Vol.2, No.4, Page 375- 380, 2024 https://doi.org/10.59628/jast.v2i4.1023 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

A hybrid Feature Selection Method Based on Binary PSO Algorithm for Microarray Data Classification

Hiba ALMarwi1 [1*] and Ghaleb H. AL-Gaphari[1]

[1] **Department of Computer Science , Faculty of Computer IT, University of Sana'a, Sana'a, Yemen**

*Corresponding author: Hebh. Almarwi@gmail. com

Abstract

Microarray technology produces data with a large number of dimensions. The abundance of dimensions in the data makes it challenging for machine learning algorithms to extract meaningful information from it. To overcome this limitation, feature selection (FS) techniques can be applied to reduce the dimensionality of the data. FS is a crucial preprocessing step that allows for the handling of high-dimensional data and facilitates more effective and efficient processing. The primary objective of this paper is to develop an efficient FS method that can effectively reduce the dimensionality of microarray data. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is utilized in the first filtering stage to extract informative features. The best feature subset is then determined by using a Particle Swarm Optimization (PSO) algorithm as a wrapper feature selector to identify ideal features. The proposed approach is evaluated using microarray datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and compared with other algorithms. The experimental findings indicate that the proposed approach outperforms the benchmark methods in terms of classification accuracy.

Article Info

Keywords:
Feature selection, Particle Swarm Optimization algorithm, Wrapper method, Dimension reduction, Multi attribute decision making

Article History:
Received: 26-June-2024,
Revised: 26-july-2024,
Accepted: 15-August-2024,
Available online: 30 August 2024.

1. INTRODUCTION

Numerous difficulties with microarray datasets make analysis and interpretation more difficult. One noticeable problem is the massive dimensionality of these datasets, where the number of genes significantly exceeds the number of samples available. In addition, the problem gets worse with redundancy and noise in the dataset. This high dimensionality can cause modeling and analytical challenges, as well as reduce the efficiency of machine learning algorithms and suffer computational costs when combined with noise and redundancy [1–3]. As such, it becomes necessary to reduce the number of features in order to handle the high-dimensionality issue [4]. One way to address the challenges caused by high-dimensional data is to employ dimensionality reduction techniques, specifically gene selection, also known as feature selection. Feature selection techniques play a crucial role in addressing these challenges by selecting the most relevant features from the high-dimensional dataset. Feature selection approaches are classified into three groups [5]: filter methods, wrapper methods, and hybrid methods. Filter methods play a fundamental role in selecting relevant features from high dimensional datasets. They operate independently of any specific learning algorithm, and they are designed to evaluate the features based on their statistical properties. Commonly employed statistical measures, such as correlation [6], information gain[7, 8], or chi-square, are utilized by filter methods to assess the relationship between each feature and the target variable. Filter methods offer several advantages, including computational efficiency and being easily understood and interpreted. However, filter methods may not capture the complex interactions between features. As a result, it is often useful to combine filter methods with additional feature selection methods, such as wrapper or embedding methods, to enhance the comprehensive selection of features. Whereas, wrapper methods incorporate a learning algorithm as part of the feature selection process [9]. Wrapper methods utilize a search algorithm, such as a genetic algorithm

[10, 11] or a recursive feature elimination [12], to explore different subsets of features and evaluate their impact on the predictive performance of the chosen learning algorithm. The literature has presented a number of hybrid approaches for addressing optimization issues. For example, in [13], the researchers developed a hybrid approach that combined Differential Evolution (DE) and the Artificial Bee Colony (ABC) algorithm. Similarly, in [14], a hybrid metaheuristic method was presented for cancer classification tasks by combining the gravitational search algorithm (GSA) with the teaching-learning-based optimization (TLBO) algorithm. In order to address global optimization issues, While Wang et al. [15] developed a hybrid technique that combines Gravitational Search Algorithm (GSA) and Particle Swarm Optimization (PSO) Algorithm. To overcome the limitations of the Simulated Annealing (SA) algorithm, Gheyas et al. in [16] presented a hybrid approach that combined SA and GA, aiming to escape from local optima and increase the overall optimization performance. These hybrid approaches provide notable improvements in optimization efficacy and efficiency by utilizing the complementary characteristics of different algorithms. Researchers in [17] propose a novel explicit representation scheme, a feature grouping strategy, and a size-adaptive expansion approach in a new ESAPSO algorithm, which demonstrates improved classification performance and computational efficiency over state-of-the-art algorithms.[18] propose a new particle swarm optimization (PSO) variant called ISPSO that integrates information gain ratio to assess feature importance and partition the feature set into groups to establish the initial population. ISPSO also replaces the traditional PSO velocity concept with a probabilistic approach and introduces a penalty term to enhance the algorithm's ability to achieve superior feature selection results. The majority of the hybrid feature selection techniques available in the literature employ a combination of one filter method and a wrapper method. Each filter method typically focuses on a specific characteristic of the dataset. As a result, it would be advantageous to combine many filter methods with a wrapper method.

