IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 April 2024, accepted 12 May 2024, date of publication 17 May 2024, date of current version 30 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402652

RESEARCH ARTICLE

Feature Selection of Gene Expression Data Using a Modified Artificial Fish Swarm Algorithm With Population Variation

ZONG-ZHENG LI, FANG-LING WANG, FENG QIN, YUSLIZA BINTI YUSOFF, AND AZLAN MOHD ZAIN, (Member, IEEE)

Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor Bahru 81310, Malaysia

Corresponding author: Azlan Mohd Zain (azlanmz@utm.my)

This work was supported in part by the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS) under Grant FRGS/1/2022/ICT02/UTM/01/1.

ABSTRACT Microarray data is of great significance for cancer identification at the gene level. In the microarray dataset, only a small number of characteristic genomes have significant classification and identification rates for cancer. How to extract a small number of characteristic genes from a large number of microarray data is a classic NP-hard problem. This paper proposes a practical hybrid approach to implement the feature selection of gene expression from the microarray by combining the F-score algorithm and an improved artificial fish swarm algorithm with population variation (FSA-PV). Firstly, the F-score algorithm eliminates a large number of useless and redundant features in the dataset. Then, FSA-PV is discussed to obtain the ability to jump out of the local optimum while retaining the excellent feature of the subset as much as possible, and the adaptive step and visual are used to adjust the search space and to move the range of the algorithm in different environments to improve the local optimization and global optimization abilities. In addition, a naive Bayesian classifier is used to test the classification accuracy of subsets. Eight classical datasets are used to verify the performance of the proposed mechanism in the experiment part. The results reveal that the classification accuracy using the FSA-PV is significant superior to other algorithms in Breast dataset, and the classification accuracy is more than 90% in 8 cases. It further indicates the robustness and feasibility of the FSA-PV in the gene selection process.

INDEX TERMS Feature selection, gene expression, microarray data, modified artificial fish swarm algorithm, population variation.

I. INTRODUCTION

Microarray technology is widely used to determine the homogeneous subtype of tumors through gene expression profiles to find differential expressed genes in tumors with different characteristics and predict the patient prognosis [1]. There are a huge number of genes, the majority of which could be considered as redundant or unrelated items. Therefore, it is difficult to get useful information from microarray data [2]. These bad genes cause dimensional overfitting in the process of sample classification or reduce the classification accuracy [3]. Therefore, how to reduce the number of genes in the cancer microarray data has become a main research topic in the bioinformatics field. To break through this limitation, various feature selection (FS) methods were proposed to obtain relevant gene subsets for maximizing the ability of the classifier to classify instances more accurately. As an effective tool to process high-dimensional data and improve learning efficiency [4], the FS approaches could be mainly divided into three categories, which are filter method, embedded method, and wrapper method [5].

The filter method mainly scores a single feature and sorts the features from good to bad through various scoring methods to remove those features with low scores [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Yafei Hou.

Commonly used methods are chi-square test [7], mutual information (MI) [8], F-score [9], minimum-redundancy maximum-relevancy (MRMR) [10] etc. The filter method can quickly and efficiently eliminate a large number of inferior features in the data, which is easy to implement, but it is difficult to accurately extract the optimal features and optimal subsets.

The embedded method is the most complex among these three methods. Through the training of the machine learning model, the most meaningful classification features are obtained [11]. The effect of this method depends on the performance of the classifier and is not easy to implement. Different classifiers will get different features. The classic methods include random forest (RF) [12], support vector machine (SVM) [13], etc.

The wrapper method treats the selection of subsets as a search optimization problem, generates different combinations, evaluates the combinations to confirm the quality of subsets, compares them with other combinations, and finally extracts subsets with better classification characteristics [14]. Wrapper methods usually take some metaheuristic algorithms as the main body. The feature selection problem is essentially a multi-objective optimization problem. In the research conducted by numerous scholars, they use the idea of decomposition to simplify the problem into a single-objective problem function with accuracy and feature number serving as weights, and accuracy is the main factor determining the fitness value. Genetic algorithm (GA) [15] and Particle Swarm Optimization algorithm (PSO) [16] and others have achieved significant results in addressing this problem. In addition, researchers have contributed many achievements in this field. For example, in the single-objective optimization field, Song et al. [17] proposed a new three-phase hybrid FS algorithm based on correlation-guided clustering and particle swarm optimization (PSO) (HFS-C-P). Hu et al. [18] introduced a fuzzy multi-objective feature selection method, named PSOMOFS, which incorporates fuzzy dominance relationships and fuzzy crowding distance measurements to identify the global optimal position of particles, Chen et al. [19] proposed a novel decomposition-based multi-objective clonal selection algorithm for feature selection (MOCSA/D_FS), while Jiao et al. [20] combined single and multi-objective algorithm to jointly improve feature selection framework performance. Compared with the other two methods, the wrapper method has been more popular in the field of feature selection in recent years because of its higher classification accuracy and more flexible application range.

Since the number of subsets grows exponentially with the increase of dimensionality, various standard classification algorithms could not achieve optimal results in some complex high-dimensional data. Recently, various modified classification algorithms have been used to improve the accuracy of classification utilizing different mechanisms. Firstly, some embedded methods and filter methods are proposed and improved. Moorthy et al. [21] achieved better accuracy in ten microarray datasets with an improved RF algorithm, but it is still far from perfect accuracy. Maldonado et al. [22] added features in the subset sequentially and used SVM as a scorer to eliminate the features with poor performance. This method can quickly extract a feature subset with good performance, but it is difficult to achieve the best accuracy. Wang et al. [23] combined naive Bayes (NB), decision tree (DT), SVM, and other machine learning algorithms and gained them to the classification of microarray datasets of acute uremia and diffuse large B-cell lymphoma. They proved that the comprehensive use of a variety of different algorithms could improve the selection rate of effective genes Halder et al. [24] proposed active learning using a rough fuzzy classifier (ALRFC). It is very effective for those uncertain, overlapping, and indiscernible data. Compared with some traditional machine learning methods, ALRFC has achieved better results. Tabakhi et al. [4] proposed an unsupervised feature scoring method based on ant Colony optimization (ACO), which exploits the redundancy and correlation between features to extract features. Moreover, there are a large number of neural network-related methods [25], [26], [27].

Then, with the bottleneck of machine learning algorithms and the low classification rate of a single method, some hybrid methods combined with metaheuristic algorithms began to appear in the FS domain. Liu et al. [28] combined the discrete PSO algorithm with F-score and SVM to improve the classification accuracy in small feature space. Shukla et al. [29] combined GA with NB and SVM and used F-score as feature preselection. Then they integrated the characteristics of teaching learning-based algorithm (TLBO) and gravitational search algorithm (GSA) to propose TLBOGSA by combining with MRMR and used the NB as the scoring classifier [2]. It achieves higher accuracy than other algorithms in several datasets. This method has achieved significant classification results on multiple cancer microarray datasets. With the rise of intelligent algorithms in gene selection, more and more hybrid algorithms have been proposed to execute the gene selection using various intelligent algorithms, such as discrete bacterial algorithm [30], hybrid binary black hole algorithm, and PSO algorithm [31].

As a metaheuristic algorithm proposed in 2001, the fish swarm algorithm (FSA) has good performance in the field of feature selection [32]. Compared with other algorithms [33], [34], [35], [36], [37], [38], [39], [40], the FSA has the characteristics of strong local optimization ability, insensitive parameter setting, and easy implementation. Moreover, Tirkolaee et al. [41] mentioned the main advantages of FSA compared to the multi-objective algorithm NSGA2 and they applied it to the Just in Time Energy Aware Flow Shop Scheduling Problem with Outsourcing option. Therefore, it is suitable for fine search in a small space. In recent years, the FSA is also successfully applied to feature selection. Luan et al. [42] have improved the step size of

FSA and reduced it based on a rough set. The algorithm has good global search ability and small-time complexity. Chen et al. [43] proposed an FSA based on a neighborhood rough set, which proved the effectiveness of FSA in dealing with feature reduction. Manikandan et al. [44] combined with PSO to improve FSA and successfully applied it in the feature selection problem of big data.

Nevertheless, similar as other optimal algorithms, the FSA still has some defects in the application of FS, such as deficiencies in global optimization capability and the imbalance between global and local optimization.

Especially in the problem of gene selection, high latitude and high redundancy data will make its defects more obvious.

To solve these defects, a novel FS framework using FSA are discussed in this paper to execute the gene selection. The highlight of this work could be summarized into following aspects.

Firstly, a novel FSA algorithm, namely FSA-PV, is proposed to search the potential optimal solution effectively from three viewpoints, which are adaptive visual and step mechanism, jumping out of local optimization based on the population variation automatically, and foraging behavior adjustment.

Then, the F-score method is used to pre-extract the features of the dataset. And the subsets re selected by FSA-PV method. Meanwhile, the gained subset is evaluated by the classification accuracy of the subset in the NB and LOOCV method.

The experiment results reveal that the comprehensive accuracy of the algorithm exceeds that of most similar methods, especially in the Breast datasets. Furthermore, the effectiveness of the algorithm and the accuracy of the results are proved by thermodynamic diagram, iterative diagram and comparison with some medical papers.

The rest part is organized as follows. Section two introduces the establishment of the FSA model in the selection problem. Section three demonstrates the improvements of the FSA-PV algorithm and the implementation process of gene selection using the FSA-PV. Section four verifies the feasibly of the proposed algorithm by analyzing the experimental results. Section V recalls the entire work.

II. FSA FOR FEATURE SELECTION

This section introduces the improved basic work, including standard FSA, binary of FSA, and feature pre-extraction method, respectively.

A. BASIC FSA

This paper draws lessons from the general idea of the basic fish swarm algorithm from the paper of Li et al. [32]. FSA is inspired by the behavior of fish while foraging. When fish are looking for food, they generally have the behavior of gathering, following and random swimming. According to these behaviors, the fish swarm algorithm obtains three optimization behaviors, foraging behavior, following behavior, clustering behavior. Figure 1 shows the basic behavior and parameters of FSA.

