



Multigranulation fusion meets evolutionary search: A dual-stage LGPSO–GWO framework for high-dimensional knowledge acquisition

Qinglong Feng , Weihua Xu*

College of Artificial Intelligence, Southwest University, Chongqing, 400715, People's Republic of China

HIGHLIGHTS

- We developed a two-stage hierarchical graph PSO–GWO hybrid optimization model.
- Weight-driven probability initialization balances feature importance.
- Graph theory restructures the information transfer rule.
- Hierarchical graph architecture enhances cooperation between PSO and GWO.
- Experiments on 12 public datasets validate the performance of our optimizer.

ARTICLE INFO

Keywords:

Feature selection
Generalized multigranulation neighborhood rough set
Layered graph particle swarm optimization
Grey wolf optimizer
High-dimensional data

ABSTRACT

High-dimensional feature selection requires reducing the search space without discarding informative feature interactions. However, filter methods may provide insufficient subset-level evaluation, while wrapper-based evolutionary methods usually incur substantial computational cost when directly searching the original feature space. To address these issues, this paper proposes a dual-stage feature selection framework, termed DSLGPSO–GWO, which integrates generalized multigranulation neighborhood rough sets (GMNRS) with layered graph particle swarm optimization and grey wolf optimization. In Stage I, adaptive feature granularities are constructed by K-means with silhouette-based granularity determination. Their decision-supporting capabilities are measured using GMNRS-based external importance, and highly ranked granularities are proportionally retained to form a reduced candidate feature space. In Stage II, a granularity-feature hierarchical prior is constructed to transfer the screening information to evolutionary search, and an adaptive multiscale initialization strategy is used to improve the quality and diversity of the initial population. A layered graph learning mechanism then exploits both fitness and structural differences among feature subsets, while upper-layer guidance and three-leader GWO learning provide complementary search directions. Diversity-oriented perturbation is further introduced to alleviate population stagnation. Experiments on twelve benchmark datasets with KNN, SVM-RBF, and MLP classifiers show that DSLGPSO–GWO achieves the best or tied-best accuracy in 24 of 36 dataset–classifier combinations. Ablation, parameter sensitivity, convergence, and non-parametric statistical analyses further demonstrate the effectiveness, robustness, and stability of the proposed framework.

1. Introduction

With the rapid advancement of data acquisition and storage technologies, high-dimensional datasets have become increasingly prevalent in machine learning, pattern recognition, bioinformatics, and intelligent decision-making. Such datasets frequently contain a large number of redundant, irrelevant, or weakly informative features, which increase computational burden and may impair the generalization performance

of learning models. Feature selection aims to identify a compact subset of informative and discriminative features while preserving essential discriminative information, and has therefore become an important technique for high-dimensional data analysis. Existing feature selection methods are generally categorized into filter, wrapper, and embedded approaches. Filter methods evaluate features independently of a specific learning model and are therefore computationally efficient, whereas

* Corresponding author.

Email addresses: fqlswu@163.com (Q. Feng), chxuwh@gmail.com (W. Xu).

<https://doi.org/10.1016/j.asoc.2026.116525>

Received 25 June 2026; Received in revised form 18 August 2026; Accepted 22 September 2026

Available online 30 September 2026

1568-4946/© 2026 Published by Elsevier B.V.

wrapper methods assess candidate feature subsets according to predictive performance and can better account for interactions among features. However, performing wrapper search directly in the original high-dimensional feature space usually entails repeated classifier evaluations and consequently incurs substantial computational cost [1,2]. Reducing the search dimensionality while retaining informative feature interactions therefore remains a fundamental challenge in high-dimensional feature selection.

