scientific reports

Check for updates

OPEN

A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning

Junyao Ding[1], Jianchao Du[1]✉, Hejie Wang[1] & Song Xiao[2]

The data acquisition methods are becoming increasingly diverse and advanced, leading to higher data dimensions, blurred classification boundaries, and overfitting datasets, affecting machine learning models' accuracy. Many studies have sought to improve model performance through feature selection. However, a single feature selection method has incomplete, unstable, or time-consuming shortcomings. Combining the advantages of various feature selection methods can help overcome these defects. This paper proposes a two-stage feature selection method based on random forest and improved genetic algorithm. First, the importance scores of the random forest are calculated and ranked, and the features are preliminarily eliminated according to the scores, reducing the time complexity of the subsequent process. Then, the improved genetic algorithm is used to search for the global optimal feature subset further. This process introduces a multi-objective fitness function to guide the feature subset, minimizing the number of features in the subset while enhancing classification accuracy. This paper also adds an adaptive mechanism and evolution strategy to improve the loss of population diversity and degeneration in the later stages of iteration, thereby enhancing search efficiency. The experimental results on eight UCI datasets show that the proposed method significantly improves classification performance and has excellent feature selection capability.

Keywords Machine learning, Data mining, Feature selection, Random forest, Improved genetic algorithm

Data mining has gained significant popularity across various industries with the diversification of data acquisition methods and the rapid growth of data volume. Applications include the classification of brain tumors in medical treatment[1], the optimization of chemical experimental processes[2], the analysis of biological data such as gene sequences[3], the prediction of crop yields[4], and the modelling and analysis of energy systems in buildings[5]. However, as the dimensionality of the data increases, the boundaries between categories become increasingly blurred. Additionally, many irrelevant and redundant features can lead to overfitting of the dataset, causing machine learning models to fail in achieving the desired outcomes[6,7]. To enhance the model's classification performance and mitigate the effects of the curse of dimensionality, feature selection has become a widely used technical tool.

The concept of feature selection involves identifying the most effective features from the original feature set to enhance the model's accuracy while minimizing the number of features in subsets[8]. Based on the various mechanisms, feature selection can be divided into three types: filter, wrapper, and embedded. Filtering methods calculate the characteristic indices of features, rank them, and assess their advantages or disadvantages based on this ranking. Common techniques include mutual information[9], correlation coefficient[10], and relief[11]. The primary advantage of these methods is their simplicity and speed. However, they also have significant drawbacks. They overlook the dependencies between features, which can result in the exclusion of valid feature combinations. Additionally, they fail to consider the interaction with the classifier, leading to a disconnection between searching for subsets and the hypothesis space, making it difficult to select features exactly beneficial to classifiers. Unlike filtering methods, wrapper methods utilize machine learning models to evaluate feature subsets. Common algorithms include recursive feature elimination[12] and swarm intelligence algorithms[13–15].

[1]School of Telecommunications Engineering, Xidian University, Xi'an 710071, China. [2]Beijing Electronic Science and Technology Institute, Beijing 100070, China. ✉email: jcdu@xidian.edu.cn

These methods consider the dependencies between features and the interactions between the search process and the model. As a result, the classification performance of the selected subset is often superior. However, these methods require the repeated construction of machine learning models through multiple rounds of iteration, training, and validation, leading to high time complexity, which makes them unsuitable for high-dimensional datasets. Embedded methods integrate the benefits of both filter and wrapper methods by incorporating the feature selection process directly into model training. These approaches facilitate rapid searching and allow for interaction with the models, enabling the identification of feature subsets within the hypothesis space. Common examples of embedded methods include lasso[16], decision tree[17], and random forest[18]. However, a notable drawback of embedded methods is that the selected feature subsets are often overly dependent on the specific model used. Thus, features that perform well for one model may not yield the same performance in another.

As mentioned above, each type of feature selection method has its advantages and disadvantages, and applying a single method often presents certain limitations. Combining the strengths of multiple feature selection methods can mitigate the shortcomings inherent in any one approach. In recent years, multi-stage feature selection methods have garnered increasing attention from researchers. Wang et al.[19] proposed a method that first screened features by calculating the weighted sum of the maximum information coefficient and fisher score. They then selected effective features using a sequential forward selection to achieve automatic sleep staging. Xu et al.[20] utilized the fisher score with a heuristic algorithm based on self-information uncertainty measures to filter out valid features sequentially. Su et al.[21] initially selected features through mutual information gain, then updated a subset of features using a recursive feature elimination method. They recorded the number of times each feature was selected throughout the process and ultimately chose the features that had been selected more than half the maximum number of iterations through a voting mechanism. These multi-stage feature selection methods offer superior screening compared to single methods. However, they do not consider how to obtain the optimal solution from the entire feature set, thus limiting the algorithm's performance to some extent. Specifically, the above methods of the first stage calculate the relationship between individual features and the target variable, which may be affected by outliers. This may introduce bias that is difficult to correct in subsequent processes, thus affecting the final feature selection results. For the second stage, the sequential forward selection in reference[19] and the recursive feature elimination in reference[21] are greedy strategies that only consider the optimal choice at the current step, making it difficult to localize optima. The method in reference[20] relies on specific heuristic criteria, which may not fully describe the complex relationships between features, thereby limiting the discovery of the globally optimal feature subset.

Based on the above research, we propose a new two-stage feature selection framework to search for the optimal feature subset from the global. In the first stage, we use random forest to initially remove irrelevant features. The random forest is not only fast, but also, due to the significance of evaluating features across multiple decision trees and the inherent randomness of the algorithm, the screening results are not easily influenced by outliers. Additionally, it effectively handles nonlinear features and high-dimensional data[22,23]. In the second stage, we propose an improved genetic algorithm that eliminates the redundant features retained from the previous step and searches for the optimal feature subset globally. By combining the strengths of both approaches, we aim to address the limited improvements associated with one-stage feature selection methods.

The main contributions of this paper are summarized as follows.

·This paper constructs a new feature selection framework. Firstly, features with low contribution to categorization are deleted based on the variable importance measure of the random forest, which also reduces the computation time of the subsequent process. Then, we use the improved genetic algorithm by modeling feature selection as a minimization problem and conducting a further search for the optimal feature subset.

·Since feature selection aims to minimize the size of the feature subset while ensuring optimal classification performance, this paper establishes a multi-objective fitness function in the improved genetic algorithm.

·This paper introduces an improved genetic algorithm to address potential diversity loss and degradation in the genetic algorithm. This enhancement is achieved through the incorporation of an adaptive mechanism for crossover and mutation, alongside the implementation of a $\mu+1$ evolutionary strategy.

The structure of the remaining content is as follows.

Section 2 outlines the feature selection framework of the random forest, the genetic algorithm, and the improved genetic algorithm in turn, culminating in the two-stage feature selection method proposed in this paper. Section 3 introduces the experiment settings, including the dataset utilized, evaluation metrics, and the parameters set for the algorithm. Section 4 is the experimental part, which shows the experimental results and analysis, and the final Sect. 5 offers a summary and future perspectives.

Method

The random forest feature selection

The random forest is characterized by aggregating votes from multiple decision trees based on the gini coefficient[24]. Beginning at the root node, each decision tree calculates the gini coefficient for the features, vertically partitioning the dataset to create successor nodes until a predetermined depth is reached. The gini coefficient quantifies sample impurity; thus, a lower value indicates a purer node, making the corresponding feature more advantageous for classification. The variable importance measure (VIM) score assesses the ability of features to reduce the gini coefficient and serves as a crucial metric for feature selection in the random forest. In contrast to the gini coefficient, a higher VIM score for a feature indicates a greater capacity to differentiate between distinct categories. Conversely, a lower score suggests that the feature is less effective in classification and may need to be eliminated. For feature x_j, the steps to calculate the VIM score are as follows:

(1) Calculate the gini coefficient of the feature x_j at the decision tree node:

$$\mathrm{Gini}(x_j) = \sum_{i=1}^{k} p_i(1-p_i) = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

Where k denotes the number of classes and p_i is the probability that the sample belongs to the ith class.

