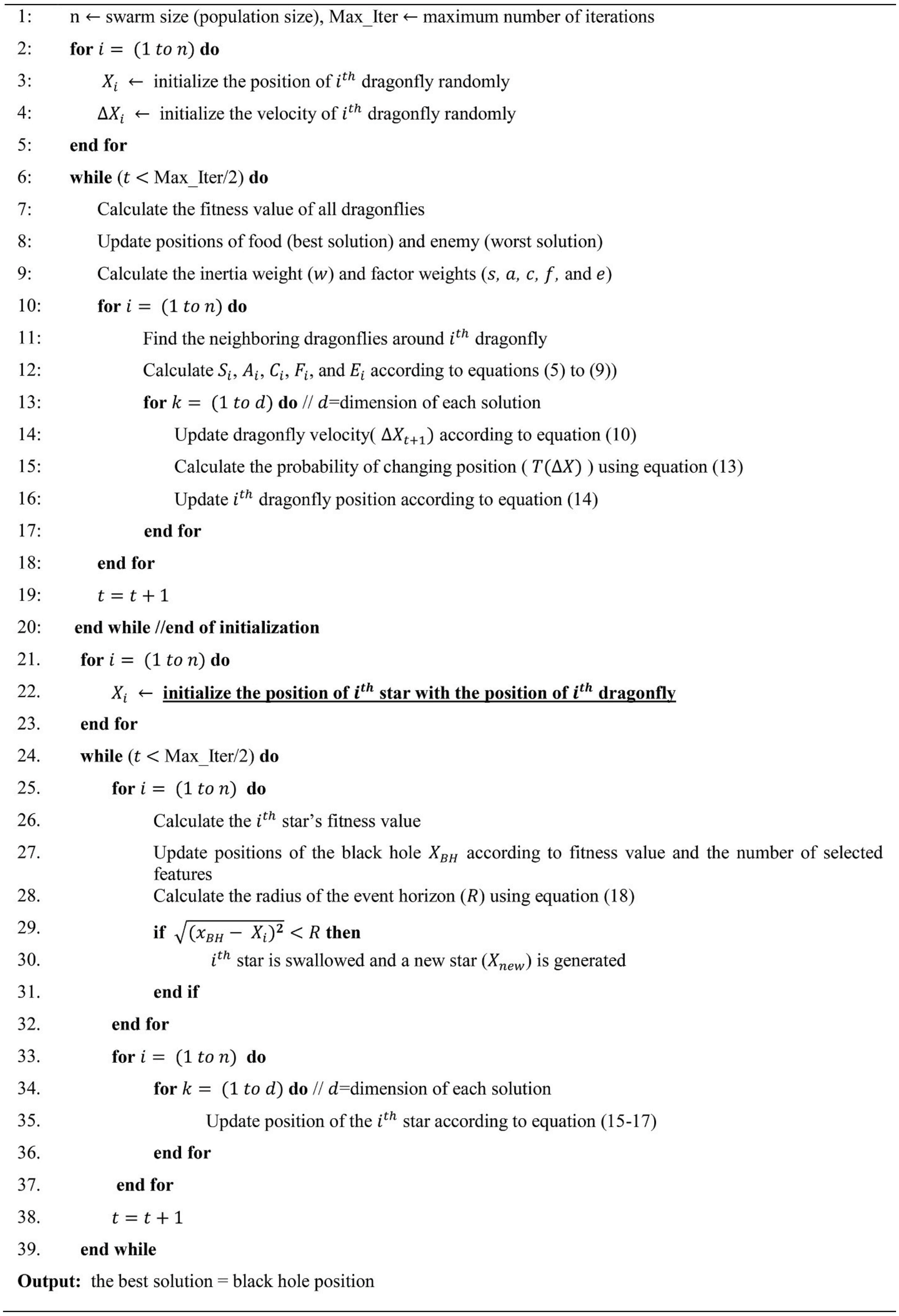


Fig. 4. LOOCV classification accuracy (%) of the filter approaches on the different number of top-ranked genes using SVM classifier.



```
n ← swarm size (population size), Max_Iter ← maximum number of iterations
for $i = (1\ to\ n)$ do
    $X_i$ ← initialize the position of $i^{th}$ dragonfly randomly
    $\Delta X_i$ ← initialize the velocity of $i^{th}$ dragonfly randomly
end for
while ($t <$ Max_Iter/2) do
    Calculate the fitness value of all dragonflies
    Update positions of food (best solution) and enemy (worst solution)
    Calculate the inertia weight ($w$) and factor weights ($s,\ a,\ c,\ f,$ and $e$)
    for $i = (1\ to\ n)$ do
        Find the neighboring dragonflies around $i^{th}$ dragonfly
        Calculate $S_i$, $A_i$, $C_i$, $F_i$, and $E_i$ according to equations (5) to (9))
        for $k = (1\ to\ d)$ do // $d$=dimension of each solution
            Update dragonfly velocity( $\Delta X_{t+1}$) according to equation (10)
            Calculate the probability of changing position ( $T(\Delta X)$ ) using equation (13)
            Update $i^{th}$ dragonfly position according to equation (14)
        end for
    end for
    $t = t + 1$
end while //end of initialization
for $i = (1\ to\ n)$ do
    $X_i$ ← initialize the position of $i^{th}$ star with the position of $i^{th}$ dragonfly
end for
while ($t <$ Max_Iter/2) do
    for $i = (1\ to\ n)$ do
        Calculate the $i^{th}$ star's fitness value
        Update positions of the black hole $X_{BH}$ according to fitness value and the number of selected
        features
        Calculate the radius of the event horizon ($R$) using equation (18)
        if $\sqrt{(x_{BH} - X_i)^2} < R$ then
            $i^{th}$ star is swallowed and a new star ($X_{new}$) is generated
        end if
    end for
    for $i = (1\ to\ n)$ do
        for $k = (1\ to\ d)$ do // $d$=dimension of each solution
            Update position of the $i^{th}$ star according to equation (15-17)
        end for
    end for
    $t = t + 1$
end while
Output: the best solution = black hole position
```

3.2.3. Hybrid BDF and BBHA: binary DBH algorithm

In this stage, BBHA is hybridized with the BDFA (DBH) to boost the BBHA's initial population development mechanism in order to create an efficient gene selection algorithm with improved exploration and exploitation capabilities as well as faster convergence. The DBH flowchart in Fig. 3 shows how the individual steps are carried out. Algorithm 3 presents the pseudocode for the proposed approach.

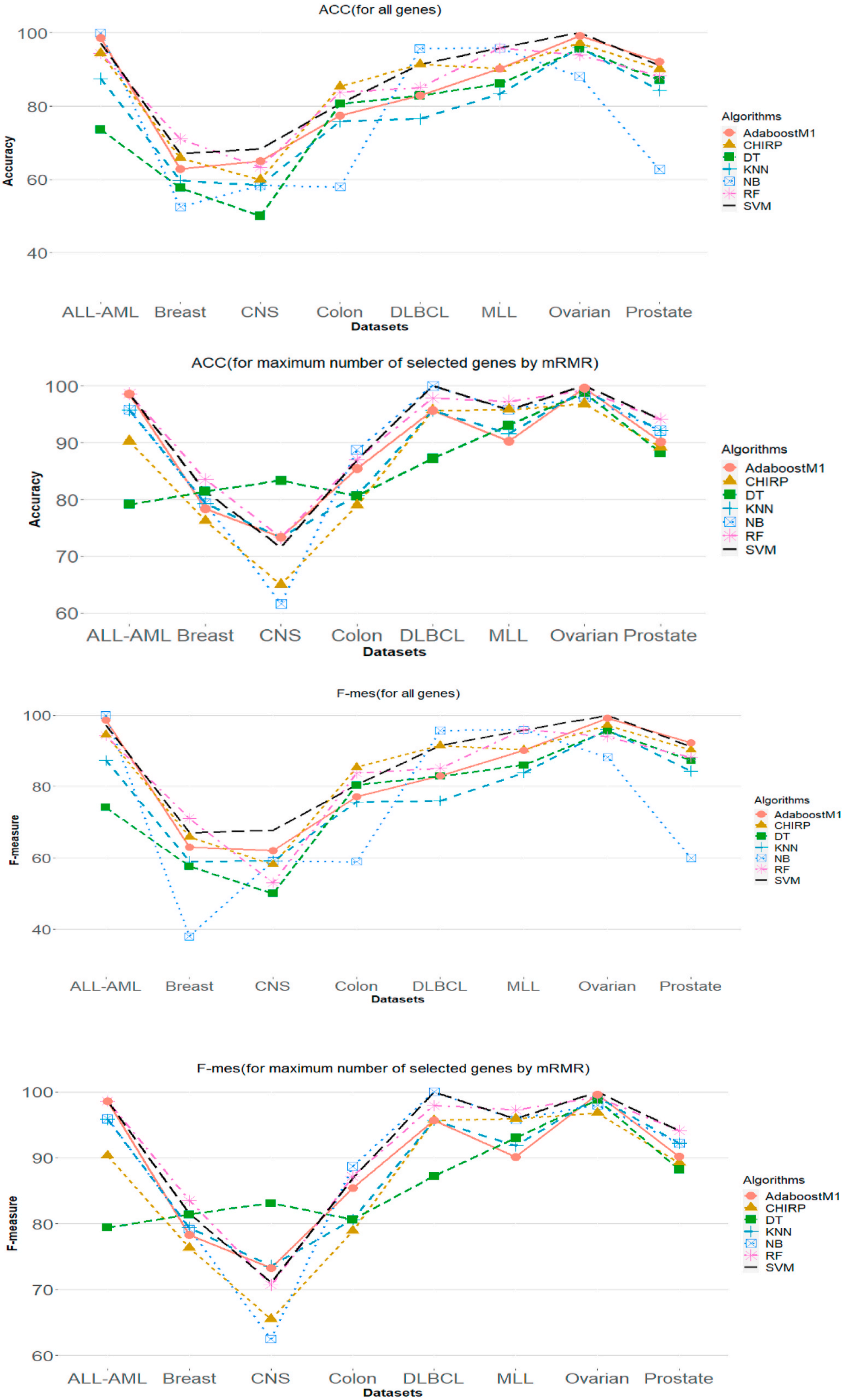
After the first stage, in which the MRMR filter approach selects *n* top-rank genes, **t**he selected genes are fed to the proposed BBHA/BDFA (DBH) algorithm in the second stage. The *I***D* population was created randomly by a binary (0/1) string. . *I* stands for the number of stars in a swarm, and *D* is the dimension of the optimal feature subset. In the DBH algorithm, first, the BDF algorithm is performed for half number of maximum iteration, then BBHA is conducted for the next half of the maximum iteration. It is worth noting that dragonflies' positions at the beginning stage of BDF are initialized randomly, while the positions of the stars in BBHA are initialized by the final dragonflies' positions in the search space (See Fig. 3). In fact, we utilized the BDF approach to optimize the initial positions of the stars in BBHA. After a predefined number of iterations, the best solution is reported by BBHA, and DBH is finished. Lastly, the final solution is used to pick the most discriminative genes. Then SVM classifier with leave one out cross-validation (LOOCV) is utilized to build the model for cancer classification. The following is a summary of this hybrid meta-heuristic algorithm:

Step 1 Generate *I***D* initial population of dragonflies using the optimal feature subset from the MRMR filtering method.