Rough set theory, originally proposed by Pawlak [3], provides a knowledge-driven framework for characterizing uncertainty, dependency, and decision information directly from data, making it well suited to feature evaluation and reduction. To accommodate numerical data, Hu et al. [4] introduced a neighborhood rough-set model for heterogeneous feature subset selection, establishing an important foundation for neighborhood-based feature evaluation. From a multi-granulation perspective, Qian et al. [5] proposed the multigranulation rough set model, in which target concepts are jointly approximated through multiple granulation structures. Building on this framework, Sun et al. [6] incorporated fuzzy neighborhood entropy-based uncertainty measures into fuzzy neighborhood multigranulation rough sets for feature selection. Xu et al. [7] subsequently developed a local generalized multigranulation neighborhood rough-set model and investigated dynamic approximation updating. Zhang et al. [8] constructed composite entropy-based uncertainty measures under generalized multigranulation fuzzy neighborhood rough sets, while Xu and Bu [9] introduced a weight-partitioned multigranulation rough-set model together with a matrix-based incremental feature selection method to exploit information from different granular structures more effectively. More recently, Wang et al. [10] proposed an adaptive fuzzy rough-set fusion model with class variance to improve feature evaluation under complex data distributions. Collectively, these studies have extended dependency-aware feature evaluation through neighborhood modeling, multigranulation mechanisms, uncertainty characterization, and adaptive information fusion. Nevertheless, existing multigranulation rough-set research has concentrated primarily on the evaluation and fusion of multiple granular structures, whereas the adaptive construction of informative feature granularities still lacks systematic investigation.

Particle swarm optimization (PSO) has also been extensively applied to feature selection because of its simple implementation, limited number of control parameters, efficient information sharing, and strong global search capability. When applied to high-dimensional feature spaces, however, conventional PSO may suffer from low-quality population initialization, inefficient particle representation, insufficient learning diversity, and premature convergence. Recent research has therefore sought improvements in population initialization, particle representation and updating, population organization, learning mechanisms, and adaptive search control. Gao et al. [11] proposed an information-gain-ratio-based subfeature grouping PSO that organizes features according to their relevance and constructs more informative initial solutions. Xue and Zhang [12] developed an importance-guided PSO based on a multilayer perceptron, incorporating learned feature importance into both population initialization and evolutionary search. To reduce the computational burden associated with conventional particle representations, Qu et al. [13] proposed an explicit and size-adaptive PSO that dynamically adjusts the representation of candidate feature subsets. He et al. [14] introduced an oscillatory update mechanism together with mutual-information-based feature evaluation to enhance exploration for hybrid data. Taken together, these studies indicate that informative initialization and adaptive particle representation play important roles in improving search efficiency in high-dimensional feature spaces.

Beyond initialization and representation, population organization and learning strategies constitute another important direction for improving PSO. Jin et al. [15] proposed an adaptive pyramid PSO that hierarchically organizes particles and employs different learning paradigms

to enhance population diversity and search capability. Yang et al. [16] developed a leader-adaptive PSO with dimensionality reduction, allowing the leadership mechanism to vary with the search state and thereby reducing excessive dependence on fixed leaders. Wei et al. [17] combined feature-informed initialization with progressive search-space reduction in a comprehensive scoring framework for high-dimensional feature selection. Wei et al. [18] further introduced reinforcement learning and adaptive scoring to adjust search behavior according to the evolutionary state. More recently, Zhang et al. [19] proposed an adaptive diversity-driven multi-strategy PSO that jointly exploits population diversity and multiple search strategies. This body of work underscores the importance of hierarchical population organization, diversified learning sources, and adaptive search control. Nevertheless, exemplar selection in many existing learning strategies remains largely driven by solution quality, while structural differences among feature subsets have not been sufficiently incorporated into the selection of learning exemplars.

Hybridization with complementary evolutionary or swarm-intelligence mechanisms provides another effective means of improving PSO. Different optimizers exhibit distinct characteristics in global exploration, local exploitation, diversity preservation, and leader guidance, and their integration can therefore compensate for the limitations of a single search mechanism. Thaher et al. [20] incorporated evolutionary population dynamics into Boolean PSO to improve population renewal and diversity during feature selection. Osei-Kwakye et al. [21] proposed a hybrid PSO and crow search algorithm with clustering-based initialization to improve both the initial population distribution and exploratory capability. Zhou et al. [22] integrated genetic search mechanisms into PSO for feature selection and parameter optimization. Dhal et al. [23] combined PSO with grey wolf optimization to exploit the complementary advantages of particle-based search and multi-leader guidance. Although these hybrid strategies can improve the balance between exploration and exploitation, their computational burden can remain substantial when wrapper optimization is performed directly in the original high-dimensional feature space, where candidate feature subsets must be repeatedly evaluated by a classifier.