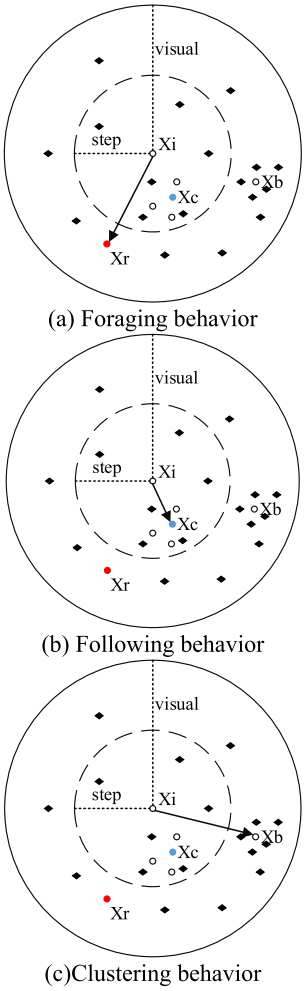


FIGURE 1. FSA behavior diagram.

In Figure 1, visual and step represent the visual field range and moving length of an individual respectively. The red dot represents a random position *Xr*, the blue dot represents the center position *Xc* of the fish swarm within the field of visual of the artificial fish Xi, the white dot represents the artificial fish, and the black diamond represents food. *Xi* represents the artificial fish that currently needs to be moved. Artificial fish will perform different behaviors to optimize, when the foraging behavior of the three behaviors is the best behavior, the current artificial fish *Xi* will move to the position of *Xr* in the field of vision as shown in (a). When the crowding behavior is the best behavior, the current artificial fish *Xi* will move to *Xc* where the artificial fish are more concentrated, as shown in (b). When the tail chasing behavior is the best behavior, the current artificial fish *Xi* will move to the artificial fish XF with the most density food as shown in (c).

In addition, there is a crowding factor δ, Individuals can judge the density of fish and food at the target location

through inequality $\frac{Y}{nf} \geq \delta Y_i$, so as to determine whether rear end behavior and group behavior occur [45].

Where, Y is fitness value of target, nf is the number of other individuals in current individual, Y_i is the fitness value of current individual.

The details of each behavior are illustrated below.

1) CLUSTERING BEHAVIOR

In the clustering behavior, the individual moves to the center point of all individuals within the field of vision, and its pseudo code is shown in Clustering behavior pseudo code. The description formula is shown in formulas (1) and (2).

$$X_i^{t+1} = X_i^t + \frac{X_c - X_i^t}{|X_c - X_i^t|} rand(0,1) \times step \quad if\, Y_i^{t+1} > Y_i^t and\, Y_c/nf \geq \delta Y_i \tag{1}$$

$$X_c = \left(\sum_{j=1}^{nf} X_j \right) /nf \tag{2}$$

where, X_i^t is the i^{th} individual in the t^{th} iteration, is the current individual, and X_c is the center point within the visual of X_i, X_j is the j^{th} individual in current individual visual, nf is the number of other individuals within the current individual visual.

2) FOLLOWING BEHAVIOR

In the following behavior, the current individual moves to the optimal individual within the visual, and its pseudo code is shown in Following behavior pseudo code. The description formula of following behavior is shown in formula (3).

$$X_i^{t+1} = X_i^t + \frac{X_b - X_i^t}{|X_b - X_i^t|} rand\ (0,1) \times step\ if\ Y_i^{t+1} > Y_i^t\ and\ Y_b/nf \geq \delta Y_i \tag{3}$$

where, X_b represent the individual with the best fitness value in others individual within the visual of current fish X_i^t.

3) FORAGING BEHAVIOR

FSA performs foraging behavior when neither clustering behavior nor tail chasing behavior can be performed. FSA mainly relies on foraging behavior to obtain more population diversity and find better areas. When the foraging behavior can't find a better position after number of *try_number*, the individual will randomly obtain a position in visual to move.

The description formula is shown in formulas (4) and (5).

$$X_i^{t+1} = X_i^t + \frac{X_r - X_i^t}{|X_r - X_i^t|} rand(0,1) \times step\ \ if\ Y_i^{t+1} > Y_i^t \tag{4}$$

$$X_i^{t+1} = X_i^t + rand(-1,1) \times step \tag{5}$$

where, X_i is the current individual and X_r is a random position within the visual of X_i.

The corresponding pseudo code of clustering behavior is summarized below.

4) FSA ALGORITHM FLOW

Through the above behavior, we can get the basic flow of FSA algorithm as follows.

Step1: Initialize the population and assign the optimal individual to the bulletin board.

Step2: Judging the behavior that an individual needs to perform.

Step3: Update fish swarm and bulletin boards.

Step4: If the output conditions are met, output the bulletin board, otherwise return to step 2.

Meanwhile, the pseudo code of standard FSA is introduced below.

B. BINARY FSA

In order to facilitate the fish swarm algorithm in dealing with the problem of feature selection, Chen et al. [46] proposed a binarization method of FSA, which has been adopted in various studies [3], [20], [57], [58]. In the FS problem, the solution space is assumed to be a large space full of feature subsets. If there are n features, there are $2^n - 1$ subsets in the space. In the FSA, each subset is defined as a position, and the position is determined by the size of the subset and the classification accuracy. To facilitate individual mobile optimization, convert the selected subset into binary representation. The selected features are marked as 1, and the unselected features are marked as 0, according to the formulas (6) and (7).

$$X_i = \{x_1, x_2, \ldots, x_{n-1}, x_n\} \tag{6}$$

$$x_i = \begin{cases} 1, & feature_i = true \\ 0, & otherwise \end{cases} \tag{7}$$

where X_i is the position which is represented by individual i, x_i represents the value which is used to reflect the feature i, and n represents the total number of features.

Then, based on the definition of binary subset in equations (6) and (7), the distance, the center and the position update mode are redefined.

1) DISTANCE

The number of difference features between two subsets can be obtained through XOR operation. The number of difference features is used to define the distance between individual I and individual J. The distance is calculated as shown in formula (8).

$$Distance_j = \sum_{k=1}^{N} X_i \oplus X_j \tag{8}$$

2) CENTER

The center will be determined by the position of other artificial fish within the current individual visual, when more than half of the individuals in the visual have the same feature, which is added to the central subset, and the central subset is also marked as the central position in the visual of the current artificial fish. The center could be calculated as shown by

formula (9).

$$Center = \{x_1, x_2, \ldots, x_{N-1}, x_N\} \; if \; \frac{1}{nf}\sum_{i=1}^{nf} X_i > 0.5$$
$$then \; x_i = 1, otherwise \; x_i = 0 \tag{9}$$

3) POSITION UPDATE

The location update method is shown in Fig 2, where X represents the individual to be moved and Y represents the target to be moved. The increase or decrease of features in the subset is used to change the position of individuals. When an individual needs to move to a certain position, the features with differences between the position and the individual are randomly selected to be changed. When the value of the *STEP* is greater than the value of the difference feature number DIF, $|STEP - DIF|$ features from their same features will be randomly selected to change.

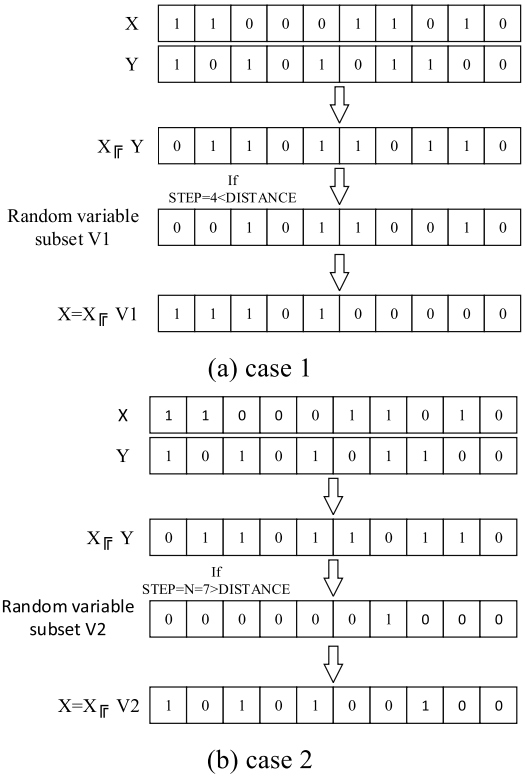


FIGURE 2. Position update diagram.

From Figure 2, it can be obtained that the distance between X and Y is 6 features. In case 1, assume the step $= 4 < 6$, and then randomly select 4 positions from the position of 1 in a as the moving variable V1. The new X is obtained by XOR operation between X and V1. In case 2, assume the step $= 7 > 6$, Randomly select $|7 - 6|$ position from the position of 0 in a as the moving variable V2. The new X is obtained by XOR operation between X and V2.

4) FITNESS FUNCTION

In the optimization problem, the fitness value is derived from a fitness function, serving as a criterion for evaluating an individual. In gene selection problems, the importance of accuracy is prioritized. Therefore, the fitness function is shown in formula (10).

$$fitness = \omega Accrary + (1 - \omega)(\frac{1}{f_{number}}) \tag{10}$$

where, ω is the weight representing importance. Accuracy is the accuracy of the individual which is calculated by embedded methods such as NB, SVM. And f_{number} is the number of characteristics of the individual.

C. FEATURE PRE-EXTRACTION METHOD

A large number of redundant or useless functions in the dataset will not only increase the processing cost, but also reduce the accuracy of classification. In order to extract effective features more effectively, the F-score algorithm and mutual information method are used to process the dataset. The appropriate pre-processing method is selected by the accuracy of the extracted subset, and the accuracy of obtaining the subset is improved.

1) F-SCORE

The F-score scores of the features could be gained by calculating the separation degree of different categories of samples and the aggregation degree of similar samples. The formula is demonstrated as shown in (11) [9].

$$F_i = \frac{\sum_{j=1}^{l} (\overline{x_i^{(j)}} - \overline{x_i})^2}{\sum_{j=1}^{l} \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (x_{k,i}^{(j)} - \overline{x_i^{(j)}})^2} \tag{11}$$

where, F_i is the F-score score of feature i, $\overline{x_i^{(j)}}$ is the average of the eigenvalues of class j samples in feature i, $\overline{x}_i$ is the average of the eigenvalues of all samples of feature i, and n_j is the total number of class j samples. $x_{k,i}^{(j)}$ is the k^{th} eigenvalue of characteristic i in class j samples.