Step 2 Perform BDFA process. Dragonflies begin to move toward the food and update their positions and velocity via Eq. (10)-(14)

Table 3
LOOCV performance of classifiers on 8 microarray datasets with complete genes and maximum selected top-ranked genes.

Dataset	Measure	Classifiers							Classifiers						
Full genes									maximum top genes						
		SVM	**NB**	**KNN**	**DT**	**RF**	**CHIRP**	**AdaboostM1**	**SVM**	**NB**	**KNN**	**DT**	**RF**	**CHIRP**	**AdaboostM1**
Breast	ACC	67	52.5	59.7	57.7	71.1	65.9	62.8	81.4	79.3	79.3	81.4	83.5	76.2	78.3
	MCC	33.8	0.70	18.9	15	42.1	32.6	25.8	62.9	58.8	58.7	62.8	67	52.4	56.6
	F-mes	67	38	59	57.6	71.1	65.9	62.9	81.5	79.2	79.4	81.4	83.5	76.3	78.3
	AUC	66.9	50.	59.1	48	77.1	66.3	67.3	81.5	87.4	79.3	85.2	91.2	76.2	87.2
Colon	ACC	80.6	58	75.8	80.6	83.8	85.4	77.4	87	88.7	80.6	80.6	87	79	85.4
	MCC	57	19.9	46.6	57	64.2	68	49.8	71.4	75.1	56.5	57.7	72.6	53.8	68
	F-mes	80.4	58.9	75.7	80.4	83.7	85.4	77.2	86.9	88.7	80.6	80.6	87.2	78.9	85.4
	AUC	77.8	64.8	73.1	66.5	86.6	83.6	85.3	84.9	91.8	76.8	83.7	91.5	76.6	91.5
DLBCL	ACC	91.4	95.7	76.5	82.9	85.1	91.4	82.9	100	100	95.7	87.2	97.8	95.7	95.7
	MCC	83.3	91.5	55.4	66.1	70.2	83	65.9	100	100	91.8	74.7	95.8	91.5	91.8
	F-mes	91.5	95.7	76	82.9	85.1	91.5	83	100	100	95.7	87.2	97.9	95.7	95.7
	AUC	91.6	98.8	76.3	76.6	93.5	91.5	93.8	100	100	95.8	82.6	100	95.7	99.8
Prostate	ACC	91.1	62.7	84.3	87.2	88.2	90.1	92.1	94.1	92.1	92.1	88.2	94.1	89.2	90.1
	MCC	82.4	28.8	69.2	74.7	76.5	80.5	84.3	88.2	84.4	84.4	76.5	88.5	78.4	80.6
	F-mes	91.2	59.9	84.3	87.3	88.2	90.2	92.2	94.1	92.2	92.2	88.2	94.1	89.2	90.2
	AUC	91.2	62.7	84.4	97.6	91.2	90.2	93.2	94.1	92.8	92.1	91.6	96.8	89.2	96.6
MLL	ACC	95.8	95.8	83.3	86.1	95.8	90.2	90.2	95.8	95.8	91.6	93	97.2	95.8	90.2
	MCC	93.8	93.7	76.7	78.8	93.9	85.7	85.4	93.7	93.9	88	89.4	95.8	94.2	85.4
	F-mes	95.8	95.9	83.8	86	95.9	90.3	90.1	95.9	95.8	91.8	93	97.2	95.9	90.1
	AUC	96.9	96.8	88.1	87.4	98.9	92.6	97.4	96.8	98.3	93.9	93.6	99.9	97.1	97.4
Ovarian	ACC	100	88.1	96	95.6	94	97.2	99.2	100	98	99.2	98.8	99.2	96.8	99.6
	MCC	100	74.5	91.4	90.5	87.1	94	98.3	100	95.7	98.3	97.5	98.3	93.1	99.1
	F-mes	100	88.2	96	95.6	94	97.2	99.2	100	98	99.2	98.8	99.2	96.8	99.6
	AUC	100	93	95.5	95.7	99.5	96.4	100	100	98.6	98.9	99.7	99.9	96.1	100
CNS	ACC	68.3	58.3	58.3	50	63.3	60	65	71.6	61.6	73.3	83.3	73.3	65	73.3
	MCC	28.2	13.4	13.4	−9.9	0	6.1	15.7	35.8	24.5	42.8	62.7	37.5	25.7	41.4
	F-mes	67.7	59.1	59.1	50	53	58.1	62	71.1	62.5	73.6	83.1	70.7	65.5	73.3
	AUC	63.6	58.4	57	49.8	60.4	52.7	62.8	67.2	63.8	71.8	83.2	86.3	63.2	80.7
ALL-AML	ACC	97.2	100	87.5	73.6	94.4	94.4	98.6	98.6	95.8	95.8	79.1	98.6	90.2	98.6
	MCC	93.9	100	72	44.9	88	88.1	97	97	90.9	90.8	55.5	97	78.8	97
	F-mes	97.2	100	87.3	74.1	94.3	94.5	98.6	98.6	95.9	95.8	79.4	98.6	90.3	98.6
	AUC	96.9	100	84.8	69.4	97.6	94.8	99.8	98	99.1	94.9	85.1	100	89.7	99.8



ACC(for all genes)

ACC(for maximum number of selected genes by mRMR)

F-mes(for all genes)

F-mes(for maximum number of selected genes by mRMR)

Fig. 5. The classification performance of seven classifiers on datasets with complete genes and maximum selected top-ranked genes in terms of Accuracy and F-measure.

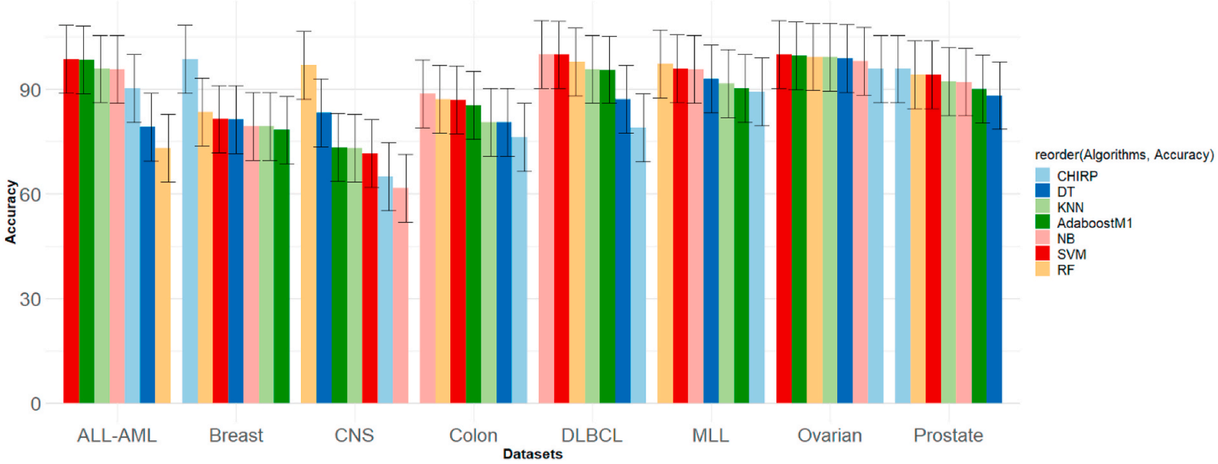


Fig. 6. The bar plot with error bars for the average classification accuracy of seven classifiers on 8 microarray datasets.

Table 4
Comparison of the proposed method with basic BBHA and BDF regarding classification accuracy (in %), number of selected genes, and execution time using SVM classifier.

Dataset	Performance		BBHA	BDF	proposed	Dataset	Performance		BBHA	BDF	proposed
Breast	ACC	*Best*	92.77	90.75	**94.77**	**DLBCL**	ACC	*Best*	100	100	100
		AVG	88.4	85.63	**90.21**			*AVG*	100	100	100
	#G	*Best*	12	23	**11**		#G	*Best*	5	7	**3**
		AVG	**13**	25	14			*AVG*	7	10	**4.05**
	CPU time		**159**	195.92	171		CPU time		218.66	334.79	**209.32**
Colon	ACC	*Best*	98.57	98.57	98.571	**Prostate**	ACC	*Best*	99.09	98.18	99.09
		AVG	95.88	96.29	**97.02**			*AVG*	98.13	97.09	**98.19**
	#G	*Best*	**6**	21	8		#G	*Best*	14	35	**12**
		AVG	**7.66**	22.33	12			*AVG*	**26.8**	42	28
	CPU time		199.3	221.18	**194.3**		CPU time		206.14	252.77	**200.1**
MLL	ACC	*Best*	100	100	100	**Ovarian**	ACC	*Best*	100	100	100
		AVG	100	100	100			*AVG*	100	100	100
	#G	*Best*	**5**	10	**5**		#G	*Best*	3	9	**2**
		AVG	5.3	12	**5.25**			*AVG*	3.2	8.33	**2.66**
	CPU time		**186.1**	248.64	208		CPU time		**157.69**	306.33	280.1
CNS	ACC	*Best*	96	95	**98.57**	**ALL-AML**	ACC	*Best*	100	100	100
		AVG	94.32	90.6	**97.19**			*AVG*	97.79	99.08	**100**
	#G	*Best*	29	54	**22**		#G	*Best*	**2**	33	3
		AVG	**33**	57.33	39.75			*AVG*	4.33	36	**4**
	CPU time		317.66	392.53	**271.2**		CPU time		238.19	196.65	**168.23**

Step 3 Check termination. If the termination condition is satisfied (maximum number of iterations = 10), go to Step 4. Otherwise, go to Step 2.
Step 4 The positions of the optimized dragonflies are passed to the BBHA algorithm as the stars' initialized population.
Step 5 Compute fitness values of all-stars and update the position of the black hole X_{BH} .
Step 6 Perform BBHA operators. Each star updates its position according to Eqs. 15–17.
Step 7 Judge termination. If satisfied, output the final solution. Otherwise, go to Step 5.
Step 8 Build SVM classifier using the final solution with a LOOCV schema

Example. consider the solution vector X_i, with 8 top-ranked genes X_i = [1,0,1,1,1,0,1,0]; where first, third, fourth, fifth and seventh feature are chosen randomly. On the other hand, second, sixth, and eighth features are not chosen. First binary DFA is performed on X_i and its position is updated according to its velocity iteratively with $X_{new} = \begin{cases} \neg X_{current} & r < T(\Delta X_{new}) \\ X_{current} & r \geq T(\Delta X_{new}) \end{cases}$ equation where $\Delta X_{new} = (sS_i + aA_i + cC_i + fF_i + eE_i) + w\Delta X_{current}$. Then, the BBHA is employed to move X_i toward the best solution using $X_{new} = \begin{cases} 1, & if \quad abs(tanh(X_{new})) > 0.6 \\ 0, & otherwise \end{cases}$ where $X_{new} = X_i(current) + rand\,[X_{BH} - X_i(current)]$.