The complementary strengths of rough-set models and population-based optimizers have further motivated their integration in feature selection. Rough-set models provide data-driven knowledge of uncertainty and dependency, whereas population-based optimizers offer strong combinatorial search capabilities. Accordingly, rough-set information has been incorporated into evolutionary optimization for feature screening, population initialization, and search guidance. Sun et al. [24] proposed a two-stage fuzzy neighborhood-based feature reduction method with binary whale optimization, in which fuzzy neighborhood information is used to reduce the feature space before subsequent optimization. Sun et al. [25] further proposed TFSFB by combining fuzzy multi-neighborhood rough sets with binary whale optimization for imbalanced data. Guo et al. [26] employed hybrid fuzzy rough-set information to construct informative initial populations for metaheuristic optimization on high-dimensional gene data. For hybrid data, Li et al. [27] proposed a cooperative swarm-intelligence method based on fuzzy β -covering and fuzzy self-information to guide population search. Chen et al. [28] incorporated fuzzy β -covering and fuzzy mutual information into ant-colony-based attribute reduction. Zhang and Zhao [29] combined a k -nearest-neighborhood rough-set model with opposition-based PSO for heterogeneous information systems. More recently, Xie et al. [30] incorporated fuzzy neighborhood information into differential evolution followed by decision-tree refinement. Collectively, these studies establish rough-set knowledge as an effective source of guidance for feature screening, population initialization, and evolutionary search. However, an explicit cross-stage mechanism for transferring granularity-level knowledge acquired during preliminary screening to subsequent feature-level evolutionary optimization has yet to be systematically investigated.

On this basis, three research issues merit further investigation:

- (1) Many PSO-based and hybrid evolutionary feature selection methods still perform wrapper optimization directly in the original high-dimensional feature space. Repeated classifier evaluations for each candidate subset cause the computational cost to increase substantially with feature dimensionality and the number of fitness evaluations.
- (2) Many existing hierarchical and adaptive learning strategies determine exemplars primarily according to solution quality. Structural differences among feature subsets are not always explicitly incorporated into exemplar selection, which may restrict the diversity of learning directions and increase the risk of premature convergence.
- (3) Although rough-set information has been employed for feature screening, population initialization, and evolutionary search guidance, existing studies provide limited support for the adaptive construction of feature-level granular structures and the explicit cross-stage transfer of multigranulation knowledge to subsequent evolutionary optimization.

To address these issues, this paper proposes a dual-stage feature selection framework termed DSLGPSO–GWO. The main contributions are summarized as follows:

- (1) A dual-stage filter–wrapper feature selection framework is developed to reduce the search burden in high-dimensional subset optimization. In Stage I, adaptive feature granularities are constructed and ranked according to their GMNRS-based external importance with respect to the decision attribute. Highly ranked granularities are then proportionally retained to construct a compact candidate feature space for subsequent optimization.
- (2) A hierarchical granularity–feature prior is formulated to transfer the granularity-level information obtained in Stage I to feature-level optimization in Stage II. Based on this prior, an adaptive multiscale initialization strategy is developed to generate candidate subsets of different sizes while balancing prior-guided exploitation and uniform exploration, thereby improving both the informativeness and structural diversity of the initial population.
- (3) An improved LGPSO–GWO optimizer is developed for feature subset search. Same-layer graph learning jointly incorporates fitness advantage and structural differences among feature subsets, while upper-layer guidance and the three-leader GWO mechanism provide complementary learning directions. A diversity-oriented perturbation mechanism is further incorporated to mitigate population stagnation and reduce the risk of premature convergence.

The remainder of this paper is organized as follows. [Section 2](#) introduces the theoretical foundations related to the proposed method. [Section 3](#) presents the proposed DSLGPSO–GWO framework and its main components. [Section 4](#) describes the experimental setup and analyzes the experimental results. Finally, [Section 5](#) concludes the paper and outlines future research directions.

2. Related work and foundations

This section introduces the theoretical foundations closely related to the proposed method. First, the generalized multigranulation neighborhood rough set is reviewed, including the neighborhood relation, support function, approximation operators, and dependency degree. Subsequently, the basic mechanisms of particle swarm optimization and grey wolf optimization are presented. These foundations provide the theoretical basis for the first-stage granularity screening and the second-stage evolutionary feature search developed in the following sections.