To the best of our knowledge, the efficacy of the Particle Swarm Optimization (PSO) algorithm in solving feature selection problems as a hybrid feature selection method has not been extensively studied in the existing literature. Therefore, the primary aim of this research paper is to develop a novel hybrid feature selection approach by integrating four distinct filter-based methods with a PSO algorithm. The proposed methodology uses the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to identify informative features within the gene dataset. Subsequently, Particle Swarm Optimization (PSO) is employed as a wrapper strategy to find an optimal feature subset. The structure of the paper is organized as follows: Section 2 provides an overview of the essential concepts and techniques employed in this paper. Section 3 presents the proposed hybrid feature selection method in detail. Subsequently, Section 4 discusses the experimental setup and presents the results obtained from the conducted experiments. Finally, in Section 5, the work is eventually concluded with a summary of the findings and recommendations for further research directions.

2. PRELIMINARIES

2.1. Multi-attribute decision making (MADM)

Multi-Attribute Decision-Making (MADM) techniques have emerged as popular tools for addressing decision-making scenarios that need the examination of different alternatives. Among these techniques, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is a well-known and commonly used methodology in the related literature. The TOPSIS method provides a quantitative approach to assessing feature significance within a dataset. TOPSIS accomplishes this by measuring each feature's performance against two hypothetical scenarios: an ideal solution and a negative-ideal solution (worst-case scenario). This evaluation process is mathematically formulated in Equations 1 and 2. The ideal solution represents the characteristics that are most desirable for a feature, while the worst-case scenario represents the least desirable characteristics.

$$S_{i-} = \left[\sum_{j=1}^{n}(v_{ij} - v_{-j})^2\right]^{0.5} \tag{1}$$

$$S_{i+} = \left[\sum_{j=1}^{n}(v_{ij} - v_{+j})^2\right]^{0.5} \tag{2}$$

The relative closeness to the ideal point can be calculated by Eq 3.

$$C_{i+} = \frac{s_{i-}}{s_{i+} + s_{i-}} \tag{3}$$

2.2. Particle Swarm Optimization (PSO) algorithm

Particle Swarm Optimization (PSO) is a population-based metaheuristic optimization technique. In the PSO algorithm, each solution is referred to as a "particle." These particles form a swarm, usually represented by N, and navigate a D-dimensional search space. The algorithm starts by randomly initializing a population of particles in the search space. Each particle represents one potential solution. Each particle maintains information about their current position and velocity. Additionally, PSO incorporates memory mechanisms: each particle tracks its personal best position (pbest), which represents the best previously visited position of the ith par-

ticle. Furthermore, the swarm maintains a global best position (gbest), which is the best solution found by any particle during the current iteration. During each iteration, particles update their velocities and positions based on their individual experiences (pbest) and knowledge of the global optimum (gbest), as formulated in Equation 4,5.

$$V[] = v[] + c1 * rand() * (pbest[] - present[]) + c2 * rand() * (gbest[] - present[]) \quad (4)$$

$$Present[] = present[] + v[] \quad (5)$$

Where:

$V[]$: is the particle velocity.

$Persent[]$: is the current particle (solution).

$Pbest[]$: is the best value for which particle has reached this moment.

$gbest[]$: It represents the best value within the swarm.

$rand()$: is a random number between (0, 1).

$c1$, $c2$ are learning factors usually $c1 = c2 = 2$.