(2) Calculate the VIM score of the feature x_j at the decision tree node: Let node n of the decision tree contain the feature x_j and the left and right successor nodes of node n are node l and node r. The VIM score of the feature x_j at node n is:

$$\mathrm{VIM}_{jn}^{(\mathrm{Gini})} = \mathrm{GI}_n - \mathrm{GI}_l - \mathrm{GI}_r \tag{2}$$

where $\mathrm{GI}_n, \mathrm{GI}_l, \mathrm{GI}_r$ denote gini coefficients of the feature x_j at node n, left successor node l and right successor node r, respectively.

(3) Calculate the VIM score of the feature x_j in the decision tree: Let the set N be all the nodes of the feature x_j that have appeared in the ith decision tree of the random forest. Then the VIM score of the feature x_j in the decision tree is the sum of the VIM scores of all the nodes in the set N:

$$\mathrm{VIM}_{ij}^{(\mathrm{Gini})} = \sum_{n \in N} \mathrm{VIM}_{jn}^{(\mathrm{Gini})} \tag{3}$$

(4) Calculate the VIM score of the feature x_j in the random forest: Let t decision trees in the random forest. Then, the VIM score of the feature x_j in the random forest is the sum of its scores in all decision trees:

$$\mathrm{VIM}_{j}^{(\mathrm{Gini})} = \sum_{i=1}^{t} \mathrm{VIM}_{ij}^{(\mathrm{Gini})} \tag{4}$$

(5) Normalization: After calculating the VIM scores of all features respectively, the VIM score of each feature is normalized. For the feature x_j, the normalized VIM score is:

$$\mathrm{VIM}_j = \frac{\mathrm{VIM}_j^{(\mathrm{Gini})}}{\sum_{i=1}^{m} \mathrm{VIM}_i^{(\mathrm{Gini})}} \tag{5}$$

where m is the total number of features.To better understand the computation of the VIM score, Algorithm 1 illustrates the whole progress of how to get the VIM score in the random forest.

```
Initialize: the trained random forest model and the dataset
for each feature in the dataset do
    for each decision tree in the forest do
        for each node in the decision tree do
            Calculate the Gini coefficient of the node according to Eq. (1)
            Calculate the Gini coefficients of the left successor node and the
right successor node according to Eq. (1)
            Calculate the VIM score of the node according to Eq. (2)
        end for
        Calculate the VIM score of the feature in the decision tree according to
Eq. (3)
    end for
    Calculate the VIM score of the feature in the random forest according to Eq. (4)
end for
Get VIM scores of all features in the random forest
Normalize VIM scores according to Eq. (5)
```

Algorithm 1 . Pseudocode for the VIM score calculation process.

After calculating the VIM score, the features are ranked in descending order based on their scores. There are two strategies for selecting features. The first involves selecting a specific number of features according to their ranking from highest to lowest, while the second entails selecting features with importance scores that exceed a predetermined threshold. The choice between the two strategies is application-dependent. The first one should be chosen when the model has a predefined number of input features. Based on the threshold, the second one is more flexible, retaining only the features that significantly contribute to the model, although this may result in different numbers of features between datasets.

The genetic algorithm feature selection

The genetic algorithm[25] is based on the principles of heredity, mutation, natural selection, and hybridization from evolutionary biology. It serves as a search algorithm designed to solve optimization problems in computational mathematics, enabling the identification of globally optimal solutions. Feature selection can be framed as an optimization problem. Figure 1 illustrates the genetic algorithm's application to feature selection.

Initialization

Feature subsets are represented in genetic algorithms using binary encoding. An encode of "1" means the feature in the corresponding position is selected, while "0" means not selected. Initialize n individuals as the initial population, denoted by $s_1 - s_n$.

Fitness calculation

First, map the individual into the feature subset. For example, in Fig. 3, an individual's genotype is "00010," and assuming the full feature set is "abcde," the feature subset for this individual is "d."

Next, calculate the individual's fitness by the fitness function. Feature selection aims to filter as many features as possible while ensuring classification accuracy. Therefore, the fitness function in this paper contains two parts: the cost of classification error and the evaluation of the number of features in the subset. The formula is as follows:

$$f(s_i) = \alpha \times (1 - accuracy) + \beta \times \frac{n}{N}, 1 \leqslant i \leqslant n \tag{6}$$

where $\alpha, \beta \in (0,1)$ are coefficients measuring the cost of classification error and the evaluation of the number of features in the subset, respectively. Accuracy is obtained by a machine learning model evaluating the feature subset. n is the number of features in the feature subset, and N is the total number of features in the full set.

The genetic algorithm optimizes the accuracy of a feature subset and the number of features in a subset as its objective. The lower the cost of classification error, the better the classification performance of the feature subset. Additionally, the smaller the evaluation of the number of features in the subset, the fewer features be selected. Therefore, searching for the feature subset is modelled as a minimization problem: A smaller fitness value indicates a better individual.

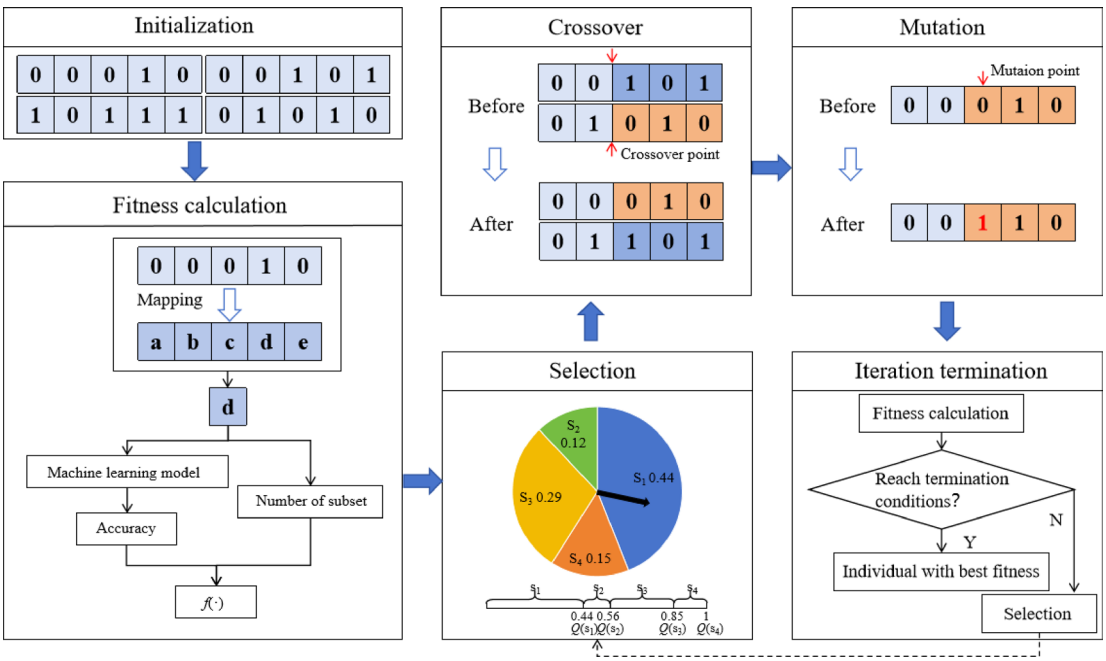


Fig. 1. The image illustrates the process of feature selection by the genetic algorithm.

Selection

n individuals are randomly selected as parents from the current population based on fitness. In this paper, we adopt a roulette mechanism for selection. Firstly, amplify the fitness value $l(s_i) = \frac{1}{1+f(s_i)}$, secondly calculate the individual selection probability $P(s_i) = \frac{l(s_i)}{\sum_{j=1}^{n} l(s_j)}$, then calculate the cumulative probability $Q(s_i) = \sum_{k=1}^{i} P(s_k)$, and finally take a random number in the interval $[0,1]$. The individual corresponding to the cumulative probability interval in which the number falls is selected.