4. Experimental setup and results

The performance of the proposed method MRMR- BDF-BBHA was evaluated using eight benchmark gene expression datasets and one real sequence read archive (SRA) dataset. Benchmark microarray datasets were taken from the http://csse.szu.edu.cn/staff/zhuzx/Datasets.html and SRA dataset was taken from the gene expression omnibus database (GEO). The algorithms in our experiments were programmed using Java and R. Except MRMR which was implemented using R (praznik package), the remains filter approaches (IG, Relief, and chi-square) were implemented in Java using the Weka tool. BDF, as well as BBHA, were both implemented using R language. The "e1071" library in R was utilized for the SVM classifier.

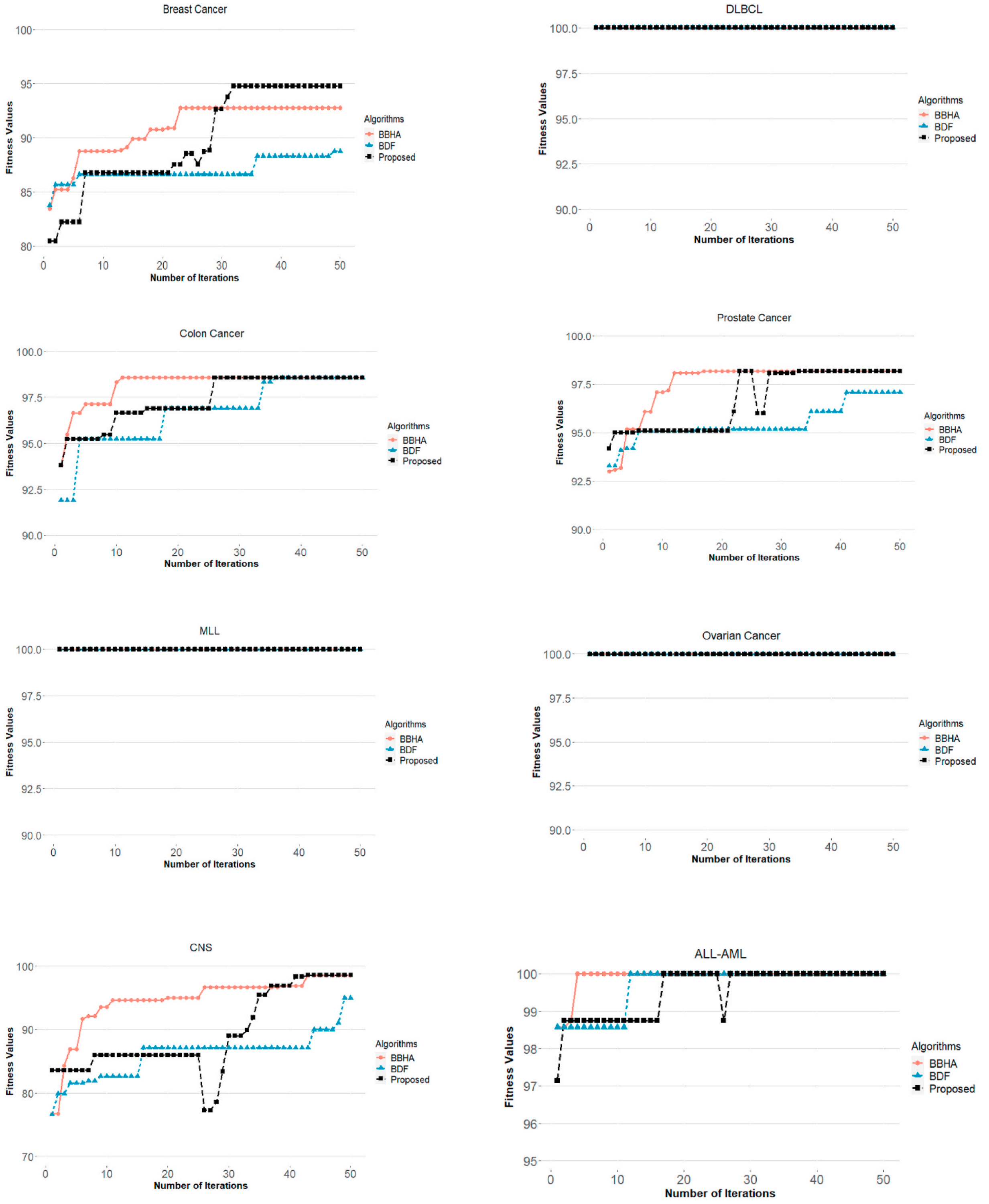


Fig. 7. The convergence behavior of BBHA, BDF, and the proposed DBH approach for a random seed.

4.1. *Dataset used*

The selected benchmark microarray datasets are common in the literature. They are “Breast Cancer”, “Colon Cancer”, Diffuse Large B-cell Lymphoma (DLBCL), “Prostate Tumor”, “Ovarian”, “MLL”, “ALL-AML”, and “CNS”. The selected real SRA dataset is “GSE149273″ which is based on COVID-19 infectious illness. The description of genes, samples, and the number of classes in these datasets are presented in Table 1.

4.2. *Parameter settings*

The parameters setting values for the filter and wrapper phases are assigned based on preliminary experiments. Accordingly, the number of top-ranked genes in MRMR is assigned to 125 for CNS, 100 for Prostate and ALL-AML, and 50 for the remainder of microarray datasets. It is worth mention that before applying MRMR, min-max normalization has been done. In all wrapper approaches the population size and the maximum number of iterations are set to 35 and 50, respectively. The

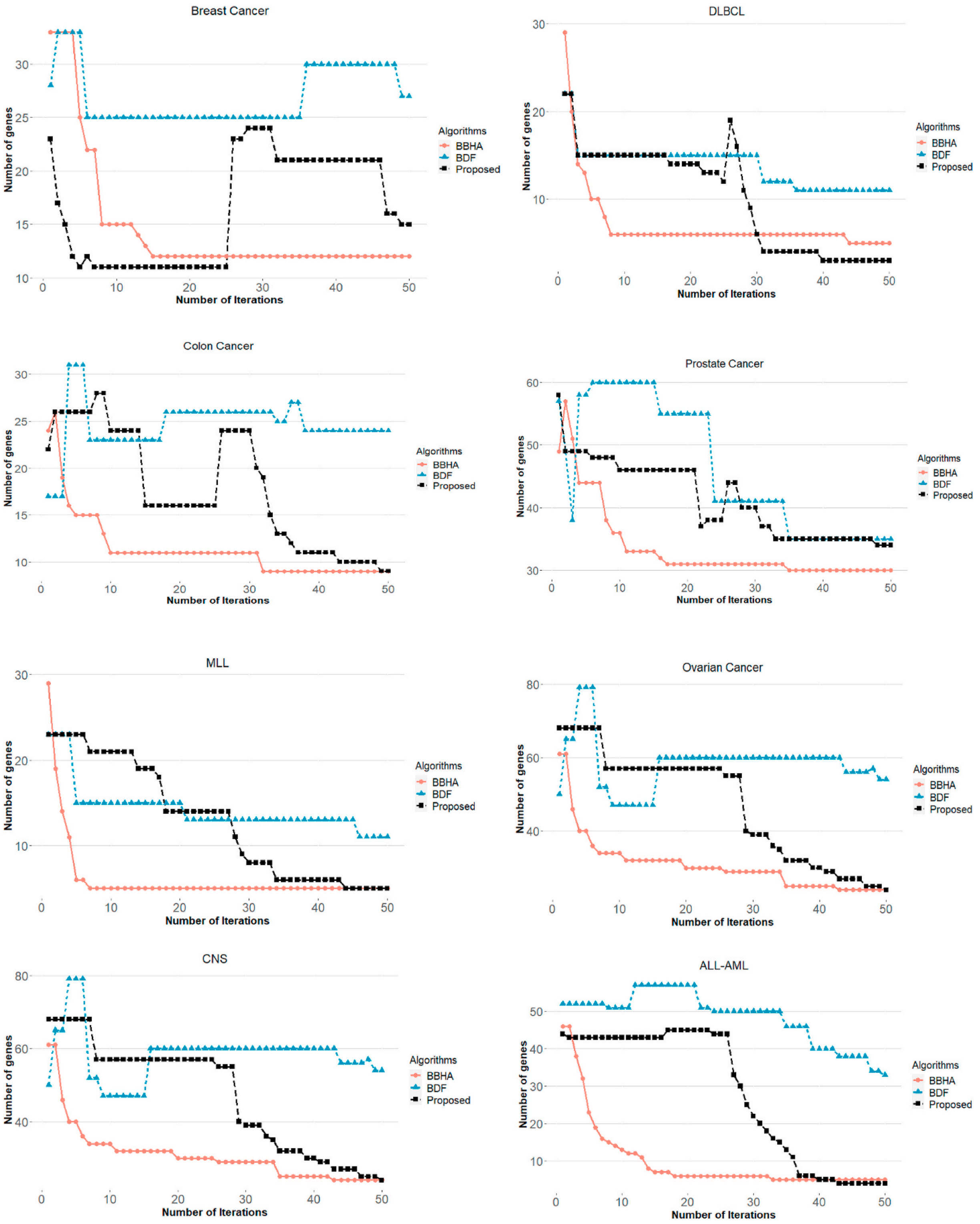


Fig. 8. The number of selected genes during the convergence of BBHA, BDF, and the proposed hybrid approach for a random seed.

SVM of the “e1071” library with the linear kernel is used to evaluate the effectiveness of the proposed feature selection method. The scale and probability parameters of SVM are set to “true” and for tuning the cost parameter of the linear kernel, a traditional grid search algorithm has been conducted. 10-fold cross-validation is used during the algorithm process to estimate the solutions’ prediction accuracy (fitness function).