As the iterations progress, particles move through the search space, gradually converging towards promising regions that potentially contain optimal solutions. The algorithm continues until a termination criterion is met, such as reaching a maximum number of iterations or achieving satisfactory solution quality.

Algorithm 1:	pseudo code of Proposed Model
1	Load dataset unknown threats
2	Apply preprocessing data (transformation, standardization)
3	Use (TOPSIS) as a filter method for extracting informative features
4	Select the highly informative subset of genes from TOPSIS result as population for pso algorithm
5	Initialize the population p, maxiter, position and velocity
6	For each particle
7	If the fitness value is better than its personal best
8	If the fitness value is better than its personal best
9	set current value as the new pBest
10	End
11	Choose the particle with the best fitness value of all as gBest
12	For each particle
13	Calculate particle velocity according equation (4)
14	Update particle position according equation (5)
15	End
16	While maximum iterations or minimum error criteria is not attained
17	Return gBest as optimal features.

3. METHODOLOGY

The proposed methodology's architecture is illustrated in Figure 1. Additionally, the pseudo-code has been outlined in algorithm 1. The workflow begins with the use of the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) as a filtering mechanism. TOPSIS identifies a subset of informative genes within the dataset. Subsequently, the Particle Swarm Optimization (PSO) algorithm refines this initial selection. PSO takes the gene subset obtained from TOPSIS as input and iteratively searches for the optimal subset of features. A detailed explanation of this two-stage process is provided in the following subsections.

3.1. Pre-processing of genes using TOPSIS

Four popular filter methods have been used to extract a subset of highly informative features from a given dataset. These methods, namely ReliefF, correlation-based, ANOVA, and information gain, collectively explore various aspects of the dataset. Each filter method is applied to the dataset, with equal weight. Based on the experimental results, the optimal weight distribution was determined to be 0.3 for each of the four filter methods. Consequently, the weight values were set as follows: w1 = 0.3, w2 = 0.3, w3 = 0.3, and w4 = 0.3. This equal weighting scheme was adopted to ensure that all four filter methods contributed equally to the overall feature selection process[19]. Consequently, based on the results of the filter methods, rankings are produced for each feature in the dataset. These rankings serve as alternative inputs for the subsequent TOPSIS technique, which is used to get each feature's final ranking. The resulting ranking is then utilized as the input for the subsequent wrapper-based feature selection technique. At this stage, a matrix is formed, comprising the rankings of each feature according to the four filter method. This matrix is treated as the decision matrix for the TOPSIS method, as visually represented in Figure 1. To identify a subset of genes that demonstrate high informativeness, the top 50% of ranked genes are partitioned into distinct segments with varying lengths. Subsequently, a neural network is employed as the classification model to evaluate the accuracy of each segment. The resulting classification accuracies for each segment are presented in Table 1. The segment exhibiting the highest accuracy is selected as the top segment, which is subsequently utilized as the input for the subsequent wrapper-based feature selection technique paper. Finally, the best particle is returned as the final solution.

3.2. Feature Selection using Particle Search algorithm

At this stage of the analysis, the top-ranked features are further reduced to select an optimal subset of features. To accomplish this, the Particle Swarm Optimization (PSO) algorithm is employed. The results obtained from the TOPSIS analysis serve as the initial population for the PSO algorithm. In the second step, particles are randomly initialized using discrete representations, where each particle can assume values of either 0 or 1. The value 1 means the selection of the desired feature,