In this step, individuals with lower fitness are more likely to be selected, meaning better feature combinations can be retained. The selection is done at n times, and all the selected individuals will proceed to the next step.

Crossover

Simulating the process of hybridization in biological evolution, exchange portions of the genes between two individuals, thereby generating new individuals that combine both characteristics. In feature selection, it is potential to yield combinations of features that are effective for classification.

Mutation

Simulating the process of genetic mutations in biological evolution, individual genes are altered randomly based on probability. This generates new gene information, which may guide the search towards the globally optimal feature subset.

Iteration termination

Recalculate the fitness of the new population and assess whether the termination conditions (reaching the maximum number of iterations or observing no further changes in population fitness) have been met. If the condition is satisfied, terminate the iteration and map the individual with optimal fitness to the feature space to obtain the feature subset. If the terminating condition is not met, return to step (3) and proceed with the next iteration.

The improved genetic algorithm feature selection

Although the genetic algorithm can search for optimal solutions on a global scale, it often experiences a loss of diversity[26] and degradation[27]. The roles of crossover and mutation are to introduce new genotypes and genetic information, thereby preserving the diversity of the population and preventing the algorithm from converging on a local optimum. However, these two processes are governed by probabilistic mechanisms. As iterations progress, the population tends to homogenize, with individuals becoming increasingly similar, which hampers the crossover process's ability to generate new genotypes. The mutation probability is typically set to a very low value to prevent the population from changing too randomly and losing its search direction. This limitation can hinder the timely introduction of new genes, resulting in the population's inability to escape local optima. Furthermore, newly generated individuals may not necessarily outperform their parents during the iterative process, leading to degradation. To address these challenges, this paper proposes the following improvements.

(1) Adaptive mechanism: The concept involves the dynamic adjustment of probability parameters. Specifically, the probabilities of crossover and mutation are increased by fitness values and the diversity of the population. The probabilities of crossover and mutation are denoted as P_c and P_m, respectively. The formulas for adaptive adjustment are as follows:

$$P_c = \begin{cases} k_1, f' < f_{avg} \\ k_2 \dfrac{(f' - f_{\min})(1 + e^{-c_1 f_{std}})}{f_{avg} - f_{\min}}, f' \geq f_{avg} \end{cases} \tag{7}$$

$$P_m = \begin{cases} k_3, f < f_{avg} \\ k_4 \dfrac{(f - f_{min})(1 + e^{-c_2 f_{std}})}{f_{avg} - f_{min}}, f \geqslant f_{avg} \end{cases} \tag{8}$$

where f' represents the parent's fitness with the superior fitness value among the two parents involved in the crossover operation. The variable f denotes the fitness of the individual subject to mutation during the mutation operation. $f_{\min}, f_{avg}, f_{std}$ correspond to the minimum, mean, and standard deviation of the population's fitness. Additionally, $k_1, k_2, k_3, k_4 \in (0,1]$ and c_1, c_2 are constants greater than zero.

(2) $\mu+1$ evolution strategy: Let the population size be denoted as μ, with 1 offspring generated in each generation. The best μ individuals are selected to form the next generation from the total of $\mu+1$ individuals, including parents and offspring. This method preserves the top individuals in each generation, accelerates the population's convergence speed, and enhances the algorithm's search performance.

In summary, the process of improved genetic algorithm feature selection is shown in Fig. 2. Moreover, Algorithm 2 better demonstrates the algorithm's workflow.

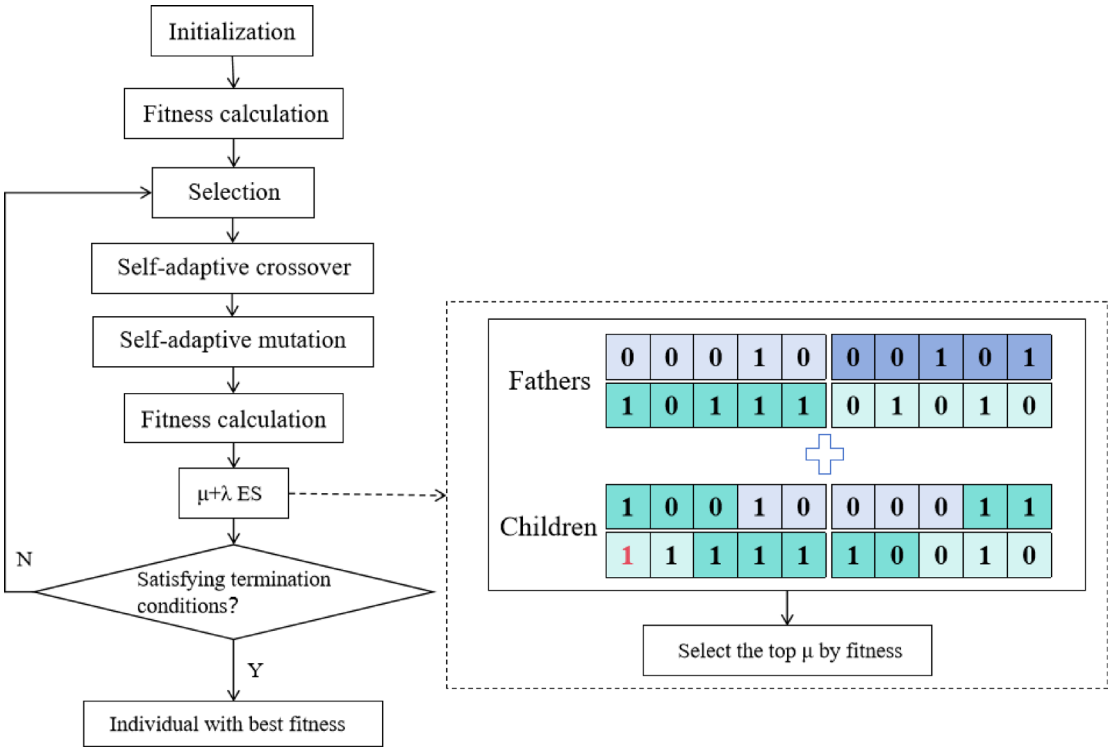


Fig. 2. The framework of the improved genetic algorithm feature selection.

Initialize: population A with μ individuals, and maximum number of iterations n
for each individual in population A **do**
    Calculate the fitness of each individual according to Eq. (6).
end for
$i = 0$
if $i < n$ **then**
    for each individual in population A **do**
        Select two parents by roulette wheel selection
        Adjust crossover probability by Eq.(7)
        Implementate crossover
    Adjust mutation probability by Eq.(8)
        Implementate mutation
    end for
    Get the population B after crossover and mutation.
    for each individual in population B **do**
        Calculate the fitness of each individual according to Eq. (6)
    end for
    Mix populations A and B, selecting the optimal λ individuals by fitness as the new population A
    $i = i+1$
end if
Select the best individual from the finished population by fitness as the final result

Algorithm 2 . Pseudocode for IGA.