2) MUTUAL INFORMATION

Mutual information is used to evaluate the amount of information contributed by the occurrence of one event to the occurrence of another event. Mutual information is used to obtain the correlation between different features and categories in order to screen out useful features. The scoring formula is shown in formula (12) [8].

$$I_i(Y; X) = \sum_{y_i \in Y} \sum_{x_i \in X} p(x_i, y_i) \log(\frac{p(x_i, y_i)}{p(x_i)p(y_i)}) \tag{12}$$

where X represents a feature, Y represents a category, x_i is the value of the feature, and y_i is category i. $p(x_i, y_i)$ is the joint probability distribution, $p(x_i)$ is the probability of x_i occurrence in the dataset, and $p(y_i)$ is the probability of y_i occurrence.

III. FSA BASED ON POPULATION VARIATION

The standard FSA has low optimization efficiency under fixed vision and step, and insufficient ability of jumping out of local optimization. To solve the above problems, a modified FSA, namely FSA-PV, has been discussed in this section. Then, the flowchart of gene selection using the FSA-PV is also illustrated.

A. MODIFICATION STRATEGIES

To overcome the mentioned bottlenecks of the standard FSA, the FSA-PV is demonstrated surrounding following three improvements, which are adaptive visual and step, jumping out of local optimization based on the population variation, and foraging behavior adjustment.

1) ADAPTIVE VISUAL AND STEP STRATEGY

The changeless step and visual may lead to the algorithm unable to converge or slow convergence speed in the standard FSA. Therefore, the search ability of the FSA could be enhanced effectively once the self- mechanism is added to auto-adjust the visual and step of the FSA.

Compared with some visual and step that change due to the number of iterations, an adaptive amendment strategy of the visual and step is used in the FSA-PV based on the search environment. When the current visual and step cannot support the fish to find a better position, it is judged as convergence stagnation, the visual and step will change according to the current value. The improvement method is shown in formulas (13)-(15).

$$ite' = \begin{cases} ite' + 1, fitness_{bt} = fitness_{bt-1} \\ 0, fitness_{bt} > fitness_{bt-1} \end{cases} \tag{13}$$

$$visual = \text{int}(visual \times \omega_{s,v}), ite' > \alpha \tag{14}$$

$$step = \text{int}(step \times \omega_{s,v}) + 1, ite' > \alpha \tag{15}$$

where ite' is the number of convergence stagnation, fitness_{bt} is the optimal value of the bulletin board in iteration, t^{th}, $fitness_{bt-1}$ is the optimal value of the bulletin board in iteration, $(t-1)^{th}$ represent the field of view, and step is the value of convergence stagnation for contraction of visual and step.

2) POPULATION VARIATION

In the early stage of the standard FSA, it is easy to eliminate some features which are not obviously helpful to the subset in the process of feature reduction because the larger subset is not sensitive to a small number of features. These features may obtain excellent performance in other subsets and can significantly improve accuracy. Therefore, how to add the eliminated excellent features into the subset in the process of jumping out of the local optimization will be the key phase to improve accuracy.

In order to jump out of the local optimum, it is necessary to disrupt the population at a later stage to increase the diversity of the population. However, in the FS problem, the local optimal position often contains some excellent features, so the random local jump behavior is easy to eliminate the excellent features in the local optimal position and to reduce the search efficiency of the algorithm. Therefore, a population variation out-of-local optimization method with feature retention mechanism is proposed to make the FSA obtain better search efficiency in gene selection problems. In this strategy, the population diversity is improved by adding new features. Meanwhile, the number of new features determines the optimization efficiency in the next stage and the probability that the population contains better subsets. The more the number of features, the lower the optimization efficiency, and the higher the probability of containing better subsets. For keep the balance the convergence rate with the probability of containing a better subset, a weight is introduced to adjust the increased number of features. The population variation behavior could be evaluated by the formulas (16)-(19).

$$ite'' = \begin{cases} ite'' + 1, & if\ fitness_{bt} = fitness_{bt-1} \\ 0, & fitness_{bt} > fitness_{bt-1} \end{cases} \tag{16}$$

$$skip = skip + 1, \quad if\ ite'' > \beta \tag{17}$$

$$m_{skip} = feature_{skip-1}/2 \tag{18}$$

$$Set_{new} = Set_{best} \bigcup Set_m, \ if\ ite'' > \beta \tag{19}$$

where, ite'' represents the number of times of convergence stagnation, the skip represents the number of times of optimization stages. After one time population variation, the algorithm will move to the next stage of optimization. mo is the initial characteristic number of the supplementary subset, m_{skip} is the number of features contained in the supplementary subset in $skip^{th}$ stage. $Featureskip$ represents the number of features of best subset in $skip^{th}$ stage. Set_{new} is the new search space, Set_{best} is the optimal subset in the previous optimization stage, and Set_m is the supplementary subset with m features. β is the tolerance value of convergence stagnation for population variation, after the number of convergence stagnation exceeds β, the populations fall into local optimization and perform the population variation.

3) FORAGING BEHAVIOR MODIFICATION

In the standard FSA, the individual will move randomly after the foraging behavior is consumed try_number times. In the FS process, randomly increasing or decreasing features through the foraging behavior will lead to the individual not moving to a better position. Hence, three mechanisms are used to accelerate the convergence speed.

Firstly, the individuals should be given up the move when they can't find a better position.

Then, the population variation mechanism will make up for the deficiency that the new random behavior can't increase the population diversity.

After the population variation behavior occurs, all random behaviors will become random reduction for speeding up the convergence of the algorithm.

B. PESUDO CODE OF THE FSA-PV AND COMPLEXITY ANALYSIS

Due to the improvements discussed above, the corresponding pseudo code of FSA-PV is summarized below.

According to introduced pseudo code above, the algorithm complexity of the proposed FSA-PV is analyzed as follows.

In each iteration, the fitness value of each individual is calculated for the first time, and the time complexity is O(N).

Before executing each behavior, the distance between each individual needs to be calculated, and the time complexity is O(N!).

When performing each behavior. The process time complexity of traversing each individual is O(N);

There are at most (N-1) individuals in visual when performing rear end behavior. Therefore, the time complexity of determining the best individual in visual is O(N-1). The current individual moves to the optimal individual in the field of vision, and the time complexity is O(1). Therefore, the time complexity of rear end collision behavior is O ((N-1) + 1 + 1)) = O (N);

When performing clustering behavior, the time complexity of obtaining the central position is O(1), and the time complexity of judging the current individual moving to the central position is O(1). Therefore, the time complexity of clustering behavior is O (1 + 1) = O(2);

When the foraging behavior is performed, the time complexity of randomly obtaining the position in the visual is try_number, the time complexity of random moving step is O(1), and the time complexity of foraging behavior is O (try_number + 1) = O (try_number).

Because of foraging behavior can be nested in tail chasing behavior or clustering behavior, the time complexity of FSA is O (N! + N + N) × ((N+ try_number)+(try_number)) = O(N!+N × try_number).

Compared to the standard FSA, the FSA-PV adds the adoptive step and visual, the foraging behavior improvement method and the population variation behavior. The corresponding time complexity of these strategies are both constant C.

To sum up, the time complexity of FSA-PV is O(N!+N×try_number+C) = O(N!+N×try_number). Obviously, the time complexity of the FSA-PV is higher than that of the basic FSA, but they are of the same order of magnitude.

C. GENE SELECTION PROCESS OF FSA-PV

As a hybrid method, the FSA-PV reduces a large number of useless features by filter method in the first stage, then a metaheuristic algorithm is used as a tool to obtain subsets in the second stage, and the machine learning model is employed as an evaluator of subsets to obtain the subset with the highest classification accuracy in the latter phase.

The whole flowchart of the gene selection process utilizing the FSA-PV is depicted in Figure 3 as follows.

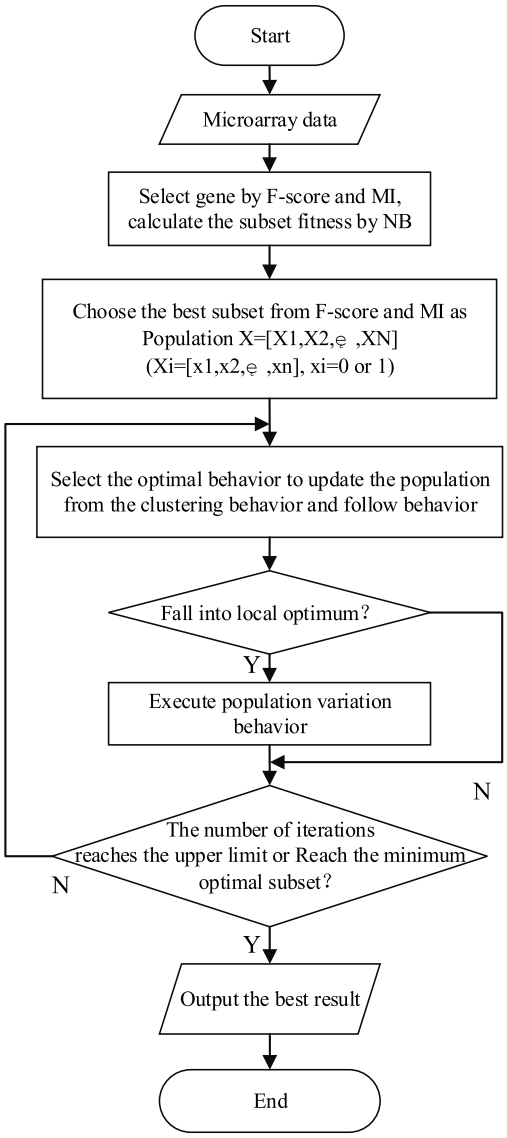


FIGURE 3. Algorithm flow diagram of FSA-PV.

Due to the mentioned flowchart, the gene selection using the FSA-PV includes following main steps.

Step 1: Put the initial microarray data into the filter F-score and MI to obtain the initial search space and generate the initial population.

Step 2: Perform the clustering Behavior of formula (1), the following behavior of formula (2) or the foraging behavior of formula (4).

Step 3: Use formula (12)-(14) to judge whether the visual and step changes adaptively.

Step 4: Use formula (15)-(18) to judge whether to perform population variation behavior.

Step 5: If the number of iterations reaches the upper limit or the minimum optimal subset is obtained, the optimal result is output. Otherwise, return to step 2.