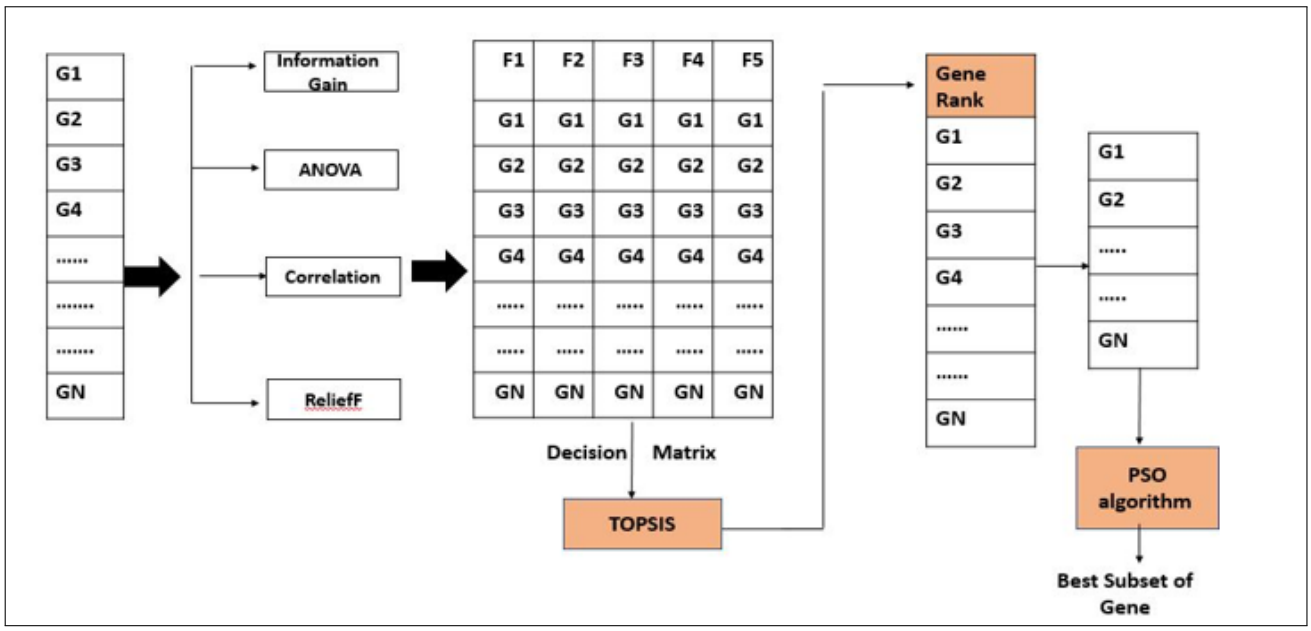

Figure 1. Overall architecture for the proposed approach

and 0 means not selecting that feature. Furthermore, the position and velocity of each particle are randomly initialized. Moving forward to the third step, the initialized particles are evaluated based on objective functions. The objective function is formulated to enhance classification accuracy. A neural network is employed as the classification model to calculate the classification accuracy of each solution at each iteration of the PSO algorithm. The objective functions are defined according to Equation 6, as provided in the paper. Finally, the best particle is returned as the final solution.

4. DESIGN OF EXPERIMENTS

In the experiments, the proposed method's performance was evaluated using genetic algorithms (GA) and ant colony optimization (ACO) as benchmark techniques. The experiments involved the utilization of the ADNI dataset, which consists of publicly available high-dimensional microarray datasets and can be accessed at the following URL:https://ida.loni.usc.edu/login.jsp#74. A summary of the datasets employed in the experiments is provided in Table 2. Furthermore, comparative results achieved by the proposed feature selection method alongside those obtained from alternative approaches is provided in Table 3. The proposed method was implemented using Python 3.7, and the simulations were executed on a computer system equipped with an Intel Core i7 CPU and 8 GB of RAM. The effectiveness of the proposed feature selection method was assessed through the utilization of the neural network functionality available in the TensorFlow library. TensorFlow is an open-source platform for building and deploying machine learning models, particularly large-scale neural networks. It was developed by Google's Brain team. The

Table 1. The classification Accuracy of each segment

Segment No.	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Acc	95.23	95.45	96.23	87.83	81.81

Table 2. Dataset Description

Number of SNPs(Features)Number of SNPs(Features)	620,901
Number of Subjects(Sample) 818 patients	818 patients
Patients Diagnosis(Number of Classes)	NC, AD
Dataset size	26 Gigabyte

performance of the proposed model is evaluated using Equation 6.

$$f_1(x) = Erorr\ \ rate = \frac{FP + FN}{FP + FN + TP + TN} \quad (6)$$

The classification accuracy of each segment recorded in Table 1 as below:

As indicated in Table 1, Segment 3 demonstrated the highest accuracy rate among all evaluated subsets. Consequently, it was determined to be the optimal subset for the subsequent analysis. A detailed description of the dataset employed in this study is presented in Table 2.