The two-stage feature selection method

The random forest randomly selects features and samples to construct decision trees, effectively reducing the influence of outliers and ensuring that the screening results are stable and reliable. Furthermore, the random forest incorporates feature selection within the modeling process, making it fast and efficient. However, it is

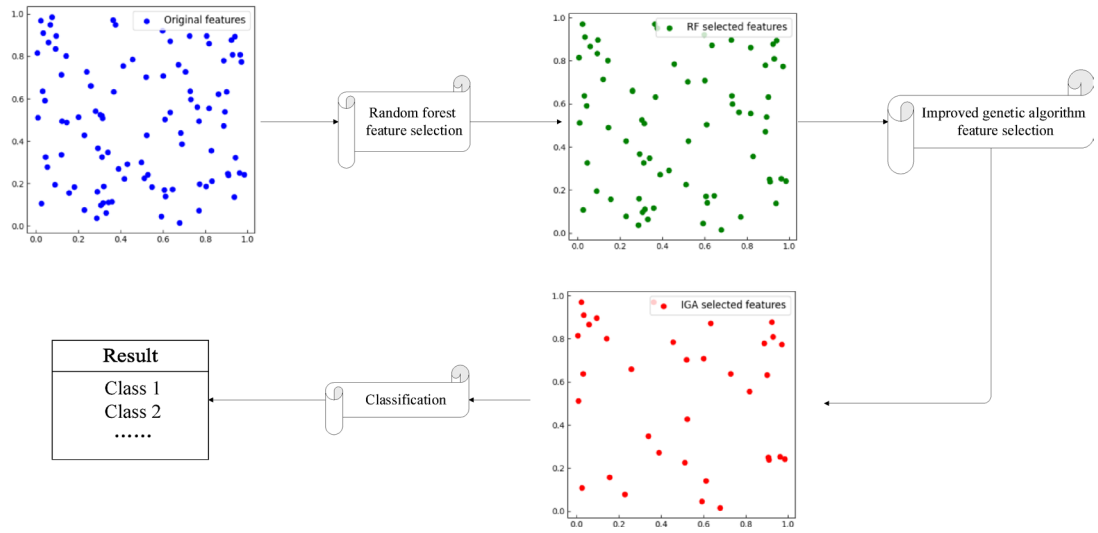


Fig. 3. The framework of the two-stage feature selection method.

1: Initialize: a dataset with m features
2: Train a random forest by the dataset
3: Calculate VIM scores for m features according to Algorithm 1
4: Based on the scores, rank the features in descending order
5: Select m_1 features by VIM scores to obtain the feature subset F_1
6: Based on F_1, the global optimal solution is searched iteratively according to Algorithm 2
7: The search is finished and the feature subset F_2 with m_2 features is obtained

Algorithm 3 . Pseudocode for the two-stage feature selection method.

important to note that the random forest cannot effectively eliminate redundant features with high VIM scores[28], necessitating a secondary screening process. The improved genetic algorithm's search mechanism retains the most effective feature subset, eliminating redundant features. Therefore, further feature selection is conducted using the improved genetic algorithm.

Above all, the two-stage feature selection method based on random forest and improved genetic algorithm (RFIGA) consists of two main steps: (1) using the variable importance measure (VIM) from random forest to eliminate features, thereby reducing the search time in the subsequent process and mitigating data overfitting; (2) based on the features retained from the first stage, the improved genetic algorithm further filters the features to identify the globally optimal subset. The algorithm framework is illustrated in Fig. 3.

For a clearer understanding of the algorithmic process, Algorithm 3 provides the pseudocode for the two-stage feature selection method.

Algorithm complexity analysis

The proposed method combined with random forest and improved genetic algorithm, so an initial analysis of the complexity of sub-modules is conducted.

The time complexity of the VIM score calculated through the gini coefficient in a random forest is primarily determined by the training phase of the decision trees. During each split in a decision tree, the gini coefficient must be computed for every feature, which requires sorting the values. This sorting process incurs a complexity of $O(n \times \log n)$, where n represents the number of samples. When considering m features, the complexity for each split increases to $O(m \times n \times \log n)$. Assuming the depth of the decision tree is d, the overall time complexity can be expressed as $O(m \times n \times \log n \times d)$. If there are T trees in the random forest, the cumulative training time complexity becomes $O(T \times m \times n \times \log n \times d)$. After the training of the random forest is completed, the complexity of calculating the VIM score for m features is $O(T \times m)$. Given that the complexity of the VIM score is primarily influenced by the training process, the overall time complexity can be characterized as $O(T \times m \times n \times \log n \times d)$.

Let μ be the population size, the dataset contains n samples and m features. Using logistic regression for evaluation, the complexity of the fitness function is $O(m \times n)$, so the complexity of calculating fitness for whole population is $O(\mu \times m \times n)$. The complexity of the roulette wheel selection mechanism is $O(\mu)$. In crossover and mutation, assuming there are g genes involved, the time complexity is $O(\mu \times g)$. After that, the parents and offspring are merged, and from a total of μ+λ individuals, the top μ optimal individuals are selected using

Datasets	Instances	Features	Classes
Glioma	839	23	2
Dermatology	366	33	6
Audiology	226	69	24
Movement Libras	360	90	15
Arrhythmia	452	279	13
Darwin	174	451	2
Period Changer	90	1177	2
Toxicity	171	1203	2

Table 1. Dataset introduction.

Ground truth \ Prediction	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Fig. 4. Confusion matrix.

quicksort, with a time complexity of $O((\mu+\lambda)\times\log(\mu+\lambda))$. This process is iterated G times, so the overall time complexity is:

$$\begin{aligned} &O(G\times(\mu\times m\times n+\mu+\mu\times g+\mu\times g+(\mu+\lambda)\times\log(\mu+\lambda)))=\\ &O(G\times(\mu\times m\times n+\mu+\mu\times g+(\mu+\lambda)\times\log(\mu+\lambda))) \end{aligned} \tag{9}$$

Finally, let's analyze the complexity of the two-stage feature selection algorithm. Assuming the dataset has n samples and m_1 features, after random forest feature selection, there are m_2 remaining features. The time complexity of the proposed algorithm is:

$$O(T\times m_1\times n\times\log n\times d)+O(G\times(\mu\times m_2\times n+\mu\times g+(\mu+\lambda)\times\log(\mu+\lambda))) \tag{10}$$

Experimental settings

Dataset

In this paper, we selected eight datasets from the UCI database[29] to validate the proposed method. The datasets include Glioma, Dermatology, Audiology, Movement libras, Arrhythmia, Darwin, Period Changer, and Toxicity. The feature dimensions of these datasets range from 23 to 1203, and the classification varies from binary to multi-class, effectively encompassing datasets with diverse characteristics. Detailed information is presented in Table 1.

Evaluation metric

Experiments in this paper use ratio and accuracy for evaluation.

The ratio denotes the ability of the feature selection method to filter features and is calculated as follows:

$$\text{Ratio}=1-\frac{n}{N} \tag{11}$$

where n denotes the number of features in the subset, and N denotes the number of features in the full feature set.

Accuracy[30] (ACC) evaluates the classification effectiveness and is calculated from the confusion matrix. The confusion matrix is shown in Fig. 4.

Accuracy indicates the proportion of correctly predicted samples to the overall sample and is calculated using the following formula:

$$\text{ACC}=\frac{1}{n}\sum_{i=1}^{n}\frac{\text{TP}_i+\text{TN}_i}{\text{TP}_i+\text{FN}_i+\text{FP}_i+\text{TN}_i} \tag{12}$$

where n is the number of categories.

Algorithm parameter setting

Table 2 shows the parameter settings of the proposed two-stage feature selection method. The random forest parameters n_estimators and max_depth represent the number of decision trees and the maximum depth of the

Method	Parameter	Value
Random forest	n_estimators	200
	max_depth	3
Improved genetic algorithm	iterations	200
	μ	30
	l	30
	k_1, k_2	0.8
	k_3, k_4	0.01
	c_1, c_2	50
	a	0.99
	b	0.01

Table 2. The setting of algorithm parameters.