Pseudocode of FSA-PV

Input: N individuals in random position and their fitness.
Output: The position and fitness of the global optimal individual.
Implementation:

Initialization parameters.
generate a population of artificial fish swarm $X = (X_1, X_2, \ldots, X_N)$, and calculate value of fitness.
record the best artificial fish on the bulletin board.
for $t = 1$ **to** T # iterations
  for $i = 1$ **to** N # artificial fish
    X_f is obtained by Following behavior.
    if Follow behavior **Fail** then
      X_f is obtained by Forage behavior.
    end
    X_h is obtained by Clustering behavior.
    if clustering behavior **Fail** Then
    X_h is obtained by forage behavior.
  end
    X_i choose better result from X_h and X_f to update.
  end
  New population were obtained and update the bulletin board.
  if $best(t) = best(t-1)$ # $best(t)$ is the best individual in t^{th} iteration
    $ite' = ite' + 1$
    $ite'' = ite'' + 1$
  else if $best(t) > best(t-1)$
    $ite' = 0$
    $ite'' = 0$
  end
  if $ite' > \alpha$
    $visual = visual \times \omega_{s,v}$
    $step = step \times \omega_{s,v}$
    $ite' = 0$
  end
  if $ite'' > \beta$
    Get a random subset $set_m = \{f_1, f_2, \ldots, f_m\}$ from all features.
    The new feature space $Subset_{new} = Set_{best} \bigcup Set_m$.
    Replace all artificial fish with the best fish in the bulletin board.
  end
end
output the best value in the all of iterative process.

IV. EXPERIMENTS AND RESULT ANALYSIS

In this section, the feature pre-extraction method and the machine learning model as a grader are determined through eight different microarray data. Furthermore, the performance of FSA-PV is verified by experiments and comparisons.

A. EXPERIMENT PREPARATION AND DATA COLLECTION

All the experiments are performed on a computer equipped with an Intel(R) Core i5-1135G7 processor with a frequency of 2.40 GHz, 16 GB of memory and a Windows 10 operating system. The algorithm is implemented using Python 3.8 in JetBrains PyCharm 2019. In order to verify the performance of the algorithm, eight microarray datasets are selected to test the performance, which are listed in Table 1. These data are from the public database http://csse.szu.edu.cn/staff/zhuzx/Datasets.html, which include Colon, Central Nervous System (CNS), Leukemia, Lung, Breast, Myeloid/Lymphoid Leukemia (MLL), Ovarian, and Small Round Blue Cell Tumors (SRBCT).

The algorithm parameter settings are shown in Table 2. The overall scale is 30. LOOCV and classifier are used to evaluate the extracted subset. Due to the randomness of meta heuristic algorithm, comparing the average results of the algorithm can better reflect the overall performance of the algorithm. The average classification accuracy obtained after 30 independent runs of the algorithm is taken as the final target value.

Due to the FSA-PV is insensitive to parameter values, it is unnecessary to set a fixed parameter value. The parameter

TABLE 1. Microarray gene dataset.

NO.	Dataset	Instances	genes	class
1	CNS	60(1:21, 0:39)	7129	2
2	Colon	62(Tumor:40, Normal:22)	2000	2
3	Leukemia	72(ALL:47, AML:25)	7129	2
4	Lung	203(1:139, 2:17, 3:6, 4:21, 5:20)	12600	5
5	MLL	72(ALL:24, MLL:20, AML:28)	12582	3
6	Ovarian	253 (Normal:91, Cancer:162)	15154	2
7	SRBCT	83(1:29, 2:11, 3:18, 4:25)	2308	4
8	Breast	97(relapse:46, non-relapse:51)	24481	2

TABLE 2. Parameter setting.

parameter	value
Population	30
Visual	1/4space
Step	1/2Visual
δ	0.75
α	2
β	4
$\omega_{s,v}$	0.75

values within a reasonable range can ensure the efficiency and accuracy of the algorithm. According to the size of population and search space, the initial visual is set as the size of 1/4 solution space and the step value of 1/2 visual to ensure the optimization efficiency of the algorithm and. Crowding factor determines the aggregation degree of fish schools. In the FS problem, at the later stage of convergence, the algorithm will optimize within a small space created by the population variation mechanism. At this time, the aggregation of fish schools will reduce the global search ability of the algorithm. We let individuals try to search their own neighborhood instead of Clustering together, so it is necessary to choose a larger one δ value.

In a small solution space, in order to make the population variation behavior play a funny role, we need to choose a smaller α value to help the visual and step shrink rapidly, so as to reach a local optimal subset faster. Setting to 2 can achieve this effect very well.

In the iterative process, when the population converges to the local optimal value, more iterations are needed to determine whether the location is the optimal location in the current region. But too large β value will reduce the convergence speed of the algorithm, and due to the existence of population variation, neighborhoods close to the local optimal value can be searched continuously in the subsequent search, which solves the problem that neighborhoods cannot be traversed. Therefore, setting β to 4 can ensure the convergence efficiency and accuracy of the algorithm.

B. SELECTION OF CLASSIFIER AND FEATURE PRE-EXTRACTION METHOD

In order to obtain more stable, efficient and high-precision classifiers. Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Decision Tree (DT) and Logistic Regression (LR) are selected to classify CNS datasets to test their stability. Their average value, optimal value, standard deviation and total running time are all obtained by running the same resource 30 times. The results are shown in Table 3.

It can be seen from Table 3 that the random forest and decision tree have better maximum classification accuracy, but their stability is not suitable for being used as a classifier to judge the quality of the extracted subset in the algorithm. KNN, SVM and NB with better results and lower time complexity are choose in this manuscript to verify the performance of FSA-PV.

Ghosh al et [6] points out that MI is a better method compared with other classical filter methods. As a widely used method in microarray gene selection, F-score is not reflected in this literature. Therefore, in order to choose a better filter method, we used NB, KNN, SVM to compare their filtered 200 top-level genes, and the results are shown in Table 4.

From table 4, F-score is better than MI in the accuracy of CNS and breast, and has the most obvious advantage in the Breast dataset. In Lung, MLL data, the accuracy of MI is better than F-score. The two methods show different filter effects in different datasets, and neither of them has significant advantages. Therefore, we combined FSA-PV to test the two filter methods, using three different classifiers to test 200 top-level genes. Each group of data was run independently for 10 times, and the average accuracy was taken. The results are shown in table 5-7.

In the results from table 5 to table 7, the accuracy of extracting subsets in datasets SRBCT, MLL, Ovarian and Leukemia by MI and F-score methods has reached 100%. In CNS, Colon, Lung and Breast datasets, the results of MI and F-score are the same as those in Table 4. F-score has advantages in CNS and Breast datasets, and MI has advantages in Colon and Lung datasets. However, the advantages of MI are not significant, and the accuracy of MI in CNS and Breast datasets is significantly lower than that of F-score. Therefore, we use F-score as the filter method of the whole feature selection framework.

Dabba et al. [52] has proved that when the number of top-level genes M = 200, SVM classifier can be used to obtain

TABLE 3. Performance of machine learning methods in CNS dataset.

Method	SD	Average	Best	Time
RF [12]	0.0291	0.6055	0.6667	60.367
SVM [13]	0	0.65	0.65	26.94
NB [47]	0	0.6333	0.6333	24.87
KNN [48]	0	0.6333	0.6333	23.451
LDA [49]	0	0.55	0.55	47.116
DT [50]	0.0336	0.6383	0.7333	47.095
LR [51]	0	0.6	0.6	47.016

TABLE 4. Accuracy in each dataset using MI and F-score methods.

		CNS	SRBCT	Colon	Lung	MLL	Ovarian	Breast	Leukemia
MI	NB	73.33	100	76.06	95.00	95.71	98.41	54.00	97.14
	SVM	81.66	100	85.30	93.09	95.71	99.2	52.57	98.57
	KNN	73.33	100	86.96	94.05	94.28	98.41	66.91	94.28
F-score	NB	80.00	100	75.90	92.04	94.64	97.61	78.55	95.71
	SVM	83.33	100	86.81	82.24	93.03	99.60	72.30	97.14
	KNN	78.33	100	85.30	90.61	93.03	97.61	73.28	92.85

TABLE 5. NB method of operation results of FSA-PV.

		CNS	SRBCT	Colon	Lung	MLL	Ovarian	Breast	Leukemia
MI	Best	93.33	100	96.90	99.00	100	100	87.74	100
	SD	0	0	8.06E-3	2.58E-3	0	0	2.08E-2	0
	avg	93.33	100	**95.73**	**98.79**	100	100	85.73	100
F-score	Best	100	100	96.90	99.02	100	100	91.66	100
	SD	5.28e-3	0	7.04E-3	4.61E-3	0	0	7.15E-3	0
	Avg	**97.94**	100	93.90	98.22	100	100	**90.70**	100

TABLE 6. SVM method of operation results of FSA-PV.

		CNS	SRBCT	Colon	Lung	MLL	Ovarian	Breast	Leukemia
MI	Best	90.00	100	93.33	98.02	100	100.00	85.66	100
	SD	8.06E-3	0	0	2.31E-3	0	0	3.28E-2	0
	Avg	89.50	100	93.33	**97.87**	100	100	83.29	100
F-score	Best	94.99	100	93.33	96.06	100	100	90.77	100
	SD	8.06E-3	0	0	2.53E-3	0	0	5.73E-3	0
	Avg	**93.32**	100	93.33	95.79	100	100	**90.10**	100

TABLE 7. KNN method of operation results of FSA-PV.

		CNS	SRBCT	Colon	Lung	MLL	Ovarian	Breast	Leukemia
MI	Best	93.33	100	98.33	99.00	100	100	93.77	100
	SD	1.75E-2	0	1.98E-2	5.13E-3	0	0	1.64E-2	0
	Avg	91.66	100	95.65	98.14	100	100	91.47	100
F-score	Best	93.33	100	95.23	98.04	100	100	97	100
	SD	7.04E-3	0	6.88E-3	3.97E-3	0	0	4.06E-3	0
	Avg	92.82	100	94.02	97.25	100	100	95.30	100

an accuracy of 100 in all datasets, and different data have different requirements for the number of the least top-level genes.

In order to obtain the number of top-level genes required by NB and KNN classifiers, we used 100 iterations to test the accuracy of FSA-PV when the number of top-level genes M = 50, 100 and 200 respectively, and the results are shown in Table 8.