4.3. Experimental results

First, we investigated the efficacy of four filter methods, including MRMR, IG, Relief, and chi-square, to assess their effects on classification process performance and choose the best one as a pre-processing stage for the wrapper algorithm. Then we selected the top-ranked genes with the highest accuracy that are chosen from the output of the MRMR method to used them as input to the proposed wrapper approach. SVM

Table 5
Average LOOCV classification accuracy ± STD (in %) and the number of genes over 10 runs of the swarm intelligence gene selection methods using SVM classifier.

Dataset	Performance	Bat	ACO	PSO	Firefly	Cuckoo	BBHA	BDF	Proposed
Breast	*ACC*	87.27 ± 0.59	85.21 ± 0.61	87.28 ± 2.37	86.59 ± 1.03	84.97 ± 0.5	88.4 ± 2.1	85.63 ± 1.8	**90.21** ± 2.2
	#G	17.66 ± 3.78	16 ± 5.29	14.66 ± 1.52	16 ± 2.64	14 ± 4.3	**13** ± 2.7	25 ± 3.2	14 ± 1.1
Colon	*ACC*	91.93 ± 1.61	90.31 ± 1.6	91.93 ± 1.61	89.51 ± 1.14	90.85 ± 1.8	95.88 ± 1.4	96.29 ± 0.9	**97.02** ± 0.5
	#G	15.33 ± 3.51	13.66 ± 5.68	9 ± 3	11.33 ± 2.08	**6.33** ± 2.08	7.66 ± 1.5	22.33 ± 1.3	12 ± 2.6
DLBCL	*ACC*	**100** ± 0.0	**100** ± 0.0	**100** ± 0.0	**100** ± 0.0	99.29 ± 1.2	**100** ± 0.0	**100** ± 0.0	**100** ± 0.0
	#G	10.66 ± 1.52	4.66 ± 0.57	4.33 ± 1.52	13.33 ± 0.57	5 ± 1.7	7 ± 1.5	10 ± 3.01	**4.05** ± 1.2
Prostate	*ACC*	97.07 ± 0.98	98.03 ± 0.01	97.70 ± 0.56	97.70 ± 0.56	97.70 ± 0.5	98.13 ± 2.5	97.09 ± 0.5	**98.19** ± 0.3
	#G	29 ± 11	18 ± 1.73	**12** ± 1	30.33 ± 1.59	14.3 ± 5.5	26.8 ± 7.4	42 ± 7.5	28 ± 6.7
MLL	*ACC*	**100** ± 0.0	99.07 ± 0.8	99.53 ± 0.8	**100** ± 0.0	99.07 ± 0.8	**100** ± 0.0	**100** ± 0.0	**100** ± 0.0
	#G	11 ± 2.64	9.66 ± 1.15	7.66 ± 0.57	13 ± 3	8.33 ± 2	5.3 ± 1	12 ± 2.2	**5.25** ± 0.5
Ovarian	*ACC*	**100** ± 0.0	**100** ± 0.0	**100** ± 0.0	**100** ± 0.0	**99.68** ± 0.5	**100** ± 0.0	**100** ± 0.0	**100** ± 0.0
	#G	13 ± 1	3.33 ± 0.57	3.33 ± 0.57	15 ± 0.81	3.33 ± 0.5	3 ± 0.0	8.33 ± 1.15	**2.66** ± 0.57
ALL-AML	*ACC*	99.07 ± 0.80	99.53 ± 0.8	99.53 ± 0.8	**100** ± 0.0	**100** ± 0.0	97.79 ± 1.4	99.08 ± 0.14	**100** ± 0.00
	#G	14.66 ± 5.5	11.66 ± 1.52	9 ± 3.6	26.66 ± 2.51	8 ± 2.64	4.33 ± 2.08	36 ± 3.6	**4** ± 1
CNS	*ACC*	87.21 ± 0.96	90.55 ± 2.54	90 ± 0.01	92.22 ± 3.47	88.33 ± 7.63	94.32 ± 1.2	90.6 ± 4.01	**97.19** ± 1.56
	#G	18.33 ± 2.51	**14.33** ± 0.57	15.25 ± 2.06	45.66±	36.33 ± 8.62	33 ± 3.46	57.33 ± 3.51	39.7 ± 10.01

Table 6
Average rankings of accuracy values among 8 algorithms on eight microarray datasets using Friedman test.

Proposed	2.12
BBHA	3.5
Firefly	4.25
BDF	4.5
PSO	4.62
ACO	5.25
BAT	5.31
Cuckoo	6.43

Table 7
Post-hoc Holm test (0.05).

Comparision	P-values	Result
Proposed vs BBHA	0.26157224	H_0 is not rejected
Proposed vs Firefly	0.08273102	H_0 is not rejected
Proposed vs BDF	0.0524795	H_0 is rejected
Proposed vs PSO	0.04122683	H_0 is rejected
Proposed vs ACO	0.01072444	H_0 is rejected
Proposed vs BAT	0.009252446	H_0 is rejected
Proposed vs Cuckoo	0.0004296932	H_0 is rejected

classifier with LOOCV was used to evaluate the performance of selected genes by filter approaches (Table 2, and Fig. 4). It can be seen from Table 2, and Fig. 4 that compared to the other approaches on most datasets, the MRMR method obtains the highest accuracy on most of the various sub-sets of top-ranked genes, except for the CNS and Prostate Tumor. IG for CNS and Relief for Prostate Tumor have the highest accuracy scores as compared to the other unsupervised filter-based approaches. For the 50 number of genes, MRMR gets significantly higher classification accuracy rates on 5 out of 8 datasets. For example, on the breast cancer dataset, MRMR obtains the highest accuracy compared to other approaches, particularly for 50 and 75 numbers of the top-ranked genes. The Relief method provides the lowest results in this dataset. According to Table 2 and Fig. 4 just 125 top rank genes for CNS, 100 for Prostate and ALL-AML, and 50 for the rest of the microarray datasets are useful for classification. An increase in the number of top-ranked genes degrades the output of classifiers by adding noisy genes.

In the second set of experiment, we compared the classification performance of seven classifiers; SVM, NB, KNN, DT, RF, CHIRP, and Adaboost1, in terms of Accuracy (ACC), F-measure, Area Under Roc Curve (AUC), and Matthew Correlation Coefficient (MCC). Table 3 reports the LOOCV classification performance of the seven classifiers for datasets with complete genes and maximum selected top-ranked genes. Fig. 5 summarize this table for F-measure and ACC measures.

The average classification accuracy of seven classifiers on 8 microarray datasets is shown in Fig. 6. According to the results presented in Table 3, Fig. 5, and Fig. 6, no classifier performs consistently better than others, and the overall performance of NB, RF, and SVM is similar.

Table 4 shows the best and average classification accuracy (ACC) of the fitness function, the best and average optimal subset of genes (#G), and execution time (CPU time) on a random seed for the proposed DBH, basic BDF, and basic BBHA. The best results are highlighted in **bold** font. From Table 4, we can observe that the proposed method is reliable compared to basic BDF and BBHA for gene selection since it has selected the least number of genes with lower execution time and higher accuracy for most datasets.

For basic BBHA, BDF, and proposed hybrid DBH, the convergence behavior on a random seed is plotted and shown in Fig. 7. On all datasets, the convergence behavior trend of suggested DBH is substantially better than BDF. Although suggested DBH performs worse than BBHA in the early stage of evolution for most datasets, proposed DBH can converge better than BBHA in the late stage of evolution. For DLBCL, Ovarian, and MLL the convergence trends of the methods are the same.

Fig. 8 displays the number of selected genes during the convergence of BBHA, BDF, and the proposed DBH for a random seed. From Fig. 8. It can be concluded that whereas the proposed DBH selects more genes than the BBHA for all datasets in the early stage of evolution, the DBH can select a minimum number of genes compared to the BBHA in the late stage of evolution on most datasets.

In the fourth series of experiments, the efficiency of the proposed method was compared to various swarm intelligence gene selection methods in terms of average classification accuracy and the average number of selected genes (Table 5). All methods were executed in 10 independent runs. As shown in Table 5, the proposed method has resulted in higher classification accuracy and a smaller number of selected genes on most datasets. The standard deviation is consistently small for both parameters on most datasets. These facts indicate that the proposed hybrid approach is a fast, consistent, and effective feature selection algorithm.

The non-parametric Friedman test has been carried out to shows whether there exists any statistically significant difference between the proposed approach and other algorithms. The Friedman test assigns average accuracy value rankings to each of the algorithms on eight datasets, which is shown in Table 6. As shown in Table 6, the proposed

Table 8
Comparing the performance of the proposed approach with the literature methods.