Dataset	Index	base	RF	GA	IGA	RFIGA
Glioma	Ratio	\	0.25	0.77	0.86	0.83
	ACC (std)	87.13	**87.48**(0.04)	87.22(0.06)	87.25(0.00)	87.25(0.00)
Dermatology	Ratio	\	0.19	0.51	0.59	0.61
	ACC (std)	96.72	98.33(0.20)	98.76(0.23)	**98.90**(0.18)	**98.90(0.00)**
Audiology	Ratio	\	0.29	0.55	0.67	0.69
	ACC (std)	77.48	79.08(0.22)	79.61(0.46)	**83.32**(0.60)	83.02(0.61)
Movement Libras	Ratio	\	0.19	0.44	0.54	0.51
	ACC (std)	63.61	66.06(0.50)	67.19(0.61)	66.72(0.47)	**69.00**(0.68)
Arrhythmia	Ratio	\	0.53	0.54	0.63	0.78
	ACC (std)	69.25	71.38(0.16)	72.05(0.80)	**74.28(0.35)**	73.90(0.55)
Darwin	Ratio	\	0.55	0.51	0.63	0.69
	ACC (std)	83.33	88.93(0.77)	89.58(2.41)	**98.57(0.41)**	98.14(0.82)
Period Changer	Ratio	\	0.99	0.51	0.90	0.95
	ACC (std)	61.11	75.89(1.39)	74.44(2.16)	82.11(1.69)	**83.00**(2.16)
Toxicity	Ratio	\	0.99	0.51	0.87	0.94
	ACC (std)	53.27	67.72(0.98)	65.24(1.23)	72.68(1.15)	**73.73**(1.58)

Table 3. Evaluation results of ablation experiments. Significance values are in bold. The bold values are the best results.

tree, respectively. Section 2.2 and Sect. 2.3 describe the significance of the parameters of the improved genetic algorithm.

Experimental results and analysis

Ablation experiment

Random forest (RF), genetic algorithm (GA), improved genetic algorithm (IGA), and the proposed method (RFIGA) were compared, and ten-fold cross-validation was taken. Due to the randomness in the computational process of these four algorithms, they were run independently ten times, and the average was taken as a result. We used the logistic regression as the classifier and regarded the classification without feature selection as the baseline. The evaluation results of the ablation experiments are shown in Table 3, with the optimal ACC results bolded.

The algorithms IGA and RFIGA presented in this paper significantly enhance classification performance while minimizing the number of features. RFIGA achieves the highest accuracy on the Dermatology and Movement Libras datasets, as well as on Period Changer and Toxicity, with accuracies of 98.90%, 69.00%, 83.00%, and 73.73%, respectively. These results surpass the baseline by 2.18%, 5.39%, 21.89%, and 20.46%. For the remaining datasets, RFIGA ranks second. In the Audiology dataset, RFIGA is 0.3% less accurate than the best IGA; however, it reduces the number of features by 2% more. In the Arrhythmia dataset, RFIGA is 1.19% lower than IGA in accuracy but has a 15% higher proportion of reduced features. On the Darwin dataset, RFIGA is 0.37% less accurate than IGA, yet it allows for 6% more features to be eliminated.

Comparison experiment

To verify the effectiveness of the proposed method for enhancing classification performance, we compared it with various feature selection techniques. The comparison methods include traditional feature selection approaches such as fisher score[31] (FS), recursive feature elimination[12] (RFE), and L_2 penalty[32]. Additionally, swarm intelligence algorithms were employed, including the whale optimization algorithm[33] (WOA), particle

swarm optimization[34] (PSO), differential evolution[35] (DE), and grey wolf optimizer[36] (GWO). Since swarm intelligence algorithms are stochastic search methods, the results of a single run can be influenced by chance. Therefore, WOA, PSO, DE, and GWO were executed independently ten times, and the average was taken as the final result. The outcomes of the comparison experiments are presented in Table 4.

According to Table 4, our method achieves the highest accuracy across all seven datasets and ranks second for one dataset. Regarding the ratio metrics, our method also leads the list, successfully reducing more than 50% of the features across all datasets. Particularly concerning the Period Changer and Toxicity datasets, which comprise over one thousand features, more than 90% of these features are eliminated, resulting in an accuracy improvement of over 20%.

We also selected four multi-stage feature selection methods for comparison with our method. The method I[19]. calculates the weighted sum of the maximum information coefficient and the fisher score in the first stage and employs a sequential forward selection in the second stage The method II[37] sequentially utilizes mRMR and an improved discrete egg algorithm to achieve feature selection. Method III[38] initially identifies features through mutual information and the reliefF, subsequently combining these selected features into a concatenated subset for the first stage. This subset is then subjected to further exploration utilizing the GWO. Method IV[39] employs a three-step process for feature selection, which includes a variance filter, extremely randomized trees, and the Harris Hawks optimization.

The experimental results are presented in Table 5. It is evident that the accuracy of the proposed method is the highest across nearly all datasets, while also maintaining high feature reduction ratios.

Algorithm convergence comparison

The enhancements to the second-stage improved genetic algorithm aim to increase search efficiency, as evidenced by the convergence process of the fitness values. Consequently, we collected fitness values during the algorithm's iterative process to construct the fitness curve for analyzing and comparing the convergence with various swarm intelligence algorithms. The involved algorithms are WOA, PSO, DE, GWO, GA, IGA, and RFIGA. In this research, the feature selection is framed as a minimization problem, with the algorithm converging to the global optimal solution when the fitness value reaches a low point and stabilizes. A steeper decline in the curve indicates a more efficient search. The fitness curves for each algorithm across the eight datasets are presented in Fig. 5.

As illustrated in Fig. 5, the fitness curves of our method decline more rapidly across all datasets. For the Glioma, Audiology, Arrhythmia, Period Changer, and Toxicity datasets, the fitness curve of RFIGA achieves the lowest convergence position, while it ranks second lowest on the remaining three datasets. In contrast, the algorithm with the lowest convergence position on the Darwin dataset is the IGA proposed in this paper. Overall, our method performs better than all other algorithms evaluated in terms of convergence speed and position.

Experiments of the proposed method on other classifiers

In this paper, four classifiers, support vector machine (SVM), ridge regression (Ridge), gaussian naive bayes (NB), and k-nearest neighbors (KNN), are constructed to evaluate the effectiveness of the proposed two-stage feature selection method across various classifiers. The ratio results are presented in Table 6, while the accuracy results are illustrated in Fig. 6.

All combinations of datasets and classifiers exhibit the ability to exclude more than fifty percent of the features, and this feature filtering capability of the proposed method is particularly pronounced in high-dimensional datasets. For example, in datasets characterized by hundreds of dimensions, such as Arrhythmia and Darwin,

Dataset	Index	base	FS	RFE	L_2	WOA	PSO	DE	GWO	RFIGA
Glioma	Ratio	\	0.00	0.65	0.65	0.87	0.76	0.85	0.87	0.83
	ACC(std)	87.13	**87.25**	**87.25**	**87.25**	87.11(0.18)	87.20(0.06)	**87.25(0.00)**	87.22(0.11)	**87.25(0.00)**
Dermatology	Ratio	\	0.00	0.76	0.58	0.34	0.47	0.65	0.69	0.61
	ACC(std)	96.72	97.27	98.36	97.26	98.46(0.19)	95.38(2.08)	98.88(0.09)	**98.90(0.00)**	**98.90(0.00)**
Audiology	Ratio	\	0.06	0.49	0.70	0.26	0.53	0.63	0.73	0.66
	ACC(std)	77.48	77.91	77.47	74.82	81.05(1.38)	81.02(1.03)	**83.37(0.42)**	81.34(1.33)	83.02(0.61)
Movement Libras	Ratio	\	0.09	0.42	0.56	0.17	0.50	0.49	0.61	0.52
	ACC(std)	63.61	66.94	67.50	55.56	65.92(0.38)	68.83(0.75)	68.19(0.27)	68.33(0.41)	**69.00**(0.68)
Arrhythmia	Ratio	\	0.32	0.80	0.80	0.39	0.56	0.53	0.80	0.78
	ACC(std)	69.25	70.81	71.70	72.81	71.53(0.29)	72.73(0.41)	73.68(0.47)	73.26(0.30)	**73.90**(0.55)
Darwin	Ratio	\	0.25	0.90	0.79	0.53	0.54	0.53	0.80	0.67
	ACC(std)	83.33	86.90	93.69	89.05	91.96(1.39)	94.89(1.55)	96.98(0.80)	97.41(1.60)	**98.14**(0.82)
Period Changer	Ratio	\	0.98	≈1	0.99	0.98	0.53	0.53	0.87	0.95
	ACC(std)	61.11	72.22	68.89	71.11	77.78(1.17)	77.11(1.27)	79.22(0.54)	80.44(1.50)	**83.00**(2.16)
Toxicity	Ratio	\	≈1	≈1	≈1	0.98	0.52	0.52	0.79	0.94
	ACC(std)	53.27	67.25	66.67	67.25	69.08(1.16)	68.99(0.79)	69.31(1.05)	70.46(1.37)	**73.73**(1.58)

Table 4. Evaluation results of comparison experiments. Significance values are in bold. The bold values are the best results.