A comparative analysis of the proposed model and alternative algorithms is presented in Table 3. As illustrated in Table 3 and visually presented in Figure 2-4, the proposed model exhibits a significant improvement in classification accuracy compared to the benchmark algorithm. The model achieved an accuracy rate of 94.58% compared to 84.68%for GA algorithm and 83% for ACO. The superior performance of the proposed model can be attributed to two primary factors. Firstly, the TOPSIS

Table 3. Comparison of our approach with other algorithms

Metaheuristic approaches	classification accuracy (CA)
Proposed approach	94.58%
GA	84.68%
ACO	83%
All	65%

method employed in the model effectively explores diverse dataset characteristics, enabling the extraction of a highly informative feature subset. Secondly, the inherent capability of the PSO algorithm to skip from local optima contributes significantly to the model's enhanced performance. These findings strongly suggest the superiority of the proposed model for the feature selection.

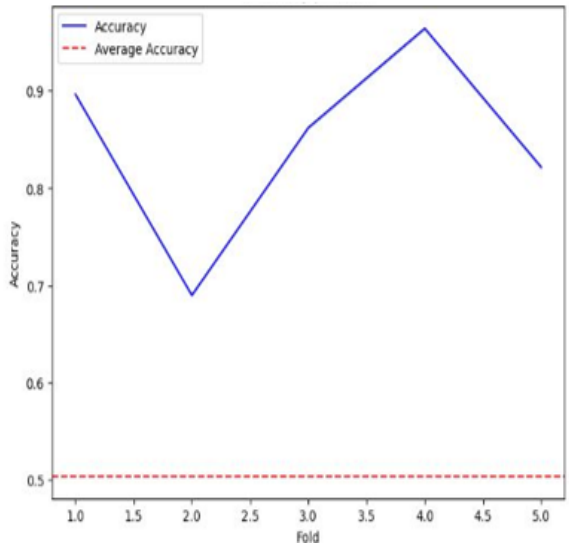

Figure 2. Result of Proposed approach.

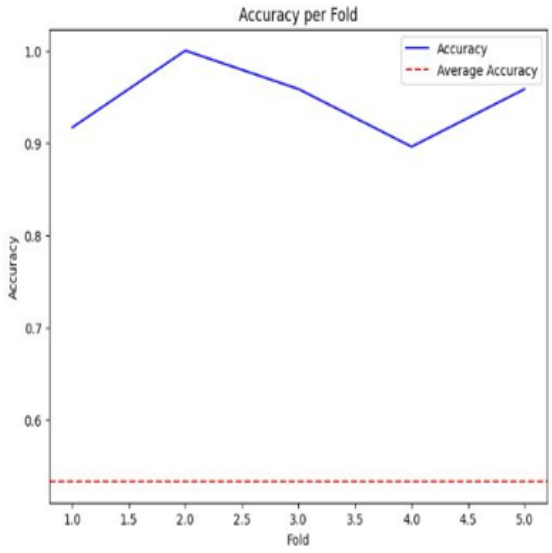

Figure 3. Result of Genetic Algorithm

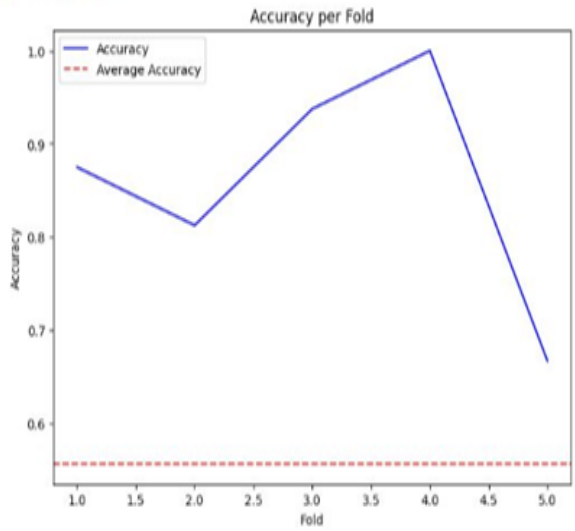

Figure 4. Result of Ant Colony algorithm

5. DISCUSSION AND RESULTS

Table 3 presents the comparative results achieved by the proposed feature selection method alongside those obtained from alternative approaches. As shown in the table, the proposed method outperforms other algorithms in terms of classification accuracy. It's possible that the proposed method's superiority comes from the fact that it employs four distinct and trustworthy filter methods as a preprocessing step, in contrast to other algorithms that only use one method for dataset preparation. By using a variety of filter techniques to enhance the quality of the chosen features before the wrapper method uses them, the proposed method effectively enhances the quality of the selected features before they are subjected to the wrapper method, thereby leading to improved classification accuracy.