Dataset	Method	Accuracy	Reference	Dataset	Method	Accuracy	Reference
Prostate Cancer	Clustering	94.71 (10)	[43]	***Colon Cancer***	TOPSIS-Jaya-NB	97.76 (18.90)	[18]
	IDGA-F-SVM	96.3 (14)	[44]		BDE-X Rankf	75 (3)	[45]
	MOBBA-LS-KNN	97.1 (6)	[46]		DRF0-CFS-SVM	90 (10)	[47]
	BFO	97.42 (29)	[48]		SFS-MB	72.21	[49]
	DRF0-CFS-SVM	97.06 (113)	[47]		RFR-IDGA-RF	93.39 (4.7)	[7]
	PSO dICA	96.77	[50]		PSO dICA	94.73 (20)	[50]
	TLBO-SA-SVM	99.13 (8)	[20]		TLBO-SA-SVM	99.01 (11)	[20]
	Proposed	98.19 ± 0.3 (28 ± 6.7)			RFR-BBHA-Bagging	93.33 (4.5)	[29]
					Proposed	97.02 ± 0.5 (14.4)	
DLBCL	BFO	98.99 (8)	[48]	***Breast cancer***	BDF	86.22 (7237)	[18]
	BDF	89.44	[18]		DRF0-CFS-SVM	84.21 (82)	[47]
	DRF0-CFS-SVM	94.67 (11)	[47]		MIM-AGA-ELM	82.47	[51]
	PSO dICA	94.73 (30)	[50]		Chi-Squared-BBHA-RF	87.77 (6.2)	[36]
	BDE-X Rank	92.9 (3)	[45]		EGS and F-score-AGA-SVM	88.64 (17)	[52]
	Proposed	100 ± 0.0 (4.05 ± 1.2)			MRMR-BA	88.8 (18.3)	[11]
					RFE- PSO-BBHA/SPLSDA	97.72 (12.9)	[38]
					Proposed	90.21 ± 2.2 (12 ± 2.6)	
MLL	RMRMR-HBA-SVM	100 (8)	[3]	***Ovarian Cancer***	RMRMR-HBA-SVM	100 (3.07)	[3]
	HAS-MB	99.55 (6.6)	[53]		HAS-MB	99.81 (5.73)	[53]
	MRMR-BA	79.86 (19.03)	[11]		MRMR-BA	100 (3.83)	[11]
	MBEGA	94.33 (32.1)	[54]		MBEGA	99.71 (9)	[54]
	CFC-iBPSO	100 (30.8)	[14]		CFC-iBPSO	100 (3.3)	[14]
	TOPSIS-Jaya-NB	99.62 (12.90)	[42]		TOPSIS-Jaya-NB	99.52 (18.50)	[42]
	Proposed	100 ± 0.0 (5.25 ± 0.5)			Proposed	100 ± 0.0 (2.66 ± 0.57	
ALL-AML	RMRMR-HBA-SVM	100 (4.07)	[3]	***CNS***	RMRMR-HBA-SVM	100 (11.2)	[3]
	HAS-MB	99.34 (5)	[53]		HAS-MB	84.17 (7.43)	[53]
	MRMR-BA	100 (5.23)	[11]		MRMR-BA	94.22 (19.2)	[11]
	MBEGA	95.89 (12.8)	[54]		MBEGA	72.21 (2.5)	[54]
	CFC-iBPSO	100 (4.3)	[14]		CFC-iBPSO	95.84 (10.5)	[14]
	TOPSIS-Jaya-NB	100 (16.10)	[42]		TOPSIS-Jaya-NB	96.22 (8.7)	[42]
	RFR- IDGA-RF	98.06 (7.4)	[7]		RFR-BBHA-Bagging	90 (3.33)	[29]
					RFE- PSO-BBHA/SPLSDA	99.16 (10.5)	[38]
	Proposed	100 ± 0.0 (4 ± 1)			Proposed	97.19 ± 1.56 (39.7 ± 10.01)	

approach has placed in rank one. In most comparisons, the p-value suggests that the null hypothesis can be rejected. This means that the performance of the proposed approach is statistically significant compared to most of the methods. The findings obtained from the “post-hoc Holm” test are shown in Table 7.

The performance of the proposed algorithm was compared with some relevant state-of-the-art methods. Table 8 indicates the average accuracy of the predictions and the optimum number of genes obtained by the proposed and other reported literature methods. Out of 8 datasets in 4 of them including DLBCL, MLL, ALL-AML, and Ovarian the proposed approach can achieve perfect accuracy with a lower number of genes than other algorithms. In the prostate cancer dataset, except the TLBO-SA-SVM method [20], the proposed approach achieved higher classification accuracy but with a slightly increasing number of selected genes than other methods. For the Breast cancer dataset, RFE-PSO-BBHA/SPLSDA [38] obtained the highest classification accuracy

Table 9
The identified best subset of genes by the proposed method.

	Gene Index	Gene names	SVM	NB	KNN	DT	RF	CHIRP	AdaboostM1
Breast	3232, 4973, 5509, 6247, 8899, 8910,10889,16629,19074,21911, 24107	NM_020123, Contig47710_RC, NM_002914, NM_021184, NM_006145, NM_013438, AL080059, Contig15714_RC, Contig12419_RC, AB014526, Contig49670_RC	93.8	82.4	68.8	85.1	91.9	78.2	90.1
Colon	249, 262, 286, 1221, 1312, 1582, 1772, 1976	M63391, control, H64489, R62549, M86934, X63629, H08393, K03474	98	93.2	83.9	81.3	92.1	83.6	89.9
DLBCL	1277, 1291, 3226	–	100	98.7	93.7	87	98.1	93.6	96.7
Prostate	1943, 4694, 4823, 5074, 5227, 6105, 7073, 7451, 7515, 7652, 8906, 10130	–	99.6	97.4	93.1	76.3	97	93.1	96.9
MLL	318, 6223, 6841,8703, 8937	31575_f_at, 37933_at, 39749_at, 36997_at, 37710_at	100	99.8	94.9	90.7	99.3	89.7	96.5
Ovarian	2239, 2314	MZ435.85411, MZ465.56916	100	99.9	100	99.4	99.9	97.4	99.9
CNS	360, 444, 1049, 1869, 2032, 2372, 2474, 2537, 2851, 3783, 4008, 4394, 4536, 4779, 5476, 5494, 5563, 6035, 6345, 6369, 6534, 6609	D31763_at, D45399_at, J00124_at, M26679_at, M55593_at, M93426_at, S71824_at, S81914_at, U17989_at, U79290_at, V00563_at, X63597_at, X74295_at, X90857_at, X70811_at, X92368_at, D13631_s_at, Z31560_s_at, M63438_s_at, X62083_s_at, U43203_s_at, Z70276_s_at	95.4	67.6	65.9	79.7	86.8	65.8	86.7
ALL-AML	3433, 4211, 4847	U57721_at, X51521_at, X95735_at	100	100	100	96	99.4	93.9	99.6

Table 10
Extracted rules by the PART classifier for identified biomarkers.

Datasets	PART rules	AUC

Breast	**IF** (NM_013438 ≤ 0.148) **AND** (AL080059 ≤ −0.199) **AND** (Contig49670_RC > −0.173) => class = non-relapse (35.0)
	IF (NM_020123 ≤ 0.071) **AND** (Contig49670_RC > −0.334) **AND** (Contig49670_RC ≤ −0.104) => class = relapse (21.0/2.0)
	IF (NM_020123 > 0.071) => class = relapse (18.0)
	IF (NM_020123 > −0.023) **AND** (NM_020123 > 0.006) **AND** (NM_006145 > −0.084) => class = non-relapse (8.0/1.0)
	IF (NM_020123 > −0.023) => class = relapse (9.0/1.0)
	=> class = non-relapse (6.0)
		73.5

Colon	**IF** (M63391 ≤ 1627.27) **AND** (H08393 > 61.59875) => class = negative (34.0) => class = positive (28.0/6.0)	76.1

DLBCL	**IF** (Gene1291<=0.43) => class = activated (21.0)
	IF (Gene1277 > 0.45) => class = germinal (22.0/1.0)
	=> class = germinal (4.0/1.0)
		86.2

Prostate	**IF** (Gene4823 ≤ 70) **AND** (Gene7652 > 441) => class = Normal (36.0)
	IF (Gene10130 ≥ 129) **AND** (Gene7515 ≤ 11) => class = Tumor (52.0/2.0)
	=> class = Normal (14.0/2.0)
		81

MLL	**IF (**37710_at ≤ 1071 **AND** 37933_at<=531) => class = AML (28.0/1.0)
	IF (31575_f_at<=216) => class = ALL (24.0)
	=> class = MLL (20.0/1.0)
		87.2

Ovarian	**IF (**MZ435.85411 > 0.345977 **AND** MZ465.56916 ≤ 0.411917) => class = Cancer (138.0)
	IF (MZ465.56916 > 0.237635 **AND** MZ435.85411 ≤ 0.495118 **AND** MZ465.56916 > 0.260951) => class = Normal (86.0)
	IF (MZ435.85411 > 0.160364) => class = Cancer (24.0)
	=> class = Normal (5.0)
		99.7

CNS	**IF (**S71824_at > 420) **AND** (M26679_at <= −12) => class = 0 (22.0)
	IF (M55593_at ≤ 1274) **AND** (U79290_at ≤ 110) **AND** (X74295_at<=747) => class = 1 (18.0)
	IF (M26679_at > −7) => class = 0 (18.0/1.0)
	=> class = 1 (2.0)
		67.2

ALL-AML	**IF (**X95735_at<=938) => class = ALL (45.0/1.0)
	IF (X51521_at<=4253) => class = AML (24.0)
	=> class = ALL (2.0)
		96

with only 12.9 average number of genes. By contrast, our proposed method selected 12 genes and achieves (90.21) classification accuracy. In the Colon cancer dataset, the proposed approach obtained better performance than the previous studies except for TLBO-SA-SVM [20] and TOPSIS-Jaya-NB [42]. For the Colon dataset, the TOPSIS-Jaya-NB method obtained 97.76 classification accuracy with 18.90 genes, while the proposed approach achieves 97.02 classification accuracy with only 14.4 genes.

In sum, the experimental results exhibit that the proposed approach can yield a smaller set of reliable genes with higher/equivalent classification accuracy than other methods on most datasets.

Table 9 shows the best subset of gens obtained by using the proposed method in each dataset. As can be seen in Table 9, all of the classifiers achieve a high area under the roc curve (AUC) on the optimal subset of gens. Table 10 shows the relationships between identified genes in each dataset using a rule-based classifier (PART). Fig. 9 displays a heatmap created for the identified best subset of gens. The heatmap correctly clusters samples and reorder the genes into blocks with similar expression patterns. The boxplots which show the expression distribution of the identified gens are plotted in Fig. 10.

4.4. GSE149273 (Covid19) dataset

A novel virus which is known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2 or CV2) was identified as the cause of coronavirus disease (COVID-19). The CV2 required ACE2 to infect the cells. ACE2 is a protein that provides the entry point for CV2 to hook into and infect human cells. In other words, ACE2 serves as a cellular gateway or a receptor for the COVID-19-causing virus. This COVID-19 infectious illness may be more severe in patients with asthma. Rhinovirus (RV) is the respiratory virus that is responsible for the majority of asthma exacerbations in children and adults. An analysis was conducted in Ref. [55] recently to examine the role of RV infection in ACE2 expression of asthmatic patients. It was found that RV infections in asthmatics lead to the overexpression of ACE2, and subsequently activate cytokine pathways that are associated with severe COVID-19 disease.

The GSE149273 is an SRA dataset that was downloaded from the GEO and composed of 25343 genes, 90 samples, and three categories (RVA, RVC, Control). After downloading the count data some preprocessing steps such as (removing lowly expressed genes, converting counts to Differentially Expressed Genes (DGEList), quality control, normalization for composition bias, finding all common genes differentially expressed in (RVA, control) and (RVC, Control), extracting ACE2 gene with 90 samples) has been done.

We keep genes that have a count per million (CPM) of 0.2 or more in at least 12 samples. The number of highly expressed genes after filtering is 17940. In the next step, we draw a multi-dimensional scaling (MDS) plot to display similarities and dissimilarities of samples in an unsupervised manner (Fig. 11(A)). Before fitting the linear model we did quality control and draw a Voom plot to check there is a need to filter more genes or not. Fig. 11(B) shows a “good” voom plot. Then we did normalization and extract genes that are DE in multiple comparisons. Fig. 11(C) is a Venn diagram that shows the number of DE genes in the comparison between controlv versus RVA, control versus RVC, and the number of genes that are DE in both comparisons (center). 4055 number of common DE genes were extracted.