The results show that when the number of top-level genes is M = 50, except for those datasets with an accuracy of 100 (SRBCT, Ovarian, Leukemia, MLL), the accuracy is lower than M = 100 and M = 200.

In the comparison of M = 100 and M = 200, we also pay attention to those datasets whose accuracy does not reach 100. In CNS, Colon and Breast, the accuracy when M = 200 is lower than that when

TABLE 8. Accuracy of different top-level genes in NB and KNN.

Dataset	NB			KNN		
	50	100	200	50	100	200
CNS	96.55	97.77	97.94	90.16	93.49	92.82
SRBCT	100	100	100	100	100	100
Colon	92.84	94.10	93.90	91.63	92.94	94.02
Lung	95.91	96.93	98.22	94.07	96.70	97.25
MLL	100	100	100	100	100	100
Ovarian	100	100	100	100	100	100
Breast	89.79	91.21	90.70	89.58	91.54	95.30
Leukemia	100	100	100	100	100	100

TABLE 9. Comparison of accuracy and number of genes obtained by FSA-PV and FSA.

Dataset	Accuracy (%)				Genes			
	Best	Worst	Average	SD	Best	Worst	Average	SD
CNS	100	98.33	99.88	0.00431	8	16	11.5	2.596
SRBCT	100	100	100	0	5	9	6.66	0.9759
Colon	96.9	95	95.50	0.00687	7	14	8.9	1.88
Lung	99.49	98.02	98.78	0.00308	17	38	25.13	5.21
MLL	100	100	100	0	3	6	4.33	0.816
Ovarian	100	100	100	0	3	4	3.2	0.4068
Breast	94	91.88	92.88	0.0065	7	16	10.8	2.49
Leukemia	100	100	100	0	3	5	4.26	0.59

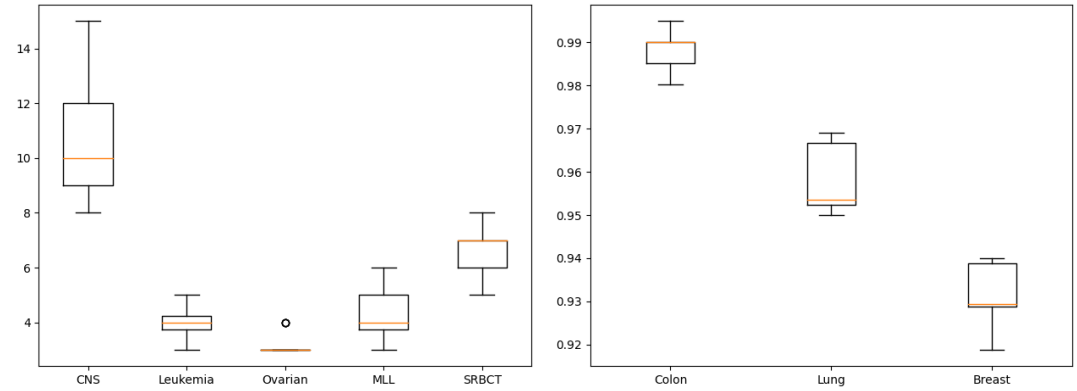


FIGURE 4. Cluster thermodynamic diagram of eight classical datasets (the left is (a), the right is (b)).

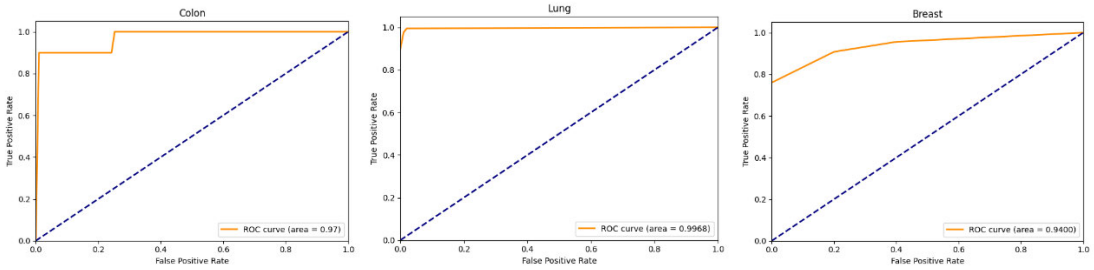


FIGURE 5. ROC curves of Colon, Lung, Breast.

M = 100. In the Lung dataset, the accuracy when M = 200 is better than that when M = 100.

Although there are redundant genes when M = 200, low scores of cancer-related genes are common in some medical literature about cancer genes [53], [54]. In the gene selection study of a Qu et al. [3], genes ranking after 150 appear in the optimal subset. Therefore, if the performance of the algorithm allows, using more top-level genes as the initial solution space is conducive to finding more optimal subsets and genes. In the subsequent experiments, we all use the number of top-level genes M = 200 as the initial solution space.

TABLE 10. The optimal gene subset obtained by FSA-PV.

Dataset	Accuracy	Feature	Genes
CNS	100	8	'AFFX-CreX-3_st', 'M18728_at', 'M55593_at', 'D16688_s_at', 'HG2994-HT4850_s_at', 'X64624_s_at', 'S66541_s_at', 'U43747_s_at'
SRBCT	100	5	'gene2', 'gene484', 'gene545', 'gene1003', 'gene1662'
Colon	96.9	7	'X87159', 'M76378', 'R36977', 'L08069', 'T47377', 'H55916', 'T57468'
Lung	99.49	40	'35414_s_at', '37490_at', '41018_at', '41385_at', '41418_at', '41423_at', '41435_at', '32650_at', '33230_at', '34265_at', '36921_at', '37192_at', '37545_at', '40093_at', '40410_at', '40421_at', '40423_at', '40783_s_at', '40863_r_at', '41197_at', '32169_at', '33340_at', '35742_at', '35835_at', '36207_at', '36606_at', '37365_at', '37406_at', '37722_s_at', '38368_at', '38408_at', '39178_at', '41288_at', '32530_at', '32569_at', '1641_s_at', '1314_at', '657_at', '409_at', '162_at'
MLL	100	3	'38242_at', '37710_at', '36122_at', '1389_at'
Leukemia	100	3	'M62762_at', 'M92287_at', 'M31523_at'
Ovarian	100	3	'MZ2.7921478', 'MZ245.53704', 'MZ4010.7341'
Breast	94.00	17	'Contig53223', 'NM_003450', 'Contig53488', 'NM_012261', 'NM_005176', 'Contig43859_RC', 'NM_014003', 'NM_013306', 'NM_005342', 'AJ011306', 'AL080059', 'Contig47544_RC', 'Contig21190_RC', 'Contig412_RC', 'Contig48208_RC', 'NM_018089', 'Contig3920_RC'

TABLE 11. The Significant genes of optimal subset.

Dataset	Genes
CNS	'AFFX-CreX-3_st', 'M18728_at', 'M55593_at', 'HG2994-HT4850_s_at', 'S66541_s_at', 'U43747_s_at'
SRBCT	'gene2', 'gene545', 'gene1662'
MLL	'38242_at', '37710_at', '36122_at', '1389_at'
Leukemia	'M31523_at
Ovarian	MZ2.7921478

TABLE 12. Comparison of result obtained by FSA-PV and FSA.

Dataset	FSA-PV		FSA	
	Accuracy	Genes	Accuracy	Gens
CNS	99.88	11.5	94.44	58.16
SRBCT	100	6.66	100	65.14
Colon	95.50	8.9	91.02	52.5
Lung	98.78	25.13	98.19	65.5
MLL	100	4.33	100	51.36
Ovarian	100	3.2	100	39
Breast	92.88	10.8	87.99	56.33
Leukemia	100	4.26	100	48.2

C. GENE SELECTION RESULTS AND ANALYSIS

It can be seen from Table 8 that when the number of iterations is 100, it is not enough to support FSA-PV to find the best gene subset. Therefore, in order to prove the search ability of the algorithm, we have expanded the number of iterations to 500. The results are shown in Table 9.

Table 9 shows the optimal value, worst value, average value and standard deviation of precision and gene number of FSA-PV under the condition of 500 iterations. For different datasets, there are great differences in the number of genes in the optimal subset obtained by the combination method. Combined with the selection results of the top-level genes in