Dataset	Index	Method I	Method II	Method III	Method IV	RFIGA
Glioma	Ratio	0.87	0.34	0.76	0.79	0.83
	ACC(std)	87.25	86.95(0.47)	**87.53**(0.15)	86.40(0.18)	87.25(0.00)
Dermatology	Ratio	0.76	0.35	0.68	0.49	0.61
	ACC(std)	98.64	97.32(0.34)	95.49(0.26)	94.13(2.24)	**98.90(0.00)**
Audiology	Ratio	0.81	0.35	0.75	0.79	0.69
	ACC(std)	77.98	75.93(3.44)	80.01(0.72)	64.19(6.88)	**83.02**(0.61)
Movement Libras	Ratio	0.78	0.33	0.69	0.55	0.51
	ACC(std)	65.00	62.86(0.75)	67.25(0.38)	62.25(1.02)	**69.00**(0.68)
Arrhythmia	Ratio	0.95	0.31	0.85	0.61	0.78
	ACC(std)	70.60	70.61(0.37)	73.09(0.43)	67.80(1.45)	**73.90**(0.55)
Darwin	Ratio	0.98	0.31	0.85	0.60	0.69
	ACC(std)	92.55	80.93(1.65)	96.50(1.32)	84.94(1.98)	**98.14**(0.82)
Period Changer	Ratio	≈ 1	0.60	0.90	0.95	0.95
	ACC (std)	77.78	60(1.81)	80.11(1.92)	66.44(3.13)	**83.00**(2.16)
Toxicity	Ratio	≈ 1	0.60	0.86	0.98	0.94
	ACC (std)	70.78	59.17(1.46)	69.45(1.14)	64.96(3.11)	**73.73**(1.58)

Table 5. Comparison results with other two-stage feature selection methods. Significance values are in bold. The bold values are the best results.

all four classifiers can discard a minimum of 80% of the features. In cases involving datasets with more than one thousand dimensions, such as Period Changer and Toxicity, the proposed method by four classifiers can eliminate at least 90% of the features.

Furthermore, the accuracy in Fig. 6 shows significant improvement across all eight datasets. The improvement ranges are as follows: SVM shows an enhancement of 4.12–35.09%, Ridge exhibits an increase of 0.12–33.29%, NB shows an improvement of 3.46–62.40%, and KNN demonstrates an enhancement of 7.15–24.84%.

Dataset	SVM	Ridge	NB	KNN
Glioma	0.57	0.70	0.78	0.70
Dermatology	0.64	0.52	0.58	0.73
Audiology	0.75	0.54	0.68	0.58
Movement Libras	0.60	0.57	0.57	0.59
Arrhythmia	0.83	0.80	0.92	0.73
Darwin	0.81	0.85	0.80	0.86
Period changer	0.99	0.95	0.96	0.90
Toxicity	0.97	0.91	0.92	0.94

Table 6. Ratio comparison of the proposed algorithm on multiple classifiers.

This paper employs five classifiers. Logistic regression and ridge regression are both linear models that frame classification as a convex optimization problem, making them straightforward to solve and robust. Feature selection can reduce model complexity and help prevent the classifier from overfitting the dataset. Naive bayes is a classification method grounded in Bayes' theorem and the assumption of conditional independence among features. This approach, which originates from classical mathematical theory, demonstrates good stability. However, in real-world data, a significant number of redundant or irrelevant features can adversely affect the accuracy of probability estimations, and the assumption of feature conditional independence is often invalid. Feature selection addresses this issue by retaining only those relevant features for classification, thereby reducing the likelihood of prediction errors. SVM obtains optimal classification by maximizing the minimum distance from the support vector to the hyperplane. When redundant and irrelevant features are reduced, SVM can find the hyperplane separating the categories more clearly. KNN is a distance-based algorithm. As the dimension of data features increases, the distinguishability of sample points is difficult to measure by distance. Along with the sharp reduction of feature dimensions, the effect of distance measures is strengthened, and the classification performance of KNN obtains significant improvement accordingly.

Conclusion

This paper proposes a two-stage feature selection method based on random forest and improved genetic algorithm. The first stage involves the removal of features deemed unimportant, utilizing the variable importance measure derived from the random forest. In the second stage, the method employs the improved genetic algorithm to identify the globally optimal feature subset. To validate the effectiveness of the proposed method, we evaluated feature selection ratios and accuracy metrics, conducting extensive experiments on eight UCI

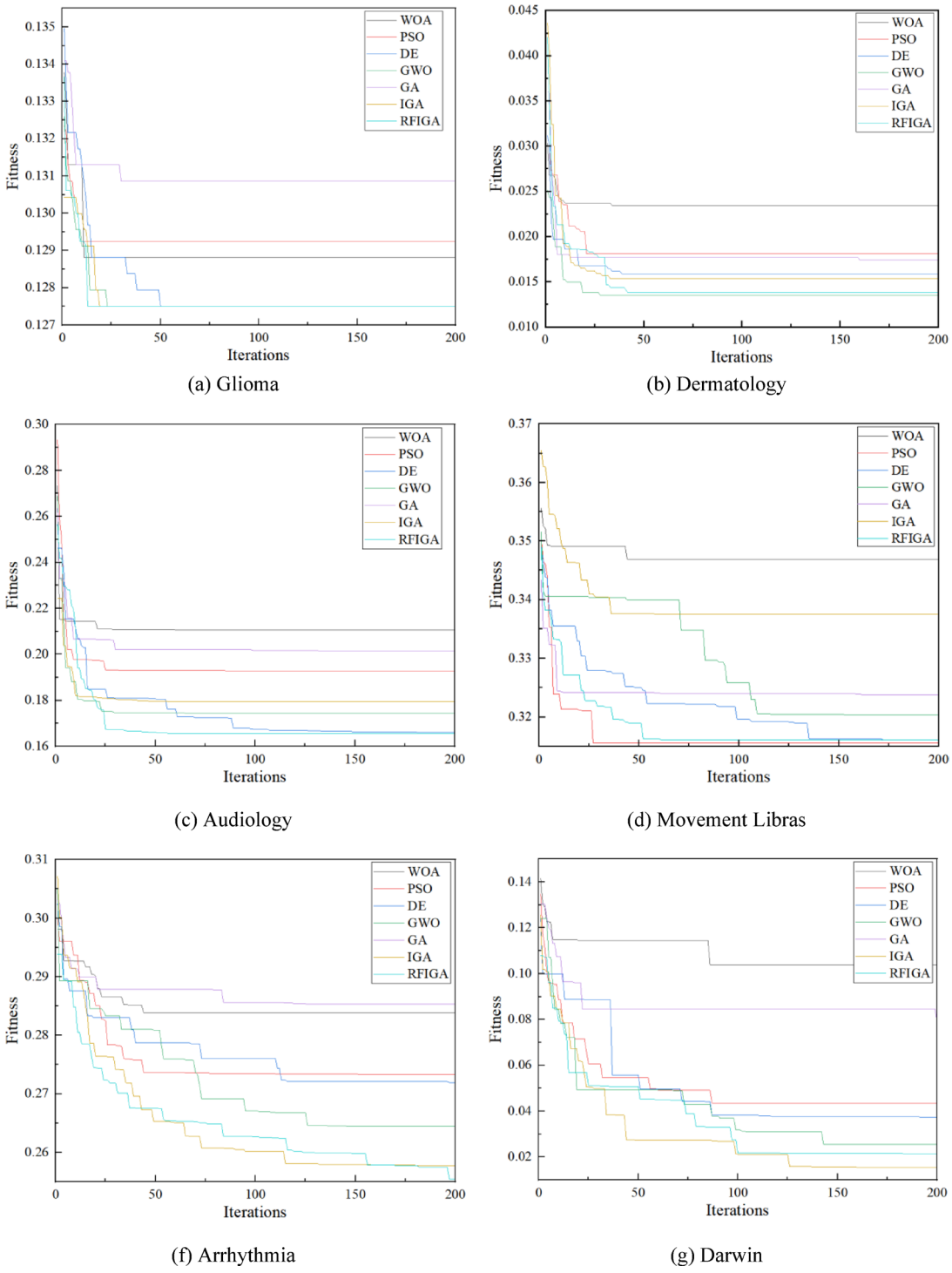
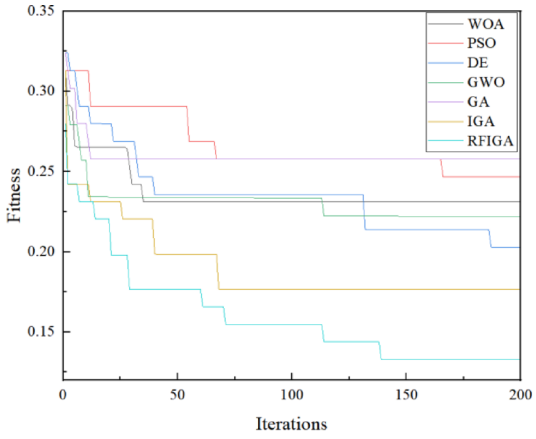