The performance of different classifiers (KNN, NB, SVM, DT, RF, CHIRP, and AdaboostM1) for 4055 common DE genes and ACE2 are shown in Table 11. Compared to all classifiers the SVM classifier obtained the highest accuracy for the ACE2 gene and the second-highest for common DE genes.

Here, we aim to build a model to improve the predictive efficiency of ACE2 in Covid 19 diagnostics. By applying the proposed approach (MRMR-BDF-BBHA-SVM) to 4055 common DE genes, we find the best subset of 9 genes with an SVM accuracy of 83.33%. This indicates that among 90 samples, only 15 of them have been misclassified (Table 12).

Table 13 shows the extracted rule by the PART classifier for the proposed model. Fig. 11 (D) displays a heatmap created for the best subset of genes. From the heat map, we observe that the expressions of **ACE2** and **HTR2B** are very similar.

Fig. 11(E) shows the expression distribution of genes. The relation of

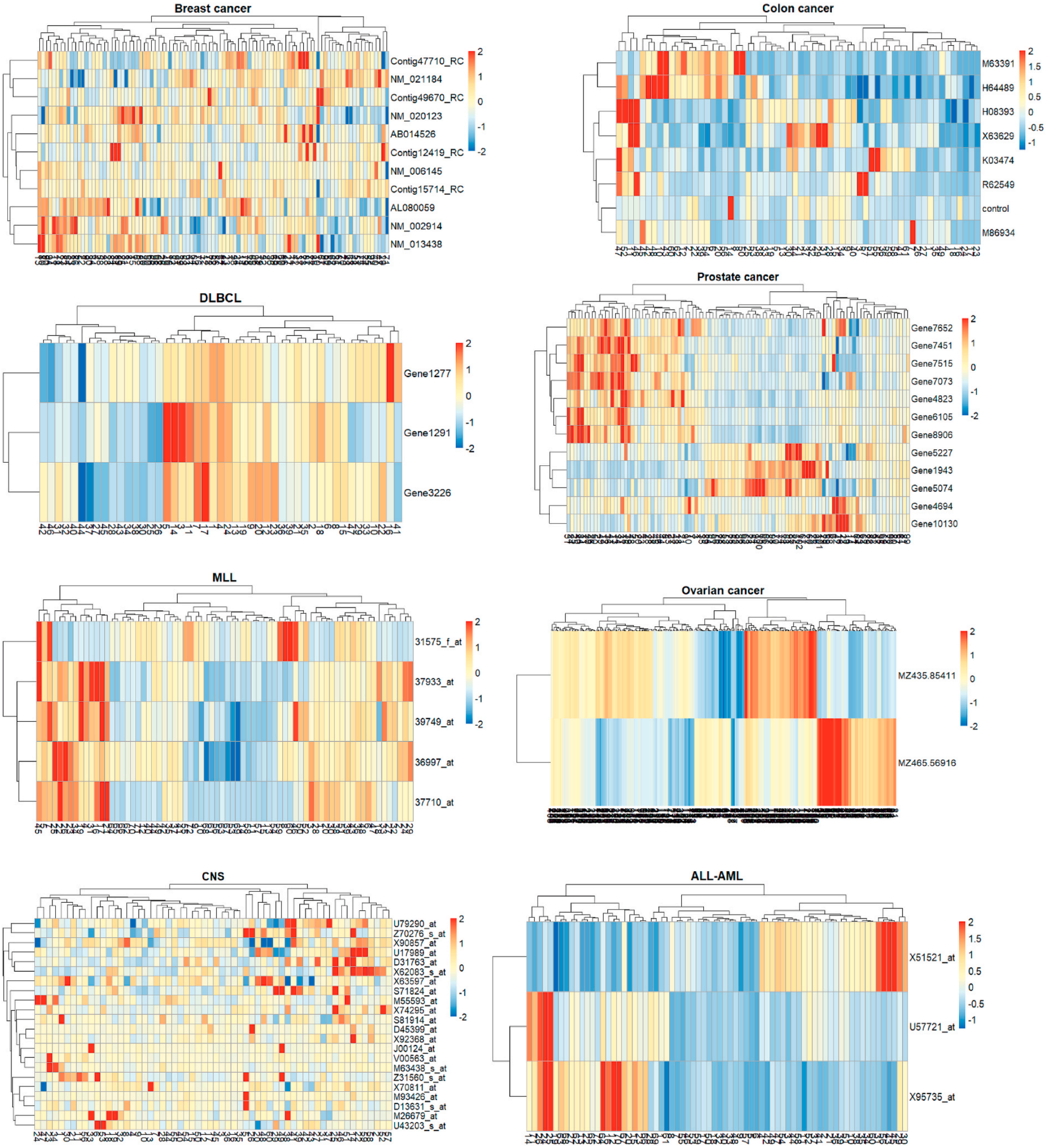


Fig. 9. The heat map of the actual expression profiles for the best subset of genes obtained from the proposed method. The heat map is generated using the “pheatmap” package in R.

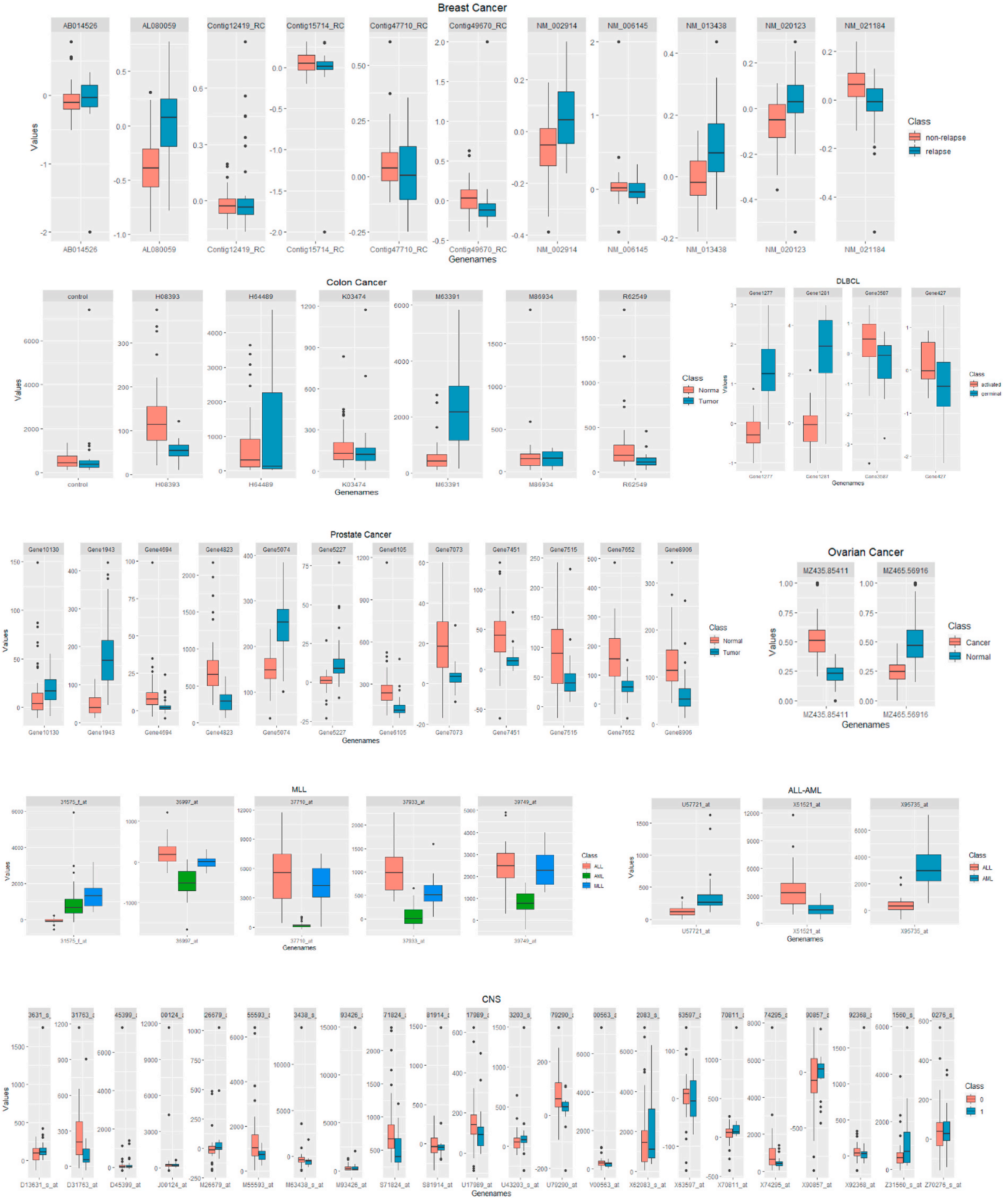


Fig. 10. Box plot for best subset of genes identified by proposed approach in each dataset.