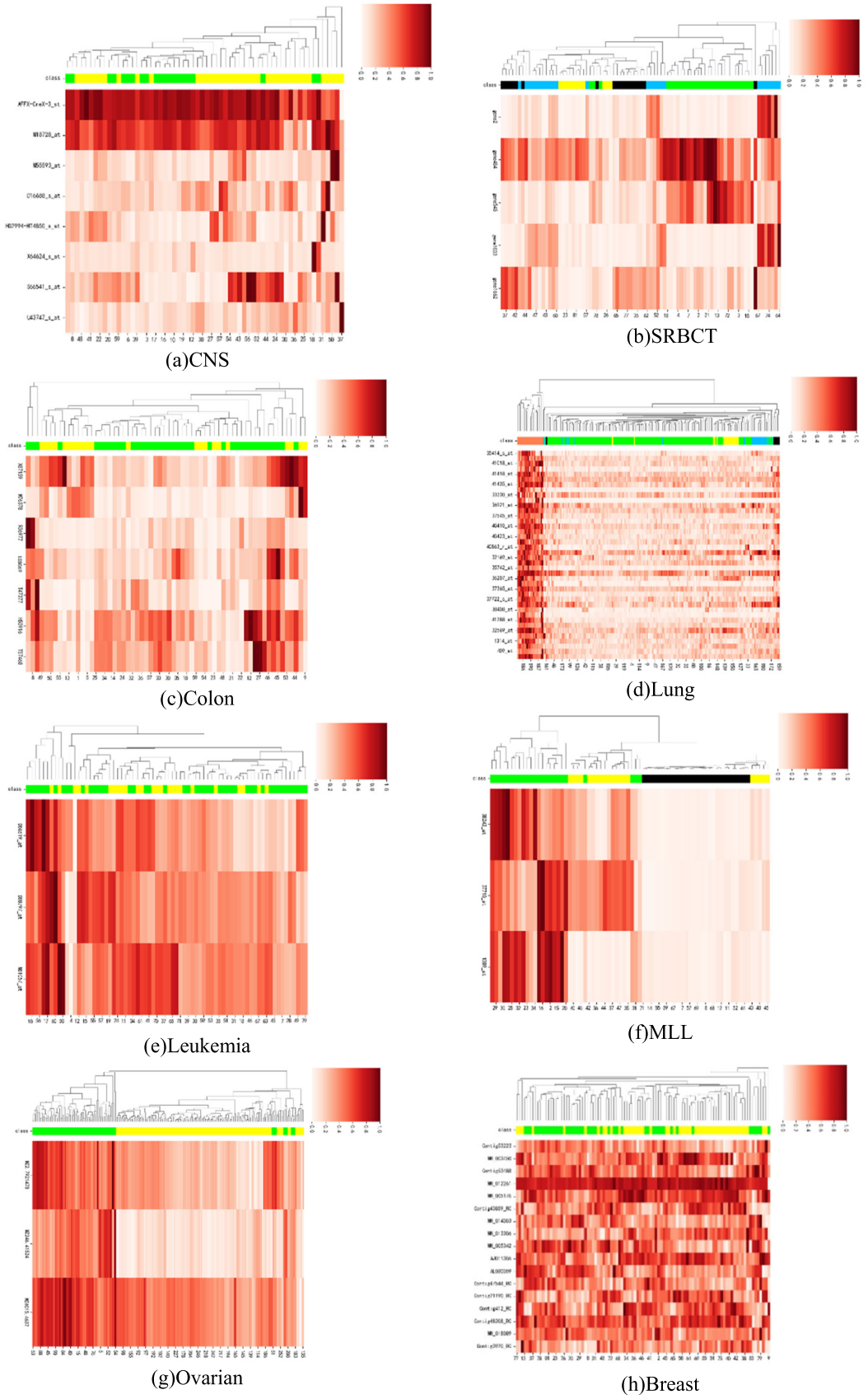


FIGURE 6. Cluster thermodynamic diagram of eight classical datasets.

TABLE 13. Comparison of AUC, Recall, F-measure and Precision obtained by FSA-PV and FSA.

Dataset	FSA-PV				FSA			
	AUC	Recall	F-measure	Precision	AUC	Recall	F-measure	Precision
CNS	99.93	100	99.77	99.66	94.99	93.71	91.88	93.05
SRBCT	100	100	100	100	100	100	100	100
Colon	96.296	99.13	95.872	94.08	91.66	97.77	91.97	88.33
Lung	99.22	96.90	97.12	97.80	98.86	97.46	97.06	97.70
MLL	100	100	100	100	100	100	100	100
Ovarian	100	100	100	100	100	100	100	100
Breast	93.60	92.38	92.62	93.62	89.82	87.24	87.87	90.99
Leukemia	100	100	100	100	100	100	100	100

TABLE 14. Parameters of other methods.

Method	Parameter
VNLHHO	Population size n=30, number of generation T=100
AGA	Population size n=30, number of generation T=100, K2=0.05, K5=0.03
rMRMR-MGWO	Number of generation T=100
MIM-mMFA	Population size n=50, number of generation T=30
QMFOA	Population size n=50, number of generation T=100, w1=0.7, w2=0.3, random angle=3
FSA-PV	Population size n=30, generation T=100

TABLE 15. Comparison of experimental results obtained by FSA-PV with other method (%).

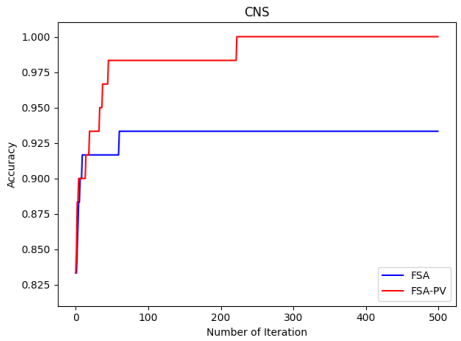
Method		CNS	SRBCT	Colon	Lung
VNLHHO [3]	Avg	99.66(12.6)	100.00(6.1)	96.12(4.7)	97.92(26)
	Best	99.99(7)	100(4)	96.77(7)	98.52(19)
AGA[29]	Avg	-	98.79(13)	98.90(16)	99.03(14)
	Best	-	99.78(13)	99.85(24)	99.85(15)
rMRMR-MGWO[57]	Avg	99.39(17.47)	100(12.3)	95.87(9.8)	97.9(15.8)
MIM-mMFA[59]	Avg	99.83(24.7)	99.4(27.3)	100(26.3)	100(35.3)
	Best	100(13)	100(23)	100(20)	100(20)
QMFOA[58]	Avg	100(31.2)	99.44(28.27)	100(30.5)	100(26.6)
	Best	100(28)	100(23)	100(27)	100(20)
FSA-PV	Avg	99.88**(11.5)**	**100**(6.6)	95.50(8.9)	98.78(25.3)
	Best	**100(7)**	**100**(5)	96.9**(7)**	99.49(40)
Method		MLL	Leukemia	Ovarian	Breast
VNLHHO	Avg	100(6.7)	100(3)	100(2.7)	-
	Best	100(3)	100(2)	100(3)	-
AGA	Avg	-	98.72(13)	-	88.64(17)
	Best	-	99.52(10)	-	91.47(13)
rMRMR-MGWO	Avg	100(8.4)	100(5.06)	100(3.57)	-
MIM-mMFA	Avg	100(33.3)	100(7.5)	98.18(35.9)	86.8(25.9)
	Best	100(19)	100(6)	100(26)	91.75(11)
QMFOA	Avg	-	100(36.4)	98.42(20.6)	74,23(27.7)
	Best	-	100(32)	100(17)	81.44(22)
FSA-PV	Avg	**100(4.3)**	**100**(4.2)	100(3.2)	**92.88(10.8)**
	Best	**100(3)**	**100**(3)	**100(3)**	**94**(17)

the initial solution space in Table 8, we can find that the more datasets contain features in the optimal subset, the greater the requirements for the initial solution space, this is reflected in the results of Lung dataset. Therefore, for different datasets, using a variety of methods can improve the accuracy of gene selection. In addition, figure 4(a) displays the box plots of the number of features for datasets with 100% accuracy, and Figure 4(b) shows the box plots of datasets with less than 100% accuracy.

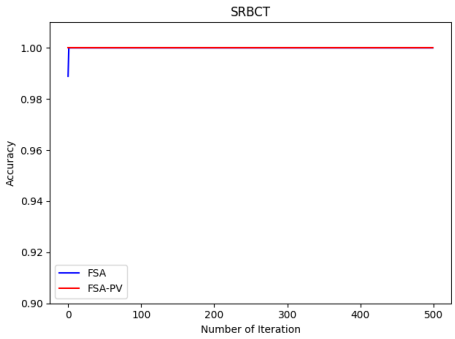
Table 10 shows the genes contained in the extracted optimal subset. In addition, FIGURE 5 shows the ROC curves of Colon, Breast, and Lung, which can more intuitively reflect the classification accuracy of the given subset.

In those classifications that do not reach 100% accuracy, it is easy to reveal that some genes have been proved to be related to disease from other studies. In literature [53], gene S66541_s_at has been shown to be related to CNS. Literature confirmed that AL080059 is related to the Breast cancer gene [54]. T47377 has been proved to be related to Colon cancer by several papers [55], [56].

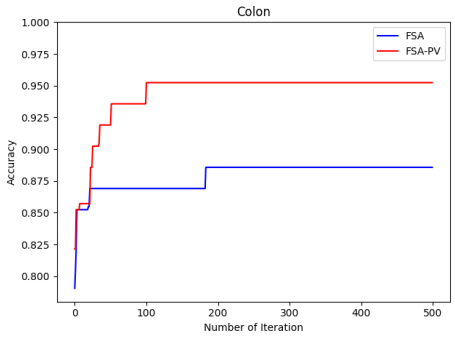
Table 11 shows those genes with a probability of more than 75% in all optimal subsets. These screened genes may be closely related to their corresponding diseases.



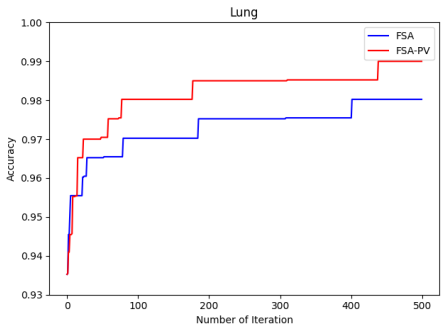
(a) CNS data



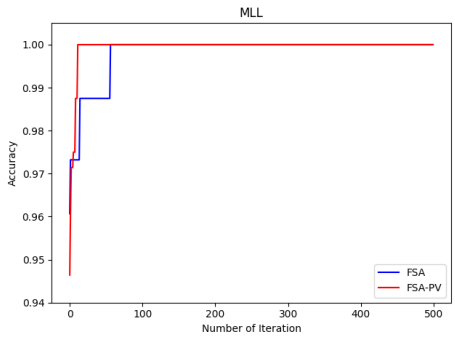
(b) SRBCT data



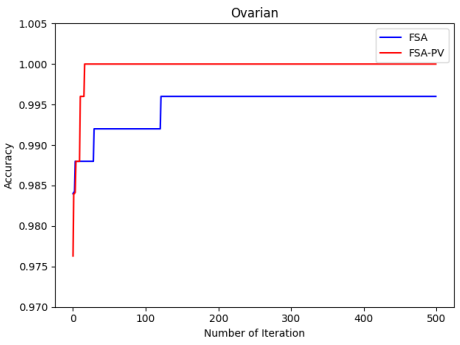
(c) Colon data



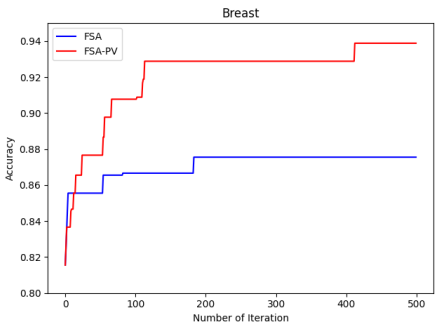
(d) Lung data



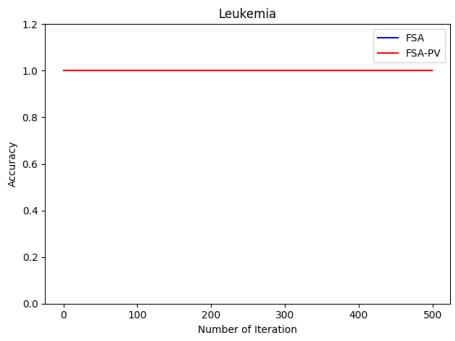
(e) MLL data



(f) Ovarian data



(g) Breast data



(h) Leukemia data

FIGURE 7. **Convergence diagram of the accuracy of FSA-PV and standard FSA.**

To verify the accuracy of the extracted subset, the genes in Table 10 were clustered using the Unweighted Pair Group Method with Arithmetic mean (UPGMA) (the data were normalized), and the results are shown in the figure 6. Among them, different colors in class represent different categories, and the clustering effect can be judged by the

TABLE 16. Comparison of AUC, Recall, F-measure and Precision obtained by FSA-PV with other method.