2.1. Generalized multigranulation neighborhood rough set

Classical rough set theory characterizes indiscernibility using equivalence relations and is therefore naturally suited to discrete information systems [3]. To process numerical data without discretization, neighborhood rough sets replace equivalence classes with metric-based neighborhoods [4]. Multigranulation rough sets further describe a target concept through multiple granulation structures rather than a single relation [5]. However, the optimistic and pessimistic models respectively require at least one or all granularities to support an object, which may produce excessively relaxed or strict approximations. The generalized multigranulation neighborhood rough set (GMNRS) introduces a controllable support proportion to provide an intermediate and more flexible decision rule [7].

2.1.1. Neighborhood decision system and neighborhood relation

Definition 1. Let $IS = (U, A \cup D, V, f)$ be a neighborhood decision system, where $U = \{x_1, x_2, \dots, x_n\}$ is a nonempty finite sample set, $A = \{a_1, a_2, \dots, a_m\}$ is the conditional attribute set, and $D = \{d\}$ is the decision attribute. The decision attribute induces the partition $U/D = \{D_1, D_2, \dots, D_C\}$, where D_c denotes the set of samples belonging to the c -th decision class.

For an attribute subset $B \subseteq A$, the Euclidean distance between samples x_p and x_q is defined as

$$d_B(x_p, x_q) = \sqrt{\sum_{a_j \in B} (\tilde{a}_j(x_p) - \tilde{a}_j(x_q))^2}, \quad (1)$$

where $\tilde{a}_j(x)$ denotes the standardized value of attribute a_j for sample x . Given a neighborhood radius $\delta > 0$, the neighborhood of x_p induced by B is

$$N_B^\delta(x_p) = \{x_q \in U \mid d_B(x_p, x_q) \leq \delta\}. \quad (2)$$

In the proposed framework, attributes are partitioned into a family of feature granularities

$$\mathcal{G} = \{G_1, G_2, \dots, G_K\},$$

where

$$G_r \cap G_s = \emptyset \quad (r \neq s), \quad \bigcup_{r=1}^K G_r = A.$$

Because different granularities may contain different numbers of features, the neighborhood distance induced by G_r is normalized as

$$d_{G_r}(x_p, x_q) = \sqrt{\frac{1}{|G_r|} \sum_{a_j \in G_r} (\tilde{a}_j(x_p) - \tilde{a}_j(x_q))^2}. \quad (3)$$

This normalization prevents large granularities from producing systematically larger distances merely because they contain more features.

2.1.2. Support function for multigranularity evaluation

GMNRS characterizes a target concept from multiple granular perspectives. Let $\mathcal{G} = \{G_1, G_2, \dots, G_K\}$ be a family of attribute granularities, where $G_r \subseteq A$ induces a neighborhood relation R_{G_r} . Here, K denotes the number of granularities and is distinguished from m , the total number of conditional attributes.

Definition 2. For a target concept $X \subseteq U$, the support function of granularity G_r with respect to an object $x \in U$ is defined as

$$P_X^{R_{G_r}}(x) = \begin{cases} 1, & N_{G_r}^\delta(x) \subseteq X, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where $N_{G_r}^\delta(x)$ denotes the neighborhood of x induced by granularity G_r under the neighborhood radius δ .

If $P_X^{R_{G_r}}(x) = 1$, the neighborhood of x under G_r is completely contained in X , indicating that granularity G_r can certainly support the classification of x into the target concept X . Otherwise, $P_X^{R_{G_r}}(x) = 0$, which means that the corresponding neighborhood contains at least one object that does not belong to X .

In the optimistic multigranulation rough set, an object can be included in the lower approximation as long as at least one granularity supports its membership in X . By contrast, the pessimistic multigranulation rough set requires all granularities to provide consistent support. The former condition may be overly relaxed, whereas the latter may be excessively strict. GMNRS introduces a flexible proportion parameter to provide an intermediate approximation mechanism between these two extreme strategies.

2.1.3. Generalized multigranulation approximation operators

To control the required degree of agreement among different granularities, a generalized multigranulation parameter $\beta_R \in (0.5, 1]$ is introduced. The subscript R is used to distinguish this rough-set parameter from the β leader in the grey wolf optimizer.