6. CONCLUSION

This paper presents a new hybrid feature selection method that employs a wrapper technique in addition to four other filters. Results from this research show that the overall ranking that results from combining these four filters is better than the results from using each filter separately. PSO algorithm is proposed as a wrapper feature selector to identify ideal features. The ADNI benchmark microarray dataset was used in the testing and comparison of the proposed method with another popular algorithm. The experimental findings demonstrate that the proposed feature selection method achieves superior classification accuracy compared to existing benchmark algorithms. This performance improvement suggests that the proposed method is a well-suited and potentially efficient technique for feature selection in the context of microarray data analysis.

REFERENCES

[1] J. H. Phan, C. F. Quo, and M. D. Wang, "Cardiovascular genomics: a biomarker identification pipeline," IEEE Trans. on Inf. Technol. Biomed. **16**, 809–822 (2012).

[2] C. C. Chen, H. Schwender, J. Keith, *et al.*, "Methods for identifying snp interactions: a review on variations of logic regression, random forest and bayesian logistic regression," IEEE/ACM transactions on computational biology bioinformatics **8**, 1580–1591 (2011).

[3] K. Kourou, T. P. Exarchos, K. P. Exarchos, *et al.*, "Machine learning applications in cancer prognosis and prediction," Comput. structural biotechnology journal **13**, 8–17 (2015).

[4] A. Ben-Dor, L. Bruhn, N. Friedman, *et al.*, "Tissue classification with gene expression profiles," in *Proceedings of the fourth annual international conference on Computational molecular biology,* (2000), pp. 54–64.

[5] P. Langley, *Elements of machine learning* (Morgan Kaufmann, 1996).

[6] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," nature **415**, 530–536 (2002).

[7] L.-Y. Chuang, C.-H. Ke, H.-W. Chang, and C.-H. Yang, "A two-stage feature selection method for gene expression data," OMICS A journal Integr. Biol. **13**, 127–137 (2009).

[8] X. Yan, M. Deng, W. K. Fung, and M. Qian, "Detecting differentially expressed genes by relative entropy," J. theoretical biology **234**, 395–402 (2005).

[9] X. Huang, L. Zhang, B. Wang, *et al.*, "Feature clustering based support vector machine recursive feature elimination for gene selection," Appl. Intell. **48**, 594–607 (2018).

[10] C. M. Macal, "Agent-based modeling and artificial life," Complex Soc. Behav. Syst. Game Theory Agent-Based Model. pp. 725–745 (2020).

[11] M. Zhu and L. Wang, "Intelligent trading using support vector regression and multilayer perceptrons optimized with genetic algorithms," in *The 2010 international joint conference on neural networks (IJCNN),* (IEEE, 2010), pp. 1–5.

[12] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," Sensors Actuators B: Chem. **212**, 353–363 (2015).

[13] E. Zorarpacı and S. A. Özel, "A hybrid approach of differential evolution and artificial bee colony for feature selection," Expert Syst. with Appl. **62**, 91–103 (2016).

[14] K.-A. Lê Cao, A. Bonnet, and S. Gadat, "Multiclass classification and gene selection with a stochastic algorithm," Comput. Stat. & Data Anal. **53**, 3601–3615 (2009).

[15] S. Mirjalili, G.-G. Wang, and L. d. S. Coelho, "Binary optimization using hybrid particle swarm optimization and gravitational search algorithm," Neural Comput. Appl. **25**, 1423–1435 (2014).

[16] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," Pattern recognition **43**, 5–13 (2010).

[17] L. Qu, W. He, J. Li, *et al.*, "Explicit and size-adaptive pso-based feature selection for classification," Swarm Evol. Comput. **77**, 101249 (2023).

[18] J. Gao, Z. Wang, T. Jin, *et al.*, "Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection," Knowledge-Based Syst. **286**, 111380 (2024).

[19] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing **300**, 70–79 (2018).