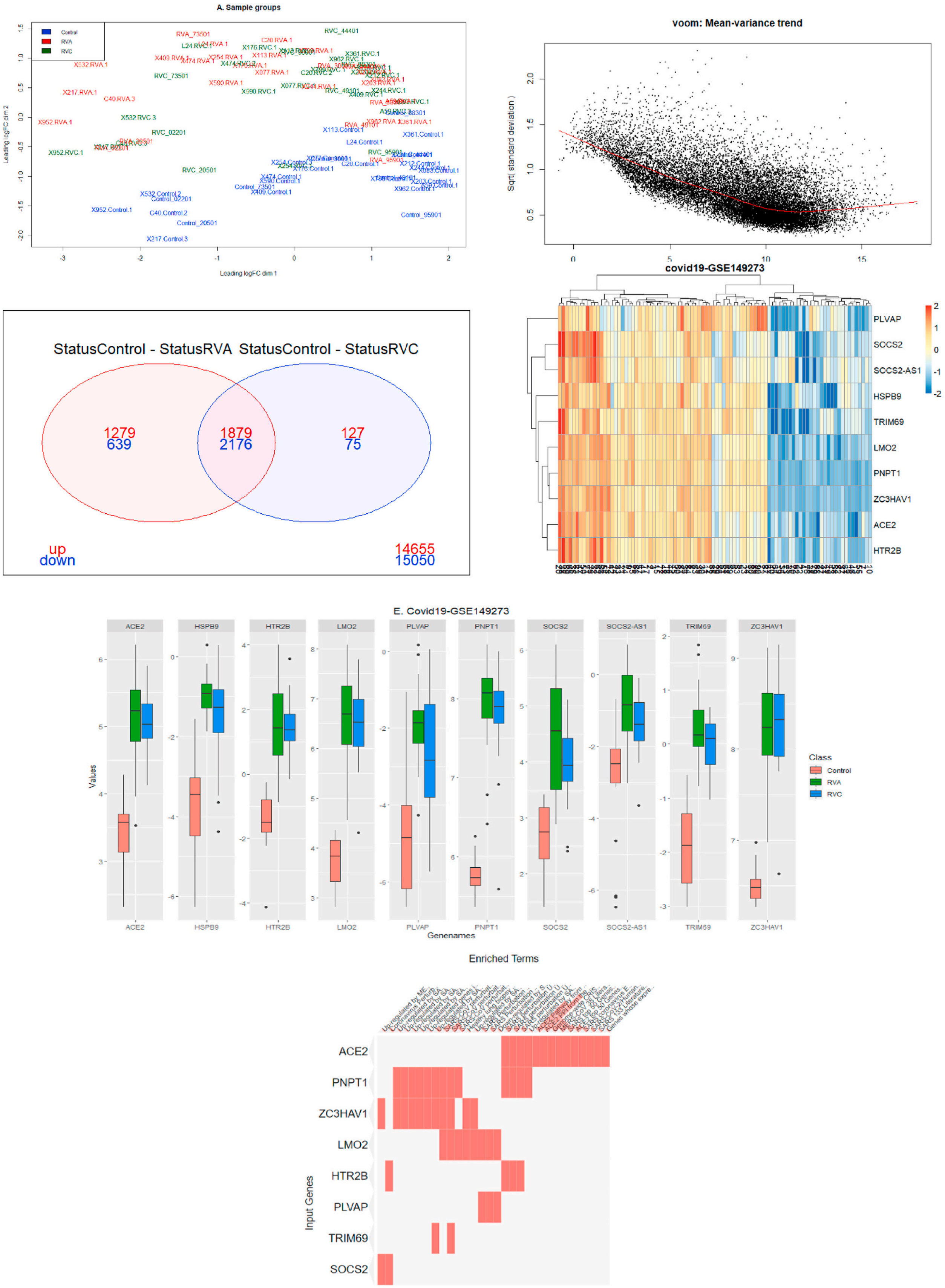


Fig. 11. A. MDS plot of log-CPM values over dimensions 1 and 2 with samples colored by categories. **B**. Mean-variance plot of each gene, **C**. Venn diagram, **D**. heat map, **E**. boxplot, **F**. Enrichment analysis for COVID-19 related gene sets.

Table 11
LOOCV classification accuracy of different classifiers for common DE genes and ACE2.

Dataset	For 4055 common DE genes							ACE2 gene						
	Algorithms							Algorithms						
	SVM	NB	KNN	DT	RF	CHIRP	Adaboost	SVM	NB	KNN	DT	RF	CHIRP	Adaboost
Covid19-GSE149273	71.11	64.44	42.22	60	60	62.22	94.44	66.66	62.22	56.66	58.88	56.66	33.33	60

Table 12
The proposed prediction model for Covid19 identification.

	Gene ındex	Gene name	SVM (c = 10	NB	KNN	DT	RF	CHIRP	AdaboostM1
Covid19-GSE149273	17, 46, 92, 126, 168, 177, 304, 460, 472, 875	PNPT1, ZC3HAV1, LMO2, TRIM69, **ACE2**, HSPB9, HTR2B, SOCS2, PLVAP, SOCS2-AS1	83.33	78.88	72.22	66.66	78.88	72.22	77.77

Table 13
Extracted rule by PART classifier for proposed model.

Datasets	PART rules	AUC

Covid19-GSE149273	**IF** (PNPT1 ≤ 6.257342) => class = Control (31.0/1.0)
	IF (SOCS2 ≤ 5.11131 **AND ACE2** > 4.198649 **AND** TRIM69 > −0.255926 **AND** PLVAP > −3.388202 **AND** RIM69 > 0.201068) => class = RVC (14.0/5.0)
	IF (PLVAP > −3.3535 AND TRIM69 > −0.255926) => class = RVA (21.0/1.0)
	IF (**ACE2** > 4.198649) => class = RVC (21.0/2.0) class = RVA (3.0)
		80.5

‘ZC3HAV1’ with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was already identified in previous studies [56,57]. Fig. 11 (F) displays the Enrichment analysis that was conducted using the Enricher web tool (https://maayanlab.cloud/Enrichr). Fig. 11 (F) shows whether or not a gene is associated with the SARS-CoV-2 term. ACE2, PNPT1, ZC3HAV1, and LMO2 are four genes to be linked to SARS-CoV-2 which have been confirmed by enrichment analysis.

5. Conclusion

Due to the curse of the dimensionality problem of microarrays, predicting strongly discriminative genes in gene expression data is a difficult task and most of the current approaches fail to deal with it effectively. This study suggests a new hybrid approach that combines BDF with BBHA, called DBH, to find the most informative genes for disease classification and diagnosis. The primary purpose of hybridization is to improve BBHA performance for a wide exploration and a deep exploitation search using the BDF algorithm. As the BDF is more firm and can easily be combined effectively with other NIOA, its efficiency has been integrated with the excellent performance of BBHA in avoiding local optima and its high convergence speed.

The proposed method consists of two stages. In the first stages, the MRMR filter method is employed to choose the top 150 features, and in the second stage, the reduced gene subset is used as an input to the DBH algorithm to extract the most discriminative genes. In the suggested approach, the SVM classifier was applied as the classification model. The efficiency of the suggested approach is tested on eight well-known benchmark microarray datasets and one real COVID-19 related gene expression dataset. According to the obtained results, the suggested MRMR- DBH approach has shown remarkable performance in terms of convergence rate, classification accuracy, and the optimal number of genes, as compared to other considered approaches. The suggested hybrid algorithm also incorporates stability and robustness into the solution. Moreover, in the RNA-Seq COVID-19 related gene expression dataset, the suggested method achieves more than 83.33% classification accuracy, which verifies that it is an effective approach in practice for selecting the most biologically relevant genes related to the disease.

As future work, the suggested hybrid approach can be extended and applied to solve various real and complex optimization problems of different domains. For Example, DF/BHA performance can be examined in numerical function optimization, image contrast enhancement, image segmentation, and text mining. It also would be interesting to hybridize the BBHA algorithm with other NIOA such as the salp optimization algorithm using many classifiers other than SVM which was adopted in this paper. Additionally, different combinations of the present method can be developed through new filtering feature selection methods to improve the classification accuracy of the BDF/BBHA method.

Author contribution statements

Elnaz Pashaei and Elham Pashaei designed the model and the computational framework. Both carried out the implementation and performed the experiment and wrote the manuscript.

Declaration of competing interest

The authors declare that they have no conflict of interest.

References

[1] M.A. Al-Betar, O.A. Alomari, S.M. Abu-Romman, A TRIZ-inspired bat algorithm for gene selection in cancer classification, Genomics 112 (2020) 114–126, https://doi.org/10.1016/j.ygeno.2019.09.015.

[2] J. Li, H. Kang, G. Sun, T. Feng, W. Li, W. Zhang, B. Ji, IBDA: improved binary dragonfly algorithm with evolutionary population dynamics and adaptive crossover for feature selection, IEEE Access 8 (2020) 108032–108051, https://doi.org/10.1109/ACCESS.2020.3001204.

[3] O.A. Alomari, A.T. Khader, M.A. Al-Betar, M.A. Awadallah, A novel gene selection method using modified MRMR and hybrid bat-inspired algorithm with β-hill climbing, Appl. Intell. 48 (2018) 4429–4447, https://doi.org/10.1007/s10489-018-1207-1.

[4] L. Gao, M. Ye, X. Lu, D. Huang, Hybrid method based on information gain and support vector machine for gene selection in cancer classification, Genom., Proteom. Bioinforma 15 (2017) 389–395, https://doi.org/10.1016/j.gpb.2017.08.002.

[5] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: introduction and review, J. Biomed. Inf. 85 (2018) 189–203, https://doi.org/10.1016/j.jbi.2018.07.014.

[6] T. Almutiri, F. Saeed, Chi square and support vector machine with recursive feature elimination for gene expression data classification, in: 2019 1st Int. Conf. Intell. Comput. Eng. Towar. Intell. Solut. Dev. Empower. Our Soc. ICOICE 2019, Institute of Electrical and Electronics Engineers Inc., 2019, https://doi.org/10.1109/ICOICE48418.2019.9035165.

[7] E. Pashaei, E. Pashaei, Gene selection using intelligent dynamic genetic algorithm and random forest, 2019 11th Int. Conf. Electr. Electron. Eng. (2019) 470–474, https://doi.org/10.23919/ELECO47770.2019.8990557.

[8] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, U. Maulik, Recursive Memetic Algorithm for gene selection in microarray data, Expert Syst. Appl. 116 (2019) 172–185, https://doi.org/10.1016/j.eswa.2018.06.057.

[9] A. Bir-Jmel, S.M. Douiri, S. Elbernoussi, Gene selection via a new hybrid ant colony optimization algorithm for cancer classification in high-dimensional data, Comput. Math. Methods Med. 2019 (2019), https://doi.org/10.1155/2019/7828590.

[10] O.A. Alomari, A.T. Khader, M.A. Al-Betar, L.M. Abualigah, MRMR BA: a hybrid gene selection algorithm for cancer classification, J. Theor. Appl. Inf. Technol. 95 (2017) 2610–2618.

[11] O. Ahmad Alomari, A. Tajudin Khader, M. Azmi Al-Betar, L. Mohammad Abualigah, Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm, Int. J. Data Min. Bioinf. 19 (2017) 32–51, https://doi.org/10.1504/IJDMB.2017.088538.

[12] O.A. Alomari, A.T. Khader, M.A. Al-Betar, Z.A. Alkareem Alyasseri, A hybrid filter-wrapper gene selection method for cancer classification, 2nd Int. Conf. BioSignal Anal. Process. Syst. ICBAPS 2018 (2018) 113–118, https://doi.org/10.1109/ICBAPS.2018.8527392.

[13] H.M. Alshamlan, Co-ABC: correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile, Saudi J. Biol. Sci. 25 (2018) 895–903, https://doi.org/10.1016/j.sjbs.2017.12.012.

[14] I. Jain, V.K. Jain, R. Jain, Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification, Appl. Soft Comput. J. 62 (2018) 203–215, https://doi.org/10.1016/j.asoc.2017.09.038.