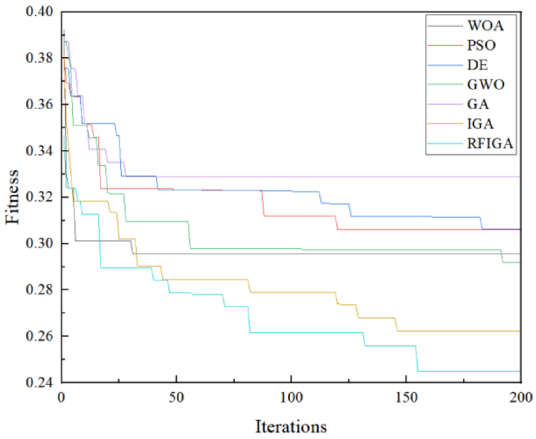
Fig. 5. Comparison of adaptation curves.

datasets. The proposed method demonstrates exceptional feature selection capabilities compared to traditional feature selection methods, swarm intelligence algorithms, and other multi-stage feature selection techniques.

However, it is important to acknowledge that the time complexity associated with the proposed method is relatively high. In scenarios involving ultra-high-dimensional data, characterized by tens of thousands of dimensions, the computational time may become prohibitive. To mitigate the challenges posed by the exponential increase in data volume, it is essential to develop a framework for feature selection that leverages GPUs. Furthermore, a more efficient method could potentially alleviate the difficulties encountered by existing algorithms in the context of feature selection for ultra-high-dimensional datasets.



(h) Period Changer



(i) Toxicity

Figure 5. (continued)

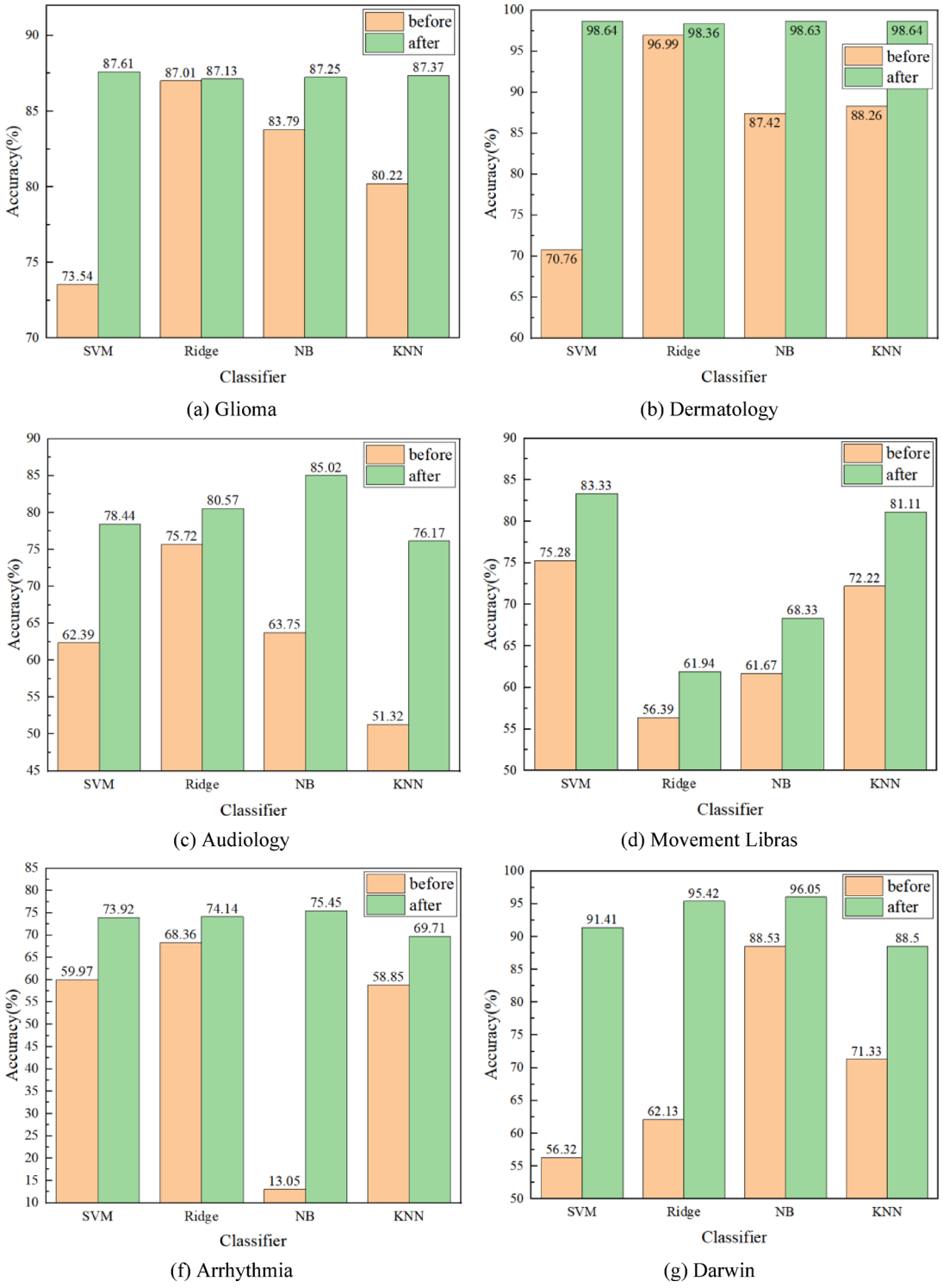
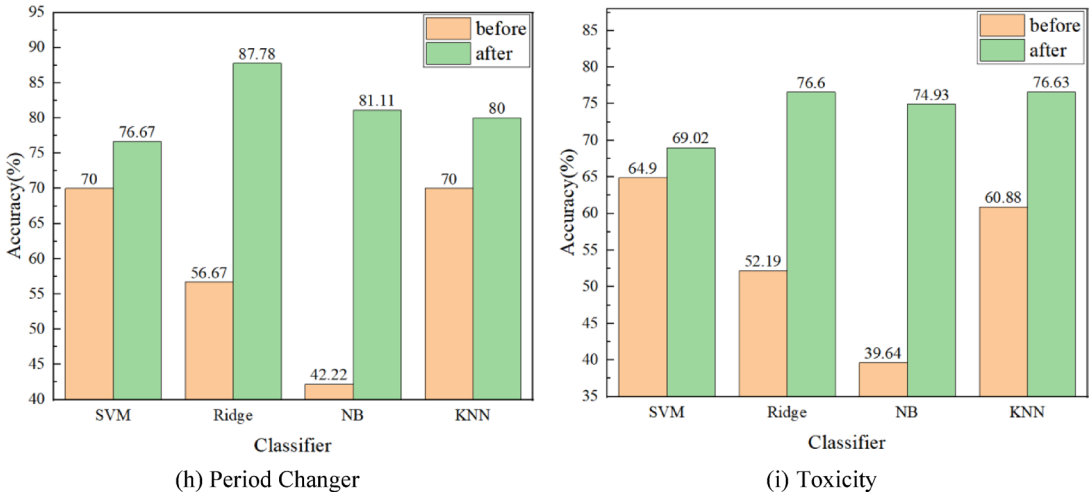


Fig. 6. Accuracy comparison of the proposed algorithm on multiple classifiers..