Dataset		VNLHHO	rMRMR-MGWO	MIM-mMFA	QMFOA	FSA-PV
CNS	AUC	0.999	-	0.852	0.944	**0.9993**
	Recall	-	0.9873	0.972	0.889	**1**
	F-measure	-	0.9953	0.958	0.842	**0.9977**
	Precision	-	**0.9974**	-	-	0.9966
SRBCT	AUC	**1**	-	**1**	0.984	**1**
	Recall	-	**1**	**1**	0.978	**1**
	F-measure	-	**1**	**1**	0.973	**1**
	Precision	-	**1**	-	-	**1**
Colon	AUC	**0.98**	-	0.958	0.910	0.9629
	Recall	-	0.9409	0.986	**1**	0.9913
	F-measure	-	0.968	**0.986**	0.991	0.9587
	Precision	-	**0.9683**	-	-	0.9408
Lung	AUC	**0.998**	-	0.985	0.991	0.9922
	Recall	-	**1**	0.976	0.988	0.9690
	F-measure	-	**1**	0.975	0.980	0.9712
	Precision	-	**1**	-	-	0.9780
MLL	AUC	**1**	-	0.975	-	**1**
	Recall	-	**1**	0.958	-	**1**
	F-measure	-	**1**	0.962	-	**1**
	Precision	-	**1**	-	-	**1**
Leukemia	AUC	**1**	-	**1**	0.972	**1**
	Recall	-	**1**	0.967	0.944	**1**
	F-measure	-	**1**	0.981	0.957	**1**
	Precision	-	**1**	-	-	**1**
Ovarian	AUC	**1**	-	0.983	1	**1**
	Recall	-	-	0.969	0.984	**1**
	F-measure	-	-	0.930	0.992	**1**
	Precision	-	-	-	-	1
Breast	AUC	-	-	0.779	0.799	**0.9360**
	Recall	-	-	0.924	0.870	**0.9238**
	F-measure	-	-	0.951	0.729	**0.9262**
	Precision	-	-	-	-	0.9362

clustering degree of the same color and the separation degree of different colors. In all data, the optimal subset extracted by FSA-PV has a good clustering effect, which verifies the effectiveness of the extracted optimal subset.

D. FSA-PV COMPARED WITH THE STANDARD FSA

Table 12 compares the accuracy of FSA-PV with standard FSA and the number of genes in the gene subset. We compared the fitness values of FSA-PV and FSA from a statistical point of view through Wilcoxon rank sum test. In all datasets, the p-value between the two is<0.05, and the performance of FSA-PV is significantly better than FSA. The effectiveness and effect of the improvement are proved.

Due to the difference in the number of positive and negative samples in the dataset, it is incomplete and unbalanced to only compare the number of genes in accuracy and subset. Therefore, FSA and FSA-PV are compared in terms of Area Under Curve (AUC), recall, F-measure and precision, as shown in Table 13. After Wilcoxon rank sum test, the AUC, recall, F-measure and precision evaluation indexes of FSA-PV are significantly better than FSA on CNS, Colon, Lung and Breast datasets (p-value<0.05). Therefore, the performance metrics are determined as the following formulas (20)–(23).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{23}$$

where, TP (True Positive) is the number of positive cases that have been classified correctly as positive. TN (True Negative) is the number of negative cases that have been classified correctly as negative. FN (False Negative) is the number of positive cases that have been classified incorrectly as negative cases. FP (False Positive) is the number of negative cases that are incorrectly classified as positive cases.

Figure 7 shows the convergence diagram of the accuracy of FSA-PV and standard FSA. In the comparison between FSA-PV and standard FSA, FSA-PV has better convergence speed and accuracy.

E. FSA-PV COMPARED WITH OTHER ADVANCED METHODS

Table 15 shows the accuracy of several most advanced feature selection methods in recent years, which adopt the framework

of filter packaging embedding. In comparison with other methods, these methods have shown more significant effects, and their parameters are shown in Table 14. The number in '()' represents the corresponding gene number, The '-' means that the data is not displayed in the corresponding method.

Where, the results of FSA-PV and VNLHHO are relatively close. The average accuracy, optimal accuracy and average feature number of FSA-PV in Lung and CNS datasets are better than VNLHHO. In Colon datasets, the two algorithms have their own advantages and disadvantages in the optimal subset and average subset respectively. In other datasets, the number of features of VNLHHO is better than that of FSA-PV. In comparison with AGA, FSA-PV is better than AGA, in general, except for a slight disadvantage in Lung data. In comparison with rMRMR-MGWO, FSA-PV has great disadvantages in Colon data, but on the whole, it is better than rMRMR-MGWO. In the comparison between MIM-mMFA and QMFOA, the number of features obtained by FSA-PV has great advantages, but there is a disadvantage in accuracy in Colon and Lung datasets. In general, FSA-PV has comprehensive and significant advantages over all other algorithms on the Breast dataset, and has advantages in the number of features compared with some algorithms. Although there are disadvantages in accuracy in Lung and Colon data, on the whole, FSA-PV algorithm is not inferior to any of the above algorithms.

In order to compare the above algorithms more comprehensively, we have compared the AUC, Recall, F-measure and Precision indicators, and the results are shown in Table 16.

Through the comparison of AUC, recall, F-measure and precision indicators, the results are roughly the same as the accuracy comparison. There are deficiencies in Lung and Colon data, and it is better than other methods in Breast data.

From Table 15 and Table 16, the different classifier or filter method can product the result differently in distinct dataset, such as the Lung and Colon dataset can generated the better result by MI and SVM in MIM-mMFA, Breast dataset can obtain best Accuracy by F-score and NB.

V. CONCLUSION

This paper proposes a feature selection algorithm for gene selection. Firstly, the algorithm uses F-score method to filter redundant and useless data on the original dataset, and then optimizes the filtered data subset using the improved artificial fish swarm algorithm, and uses NB classifiers to evaluate the results. The classification accuracy is 99.88% on CNS, 92.88% on Breast, 95.50% on Colon and 98.78% on Lung. The other four datasets achieved a classification accuracy of 100%. In addition, the effectiveness of the improvement strategy is proved by comparison with using different evaluation measures of the standard FSA and others advanced methods. Further, the most significant result is obtained by FSA-PV in Breast dataset.

Although the proposed feature selection framework has achieved significant results in most datasets, considering the improvement strategies of the entire algorithm, the experimental process, and the results obtained, the proposed method still has the drawback of slow optimization speed. This problem is mainly reflected in the evaluation process of the embedded method on the extracted subsets, where each evaluation means a training process of the machine learning algorithm, which greatly increases the overall time complexity. In addition, there are some datasets such as Lung, Colon and Breast have not achieved the 100% accuracy of classification because the F-score method and NB method have a low degree of fitting for these datasets. From table 5-7, we can see the KNN and MI have better results in these 3 datasets clearly. Therefore, to solve them, our further work will focus on adaptive improving of embedded method and filter method so that the framework can fit more datasets.

REFERENCES

[1] S. Michiels, S. Koscielny, and C. Hill, "Interpretation of microarray data in cancer," *Brit. J. Cancer*, vol. 96, no. 8, pp. 1155–1158, Apr. 2007.

[2] A. K. Shukla, P. Singh, and M. Vardhan, "Gene selection for cancer types classification using novel hybrid metaheuristics approach," *Swarm Evol. Comput.*, vol. 54, May 2020, Art. no. 100661.

[3] C. Qu, L. Zhang, J. Li, F. Deng, Y. Tang, X. Zeng, and X. Peng, "Improving feature selection performance for classification of gene expression data using Harris Hawks optimizer with variable neighborhood learning," *Briefings Bioinf.*, vol. 22, no. 5, Sep. 2021, Art. no. bbab097.

[4] S. Tabakhi, A. Najafi, R. Ranjbar, and P. Moradi, "Gene selection for microarray data classification using a novel ant colony optimization," *Neurocomputing*, vol. 168, pp. 1024–1036, Nov. 2015.

[5] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Comput. Biol. Med.*, vol. 140, Jan. 2022, Art. no. 105051.

[6] K. Kanti Ghosh, S. Begum, A. Sardar, S. Adhikary, M. Ghosh, M. Kumar, and R. Sarkar, "Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114485.

[7] H. N. Alshaer, M. A. Otair, L. Abualigah, M. Alshinwan, and A. M. Khasawneh, "Feature selection method using improved CHI square on Arabic text classifiers: Analysis and application," *Multimedia Tools Appl.*, vol. 80, no. 7, pp. 10373–10390, Mar. 2021.

[8] H. Lim and D.-W. Kim, "MFC: Initialization method for multi-label feature selection based on conditional mutual information," *Neurocomputing*, vol. 382, pp. 40–51, Mar. 2020.

[9] Y. W. Chen and C. J. Lin, *Combining SVMs With Various Feature Selection Strategies* (Feature extraction). Berlin, Germany: Springer, 2006, pp. 315–324.

[10] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, Apr. 2005.

[11] T. N. Lal, O. Chapelle, and J. Weston, "Embedded methods," in *Feature Extraction*. Berlin, Germany: Springer, 2006, pp. 137–165.

[12] T. N. Nuklianggraita, A. Adiwijaya, and A. Aditsania, "On the feature selection of microarray data for cancer detection based on random forest classifier," *Jurnal Infotel*, vol. 12, no. 3, pp. 89–96, 2020.