[15] R. Dash, An adaptive harmony search approach for gene selection and classification of high dimensional medical data, J. King Saud Univ. - Comput. Inf. Sci. 33 (2021) 195–207, https://doi.org/10.1016/j.jksuci.2018.02.013.

[16] P. Tumuluru, B. Ravi, Chronological grasshopper optimization algorithm-based gene selection and cancer classification, J. Adv. Res. Dyn. Control Syst. 10 (2018) 80–94.

[17] A.K. Shukla, P. Singh, M. Vardhan, An adaptive inertia weight teaching-learning-based optimization algorithm and its applications, Appl. Math. Model. 77 (2020) 309–326, https://doi.org/10.1016/j.apm.2019.07.046.

[18] S.A. Medjahed, T.A. Saadi, A. Benyettou, M. Ouali, Kernel-based learning and feature selection analysis for cancer diagnosis, Appl. Soft Comput. J. 51 (2017) 39–48, https://doi.org/10.1016/j.asoc.2016.12.010.

[19] S. Tabakhi, P. Moradi, Relevance-redundancy feature selection based on ant colony optimization, Pattern Recogn. 48 (2015) 2798–2811, https://doi.org/10.1016/j.patcog.2015.03.020.

[20] A.K. Shukla, P. Singh, M. Vardhan, A new hybrid wrapper TLBO and SA with SVM approach for gene expression data, Inf. Sci. (Ny) 503 (2019) 238–254, https://doi.org/10.1016/j.ins.2019.06.063.

[21] A.K. Shukla, P. Singh, M. Vardhan, Gene selection for cancer types classification using novel hybrid metaheuristics approach, Swarm Evol. Comput. 54 (2020), 100661, https://doi.org/10.1016/j.swevo.2020.100661.

[22] M. Mafarja, S. Mirjalili, Whale optimization approaches for wrapper feature selection, Appl. Soft Comput. J. 62 (2018) 441–453, https://doi.org/10.1016/j.asoc.2017.11.006.

[23] S. Mirjalili, Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems, Neural Comput. Appl. 27 (2016) 1053–1073, https://doi.org/10.1007/s00521-015-1920-1.

[24] C.M. Rahman, T.A. Rashid, Dragonfly Algorithm and its Applications in Applied Science Survey, 2019, p. 2019, https://doi.org/10.1155/2019/9293617.

[25] Y. Meraihi, A. Ramdane-Cherif, D. Acheli, M. Mahseur, Dragonfly algorithm: a comprehensive review and applications, Neural Comput. Appl. 32 (2020) 16625–16646, https://doi.org/10.1007/s00521-020-04866-y.

[26] M.M. Mafarja, D. Eleyan, I. Jaber, A. Hammouri, S. Mirjalili, Binary dragonfly algorithm for feature selection, in: Proc. - 2017 Int. Conf. New Trends Comput. Sci. ICTCS 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 12–17, https://doi.org/10.1109/ICTCS.2017.43.

[27] A.I. Hammouri, M. Mafarja, M.A. Al-Betar, M.A. Awadallah, I. Abu-Doush, An improved Dragonfly Algorithm for feature selection, Knowl. Base Syst. 203 (2020), 106131, https://doi.org/10.1016/j.knosys.2020.106131.

[28] G.I. Sayed, A. Tharwat, A.E. Hassanien, Chaotic dragonfly algorithm: an improved metaheuristic algorithm for feature selection, Appl. Intell. 49 (2019) 188–205, https://doi.org/10.1007/s10489-018-1261-8.

[29] E. Pashaei, M. Ozen, N. Aydin, Gene selection and classification approach for microarray data based on Random Forest Ranking and BBHA, in: 3rd IEEE EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2016, Institute of Electrical and Electronics Engineers Inc., 2016, pp. 308–311, https://doi.org/10.1109/BHI.2016.7455896.

[30] A. Hatamlou, Black hole: a new heuristic optimization approach for data clustering, Inf. Sci. (Ny) 222 (2013) 175–184, https://doi.org/10.1016/j.ins.2012.08.023.

[31] R. Munoz, R. Olivares, C. Taramasco, R. Villarroel, R. Soto, T.S. Barcelos, E. Merino, M.F. Alonso-Sánchez, Using black hole algorithm to improve EEG-based emotion recognition, Comput. Intell. Neurosci. 2018 (2018), https://doi.org/10.1155/2018/3050214.

[32] H.A. Abdulwahab, A. Noraziah, A.A. Alsewari, S.Q. Salih, An enhanced version of black hole algorithm via Levy flight for optimization and data clustering problems, IEEE Access 7 (2019) 142085–142096, https://doi.org/10.1109/access.2019.2937021.

[33] E. Pashaei, E. Pashaei, Training feedforward neural network using enhanced black hole algorithm: a case study on COVID-19 related ACE2 gene expression classification, Arabian J. Sci. Eng. 46 (2021) 3807–3828, https://doi.org/10.1007/s13369-020-05217-8.

[34] W. Xie, J.S. Wang, Y. Tao, Improved black hole algorithm based on golden sine operator and Levy flight operator, IEEE Access 7 (2019) 161459–161486, https://doi.org/10.1109/ACCESS.2019.2951716.

[35] J.S. Pan, Q.W. Chai, S.C. Chu, N. Wu, 3-D terrain node coverage of wireless sensor network using enhanced black hole algorithm, Sensors 20 (2020) 2411, https://doi.org/10.3390/s20082411.

[36] E. Pashaei, N. Aydin, Binary black hole algorithm for feature selection and classification on biological data, Appl. Soft Comput. J. 56 (2017) 94–106, https://doi.org/10.1016/j.asoc.2017.03.002.

[37] O.S. Qasim, N.A. Al-Thanoon, Z.Y. Algamal, Feature selection based on chaotic binary black hole algorithm for data classification, Chemometr. Intell. Lab. Syst. 204 (2020), 104104, https://doi.org/10.1016/j.chemolab.2020.104104.

[38] E. Pashaei, E. Pashaei, N. Aydin, Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization, Genomics 111 (2019) 669–686, https://doi.org/10.1016/j.ygeno.2018.04.004.

[39] M. Radovic, M. Ghalwash, N. Filipovic, Z. Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC Bioinf. 18 (2017) 1–14, https://doi.org/10.1186/s12859-016-1423-9.

[40] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, M. Lang, Benchmark for filter methods for feature selection in high-dimensional classification data, Comput. Stat. Data Anal. 143 (2020), 106839, https://doi.org/10.1016/j.csda.2019.106839.

[41] A. Statnikov, L. Wang, C.F. Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, BMC Bioinf. 9 (2008) 1–10, https://doi.org/10.1186/1471-2105-9-319.

[42] A. Chaudhuri, T.P. Sahu, A hybrid feature selection method based on Binary Jaya algorithm for micro-array data classification, Comput. Electr. Eng. 90 (2021), 106963, https://doi.org/10.1016/j.compeleceng.2020.106963.

[43] H. Chen, Y. Zhang, I. Gutman, A kernel-based clustering method for gene selection with gene expression data, J. Biomed. Inf. 62 (2016) 12–20, https://doi.org/10.1016/j.jbi.2016.05.007.

[44] M. Dashtban, M. Balafar, Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts, Genomics 109 (2017) 91–107, https://doi.org/10.1016/j.ygeno.2017.01.004.

[45] J. Apolloni, G. Leguizamón, E. Alba, Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments, Appl. Soft Comput. J. 38 (2016) 922–932, https://doi.org/10.1016/j.asoc.2015.10.037.

[46] M. Dashtban, M. Balafar, P. Suravajhala, Gene selection for tumor classification using a novel bio-inspired multi-objective approach, Genomics 110 (2018) 10–17, https://doi.org/10.1016/j.ygeno.2017.07.010.

[47] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Distributed feature selection: an application to microarray data classification, Appl. Soft Comput. J. 30 (2015) 136–150, https://doi.org/10.1016/j.asoc.2015.01.035.

[48] A. Wang, N. An, J. Yang, G. Chen, L. Li, G. Alterovitz, Wrapper-based gene selection with Markov blanket, Comput. Biol. Med. 81 (2017) 11–23, https://doi.org/10.1016/j.compbiomed.2016.12.002.

[49] A. Wang, N. An, G. Chen, L. Li, G. Alterovitz, Accelerating wrapper-based feature selection with K-nearest-neighbor, Knowl. Base Syst. 83 (2015) 81–91, https://doi.org/10.1016/j.knosys.2015.03.009.

[50] M. Mollaee, M.H. Moattar, A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification, Biocybern. Biomed. Eng. 36 (2016) 521–529, https://doi.org/10.1016/j.bbe.2016.05.001.

[51] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, Z. Gao, A hybrid feature selection algorithm for gene expression data classification, Neurocomputing 256 (2017) 56–62, https://doi.org/10.1016/j.neucom.2016.07.080.

[52] A.K. Shukla, P. Singh, M. Vardhan, A hybrid gene selection method for microarray recognition, Biocybern. Biomed. Eng. 38 (2018) 975–991, https://doi.org/10.1016/j.bbe.2018.08.004.

[53] S.S. Shreem, S. Abdullah, M.Z.A. Nazri, Hybridising harmony search with a Markov blanket for gene selection problems, Inf. Sci. (Ny) 258 (2014) 108–121, https://doi.org/10.1016/j.ins.2013.10.012.

[54] Z. Zhu, Y.S. Ong, M. Dash, Markov blanket-embedded genetic algorithm for gene selection, Pattern Recogn. 40 (2007) 3236–3248, https://doi.org/10.1016/j.patcog.2007.02.007.

[55] E.H. Chang, A.L. Willis, C.E. Romanoski, D.A. Cusanovich, N. Pouladi, J. Li, Y. A. Lussier, F.D. Martinez, RV infections in asthmatics increase ACE2 expression and cytokine pathways implicated in COVID-19, Am. J. Respir. Crit. Care Med. 202 (2020) 753–755, https://doi.org/10.1164/rccm.202004-1343LE.

[56] R. Nchioua, D. Kmiec, J.A. Müller, C. Conzelmann, R. Groß, C.M. Swanson, S.J. D. Neil, S. Stenger, D. Sauter, J. Münch, K.M.J. Sparrer, F. Kirchhoff, Sars-cov-2 is restricted by zinc finger antiviral protein despite preadaptation to the low-cpg environment in humans, mBio 11 (2020) 1–19, https://doi.org/10.1128/mBio.01930-20.

[57] Y. Wei, J. Silke, P. Aris, X. Xia, Coronavirus genomes carry the signatures of their habitats, BioRxiv 15 (2020), e0244025, https://doi.org/10.1101/2020.06.13.149591.