(h) Period Changer (i) Toxicity

Figure 6. (continued)

Data availability

The datasets generated during and/or analysed during the current study are available in the UCI dataset repository, https://archive.ics.uci.edu.

Received: 21 November 2024; Accepted: 8 May 2025
Published online: 14 May 2025

References

1. Kang, J., Ullah, Z. & Gwak, J. MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors* **21** (6), 2222 (2021).
2. Ureel, Y. et al. Active machine learning for chemical engineers: a bright future lies ahead! *Engineering*. 27, 23–30 (2023).
3. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **23** (1), 40–55 (2022).
4. Rehman, T. U., Mahmud, M. S., Chang, Y. K., Jin, J. & Shin, J. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Comput. Electron. Agric.* **156**, 585–605 (2019).
5. Zhang, L. et al. A review of machine learning in Building load prediction. *Appl. Energy*. **285**, 116452 (2021).
6. Wang, H. A novel feature selection method based on quantum support vector machine. *Phys. Scr.* **99** (5), 056006 (2024).
7. Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. When is nearest neighbor meaningful? In Database Theory—ICDT'99: In *7th International Conference Jerusalem, Israel, January 10–12, 1999. Proceedings* (eds. Beeri, C., Buneman, P.) 217–235Springer, (1999).
8. Saeys, Y., Inza, I. & Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23** (19), 2507–2517 (2007).
9. Gu, X., Guo, J., Xiao, L. & Li, C. Conditional mutual information-based feature selection algorithm for maximal relevance minimal redundancy. *Appl. Intell.* **52** (2), 1436–1447 (2022).
10. Kaya, U. & Fidan, M. Parametric and nonparametric correlation ranking based supervised feature selection methods for skin segmentation. *J. Ambient Intell. Humaniz. Comput.* **13** (2), 821–833 (2022).
11. Cui, X., Li, Y., Fan, J. & Wang, T. A novel filter feature selection algorithm based on relief. *Appl. Intell.* **52** (5), 5063–5081 (2022).
12. Yan, K. & Zhang, D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sens. Actuators B*. **212**, 353–363 (2015).
13. Houssein, E. H., Abdalkarim, N., Samee, N. A., Alabdulhafith, M. & Mohamed, E. Improved Kepler optimization algorithm for enhanced feature selection in liver disease classification. *Knowl. Based Syst.* **297**, 111960 (2024).
14. Saif Alghawli, A. & Taloba, A. I. An enhanced ant colony optimization mechanism for the classification of depressive disorders. *Comput. Intell. Neurosci.* **2022** (1), 1332664 (2022).
15. Kazerani, R. Improving breast Cancer diagnosis accuracy by particle swarm optimization feature selection. *Int. J. Comput. Intell. Syst.* **17** (1), 44 (2024).
16. Xie, Z. & Xu, Y. Sparse group LASSO based uncertain feature selection. *Int. J. Mach. Learn. Cybernet.* **5**, 201–210 (2014).
17. Wang, Y. Y. & Li, J. Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data. *Int. J. Remote Sens.* **29** (10), 2993–3010 (2008).
18. Genuer, R., Poggi, J. M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **31** (14), 2225–2236 (2010).
19. Wang, W., Li, J., Fang, Y., Zheng, Y. & You, F. An effective hybrid feature selection using entropy weight method for automatic sleep staging. *Physiol. Meas.* **44** (10), 105008 (2023).
20. Xu, J., Qu, K., Sun, Y. & Yang, J. Feature selection using self-information uncertainty measures in neighborhood information systems. *Appl. Intell.* **53** (4), 4524–4540 (2023).
21. Su, Y. et al. An efficient task implementation modeling framework with Multi-Stage feature selection and automl: A case study in forest fire risk prediction. *Remote Sens.* **16** (17), 3190 (2024).
22. Mesiar, R. & Sheikhi, A. Nonlinear random forest classification, a copula-based approach. *Appl. Sci.* **11** (15), 7140 (2021).
23. Ghosh, D. & Cabrera, J. Enriched random forest for high dimensional genomic data. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **19** (5), 2817–2828 (2021).
24. Sundhari, S. S. A knowledge discovery using decision tree by Gini coefficient. In *2011 International Conference on Business, Engineering and Industrial Applications, Kuala Lumpur, Malaysia, June 05–07, 2011. Proceedings* 232–235(IEEE, 2011). (2011), June.
25. Man, K. F., Tang, K. S. & Kwong, S. *Genetic Algorithms: Concepts and Designs* (Springer Science & Business Media, 2001).

26. Srinivas, M. & Patnaik, L. M. Adaptive probabilities of crossover and mutation in genetic algorithms. *IEEE Trans. Syst. Man. Cybernetics*. **24** (4), 656–667 (1994).
27. Mingxiao, S., Tiantian, L. & Jun, X. Amphibious Vehicle Layout Optimization based on Adaptive Elite Genetic Algorithm. In *IEEE International Conference on Mechatronics and Automation (ICMA), Tianjin, China, August 4–7, 2019. Proceedings* 491–496 (IEEE, 2019). (2019).
28. Archer, K. J. & Kimes, R. V. Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* **52** (4), 2249–2260 (2008).
29. Markelle, K. & Rachel, L. & Kolby Nottingham. The UCI Machine Learning Repository. *figshare* https://archive.ics.uci.edu
30. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45** (4), 427–437 (2009).
31. Aran, O. & Akarun, L. A multi-class classification strategy for fisher scores: application to signer independent sign Language recognition. *Pattern Recogn.* **43** (5), 1776–1788 (2010).
32. Lemmon, J. et al. Evaluation of feature selection methods for preserving machine learning performance in the presence of Temporal dataset shift in clinical medicine. *Methods Inf. Med.* **62** (01/02), 060–070 (2023).
33. Mirjalili, S. & Lewis, A. The Whale optimization algorithm. *Adv. Eng. Softw.* **95**, 51–67 (2016).
34. Wang, D., Tan, D. & Liu, L. Particle swarm optimization algorithm: an overview. *Soft. Comput.* **22** (2), 387–408 (2018).
35. Storn, R. & Price, K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11**, 341–359 (1997).
36. Mirjalili, S., Mirjalili, S. M. & Lewis, A. Grey Wolf optimizer. *Adv. Eng. Softw.* **69**, 46–61 (2014).
37. Daneshfar, F. & Aghajani, M. J. Enhanced text classification through an improved discrete laying chicken algorithm. *Expert Systems*. e13553; (2024). https://doi.org/10.1111/exsy.13553
38. Moustafa, M. S., Mahmoud, A. S. & Farg, E. Bi-stage feature selection for crop mapping using grey Wolf metaheuristic optimization. *Adv. Space Res.* **73** (10), 5005–5016 (2024).
39. Liu, J., Feng, H. & Tang, Y. A novel hybrid algorithm based on Harris Hawks for tumor feature gene selection. *PeerJ Comput. Sci.* **9**, e1229 (2023).

Acknowledgements

This research was supported by National Nature Science Foundation of China (No.62476013).

Author contributions

Junyao Ding is responsible for method design, experimentation, and manuscript. Jianchao Du is responsible for review and amendment. Hejie Wang is responsible for data analysis. Song Xiao is responsible for review and supervision.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.