[13] C. Kang, Y. Huo, and L. Xin, "Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine," *J. Theor. Biol.*, vol. 463, pp. 77–91, Jul. 2019.

[14] R. Kohavi and G. H. John, "The wrapper approach," in *Feature Extraction, Construction and Selection*. Boston, MA, USA: Springer, 1998, pp. 33–50.

[15] S. Sayed, M. Nassef, and A. Badr, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets," *Expert Syst. Appl.*, vol. 121, pp. 233–243, Jul. 2019.

[16] M. Ye, W. Wang, and C. Yao, "Gene selection method for microarray data classification using particle swarm optimization and neighborhood rough set," *Current Bioinf.*, vol. 14, no. 5, pp. 422–431, 2019.

[17] X.-F. Song, Y. Zhang, D.-W. Gong, and X.-Z. Gao, "A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9573–9586, Sep. 2022.

[18] Y. Hu, Y. Zhang, and D. Gong, "Multiobjective particle swarm optimization for feature selection with fuzzy cost," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 874–888, Feb. 2021.

[19] C. Chen, Y. Wan, A. Ma, L. Zhang, and Y. Zhong, "A decomposition-based multiobjective clonal selection algorithm for hyperspectral image feature selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541516.

[20] R. Jiao, B. Xue, and M. Zhang, "Benefiting from single-objective feature selection to multiobjective feature selection: A multiform approach," *IEEE Trans. Cybern.*, vol. 53, no. 12, pp. 7773–7786, Dec. 2023.

[21] K. Moorthy and M. S. Mohamad, "Random forest for gene selection and microarray data classification," in *Knowledge Technology Week*. Berlin, Germany: Springer, 2011, pp. 174–183.

[22] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Inf. Sci.*, vol. 179, no. 13, pp. 2208–2217, Jun. 2009.

[23] Y. Wang, I. V. Tetko, and M. A. Hall, "Gene selection from microarray data for cancer classification-a machine learning approach," *Comput. Biol. Chem.*, vol. 29, no. 1, pp. 37–46, 2005.

[24] A. Halder and A. Kumar, "Active learning using rough fuzzy classifier for cancer prediction from microarray gene expression data," *J. Biomed. Informat.*, vol. 92, Apr. 2019, Art. no. 103136.

[25] B. I. Grisci, B. C. Feltes, and M. Dorn, "Neuroevolution as a tool for microarray gene expression pattern identification in cancer research," *J. Biomed. Informat.*, vol. 89, pp. 122–133, Jan. 2019.

[26] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene selection and classification of microarray data using convolutional neural network," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, Oct. 2018, pp. 145–150.

[27] S. Kilicarslan, K. Adem, and M. Celik, "Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network," *Med. Hypotheses*, vol. 137, Apr. 2020, Art. no. 109577.

[28] Y. N. Liu, G. Wang, and H. L. Chen, "An improved particle swarm optimization for feature selection," *J. Bionic Eng.*, vol. 8, no. 2, pp. 191–200, 2011.

[29] A. K. Shukla, P. Singh, and M. Vardhan, "A hybrid gene selection method for microarray recognition," *Biocybernetics Biomed. Eng.*, vol. 38, no. 4, pp. 975–991, 2018.

[30] H. Wang, X. Jing, and B. Niu, "A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data," *Knowl.-Based Syst.*, vol. 126, pp. 8–19, Jun. 2017.

[31] E. Pashaei, E. Pashaei, and N. Aydin, "Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization," *Genomics*, vol. 111, no. 4, pp. 669–686, Jul. 2019.

[32] X. L. Li, "An optimizing method based on autonomous animats: Fish-swarm algorithm," *Syst. Eng.-Theory Pract.*, vol. 22, no. 11, pp. 32–38, 2002.

[33] S. Q. Ye, K. Q. Zhou, A. M. Zain, F. L. Wang, and Y. Yusoff, "A modified harmony search algorithm and its applications in weighted fuzzy production rule extraction," *Frontiers Inf. Technol. Electron. Eng.*, vol. 24, no. 11, pp. 1574–1590, 2023.

[34] S. Q. Ye, K. Q. Zhou, C. X. Zhang, A. M. Zain, and Y. Ou, "An improved multi-objective cuckoo search approach by exploring the balance between development and exploration," *Electronics*, vol. 11, no. 5, p. 704, 2022.

[35] X.-Y. Zhang, K.-Q. Zhou, P.-C. Li, Y.-H. Xiang, A. M. Zain, and A. Sarkheyli-Hagele, "An improved chaos sparrow search optimization algorithm using adaptive weight modification and hybrid strategies," *IEEE Access*, vol. 10, pp. 96159–96179, 2022.

[36] A. Yaqoob, N. K. Verma, and R. M. Aziz, "Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm," *J. Med. Syst.*, vol. 48, no. 1, p. 10, Jan. 2024.

[37] A. A. Joshi and R. M. Aziz, "Deep learning approach for brain tumor classification using metaheuristic optimization with gene expression data," *Int. J. Imag. Syst. Technol.*, vol. 34, no. 2, 2024, Art. no. e23007.

[38] R. Mahto, S. U. Ahmed, R. U. Rahman, R. M. Aziz, P. Roy, S. Mallik, A. Li, and M. A. Shah, "A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection," *BMC Bioinf.*, vol. 24, no. 1, p. 479, Dec. 2023.

[39] A. A. Joshi and R. M. Aziz, "A two-phase cuckoo search based approach for gene selection and deep learning classification of cancer disease using gene expression data with a novel fitness function," *Multimedia Tools Appl.*, pp. 1–32, 2024.

[40] Y. Ou, P. Yin, and L. Mo, "An improved grey wolf optimizer and its application in robot path planning," *Biomimetics*, vol. 8, no. 1, p. 84, 2023.

[41] E. Babaee Tirkolaee, A. Goli, and G.-W. Weber, "Fuzzy mathematical programming and self-adaptive artificial fish swarm algorithm for just-in-time energy-aware flow shop scheduling problem with outsourcing option," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, pp. 2772–2783, Nov. 2020.

[42] X.-Y. Luan, Z.-P. Li, and T.-Z. Liu, "A novel attribute reduction algorithm based on rough set and improved artificial fish swarm algorithm," *Neurocomputing*, vol. 174, pp. 522–529, Jan. 2016.

[43] Y. Chen, Z. Zeng, and J. Lu, "Neighborhood rough set reduction with fish swarm algorithm," *Soft Comput.*, vol. 21, no. 23, pp. 6907–6918, Dec. 2017.

[44] R. P. S. Manikandan and A. M. Kalpana, "Feature selection using fish swarm optimization in big data," *Cluster Comput.*, vol. 22, no. 5, pp. 10825–10837, 2019.

[45] W.-H. Tan and J. Mohamad-Saleh, "Normative fish swarm algorithm (NFSA) for optimization," *Soft Comput.*, vol. 24, no. 3, pp. 2083–2099, Feb. 2020.

[46] Y. Chen, Q. Zhu, and H. Xu, "Finding rough set reducts with fish swarm algorithm," *Knowl.-Based Syst.*, vol. 81, pp. 22–29, Jun. 2015.

[47] K. M. Leung, "Naive Bayesian classifier," Dept. Comput. Sci./Finance Risk Eng., Polytech. Univ., 2007, pp. 123–156, vol. 2007.

[48] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jul. 1967.

[49] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis—A brief tutorial," *Inst. Signal Inf. Process.*, vol. 18, no. 1998, pp. 1–8, 1998.

[50] G. De'ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, p. 3178, Nov. 2000.

[51] H.-F. Yu, F.-L. Huang, and C.-J. Lin, "Dual coordinate descent methods for logistic regression and maximum entropy models," *Mach. Learn.*, vol. 85, nos. 1–2, pp. 41–75, Oct. 2011.

[52] A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm," *Exp. Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114012.

[53] S. L. Pomeroy, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, Jan. 2002.

[54] L. J. Van't Veer, H. Dai, and M. J. Van De Vijver, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[55] A. Sharma and R. Rani, "C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods," *Comput. Methods Programs Biomed.*, vol. 178, pp. 219–235, Sep. 2019.

[56] A. Kulkarni, B. S. C. N. Kumar, V. Ravi, and U. S. Murthy, "Colon cancer prediction with genetics profiles using evolutionary techniques," *Exp. Syst. Appl.*, vol. 38, no. 3, pp. 2752–2757, Mar. 2011.

[57] O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar, Z. A. A. Alyasseri, I. A. Doush, A. K. Abasi, M. A. Awadallah, and R. A. Zitar, "Gene selection for microarray data classification based on gray wolf optimizer enhanced with TRIZ-inspired operators," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 107034.

[58] A. Dabba, A. Tari, and S. Meftali, "Hybridization of moth flame optimization algorithm and quantum computing for gene selection in microarray data," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 2731–2750, Feb. 2021.

ZONG-ZHENG LI was born in Yiyang, China, in 1997. He received the B.E. degree in communications engineering and the M.E. degree in electronics and communications engineering from Jishou University, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in computer science with Universiti Teknologi Malaysia. His research interests include machine learning, soft computing, and evolutionary computation.

FANG-LING WANG was born in Yongcheng, China, in 1996. He received the B.E. degree in communications engineering and the M.E. degree in electronics and communications engineering from Jishou University, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in computer science with Universiti Teknologi Malaysia. His research interests include machine learning, disease diagnosis, and evolutionary computation.

FENG QIN was born in Changsha, China, in 1994. He received the B.E. degree in computer science and technology from Jishou University, in 2018, and the M.E. degree in human–computer interaction from the University of Nottingham, in 2019. He is currently pursuing the Ph.D. degree in computer science with Universiti Teknologi Malaysia. His research interests include fuzzy Petri nets and its applications, and knowledge graphs.

YUSLIZA BINTI YUSOFF received the Ph.D. degree in computer science from Universiti Teknologi Malaysia, in 2017. She is currently a Senior Lecturer with the Faculty of Computing, Universiti Teknologi Malaysia. Her main research interests include computational intelligence, modeling and optimization, and artificial intelligence.

AZLAN MOHD ZAIN (Member, IEEE) received the Ph.D. degree in computer science from Universiti Teknologi Malaysia, in 2010. He is currently a Professor of computer science with Universiti Teknologi Malaysia. His main research interests include artificial intelligence, modeling and optimization, machining, and statistical process control.