Definition 3. Given a target concept $X \subseteq U$ and the granularity family $\mathcal{G} = \{G_1, G_2, \dots, G_K\}$, the generalized multigranulation lower approximation and upper approximation of X are respectively defined as

$$\underline{P}_{\beta_R}^{\mathcal{G}}(X) = \left\{ x \in U \mid \frac{1}{K} \sum_{r=1}^K P_X^{R_{G_r}}(x) \geq \beta_R \right\}, \quad (5)$$

and

$$\overline{P}_{\beta_R}^{\mathcal{G}}(X) = \left\{ x \in U \mid \frac{1}{K} \sum_{r=1}^K P_X^{R_{G_r}}(x) > 1 - \beta_R \right\}. \quad (6)$$

Accordingly, the boundary region is defined as

$$Bn_{\beta_R}^{\mathcal{G}}(X) = \overline{P}_{\beta_R}^{\mathcal{G}}(X) \setminus \underline{P}_{\beta_R}^{\mathcal{G}}(X). \quad (7)$$

If $\underline{P}_{\beta_R}^{\mathcal{G}}(X) = \overline{P}_{\beta_R}^{\mathcal{G}}(X)$, then X is definable with respect to $\mathcal{G}$; otherwise, X is a rough set with a nonempty uncertainty region.

The lower approximation $\underline{P}_{\beta_R}^{\mathcal{G}}(X)$ contains the objects for which the proportion of supporting granularities is no less than β_R , and therefore represents the certain region of X . The upper approximation $\overline{P}_{\beta_R}^{\mathcal{G}}(X)$ contains the objects whose support proportion is greater than $1 - \beta_R$, and therefore represents the possible region of X .

The parameter β_R controls the strictness of the approximation. A larger value of β_R imposes a stricter requirement on the lower approximation, since an object must be supported by a larger proportion of granularities to enter the certain region. In contrast, a smaller value provides a more tolerant multigranulation decision rule.

2.1.4. Attribute dependency in GMNRS

Definition 4. Let the decision attribute induce the decision partition $U/D = \{D_1, D_2, \dots, D_C\}$. The positive region of D with respect to $\mathcal{G}$ is defined as

$$POS_{\mathcal{G}}^{\beta_R}(D) = \bigcup_{c=1}^C \underline{P}_{\beta_R}^{\mathcal{G}}(D_c). \quad (8)$$

Based on the positive region, the dependency degree of the decision attribute D on the granularity family $\mathcal{G}$ is defined as

$$\gamma_{\mathcal{G}}^{\beta_R}(D) = \frac{|POS_{\mathcal{G}}^{\beta_R}(D)|}{|U|}. \quad (9)$$

The positive region contains all objects that can certainly be assigned to their corresponding decision classes by a sufficient proportion of granularities.

The dependency degree satisfies

$$0 \leq \gamma_{\mathcal{G}}^{\beta_R}(D) \leq 1. \quad (10)$$

When $\gamma_{\mathcal{G}}^{\beta_R}(D) = 1$, all objects can certainly be classified by the granularity family, indicating that the decision attribute completely depends on $\mathcal{G}$ under the specified parameters. When $\gamma_{\mathcal{G}}^{\beta_R}(D) = 0$, no object belongs to the positive region, indicating that the current granularities cannot provide certain support for the decision classification. When $0 < \gamma_{\mathcal{G}}^{\beta_R}(D) < 1$, the decision attribute partially depends on the granularity family.

The dependency degree provides an algebraic measure of the joint decision-preserving capability of multiple granularities. In the proposed method, it is further used to quantify the contribution of an individual feature granularity by comparing the dependency degrees before and after that granularity is removed. Granularities whose removal causes a larger reduction in dependency are considered to provide stronger decision support.

2.2. Particle swarm optimization algorithm

Particle swarm optimization (PSO), proposed by Kennedy and Eberhart, is a population-based optimization algorithm inspired by collective foraging behavior [31]. A swarm consists of P particles, and the position of each particle represents a candidate solution. In feature selection, the i -th particle is encoded as

$$\mathbf{X}_i^t = (x_{i1}^t, x_{i2}^t, \dots, x_{im'}^t), \quad x_{ij}^t \in \{0, 1\}, \quad (11)$$

where m' is the number of candidate features retained after the first-stage screening. The value $x_{ij}^t = 1$ indicates that the j -th candidate feature is selected.

In standard PSO, the continuous velocity is updated using the particle's personal best position and the global best position:

$$\mathbf{V}_i^{t+1} = \omega \mathbf{V}_i^t + c_1 \mathbf{R}_1 \odot (\mathbf{P}_i^{\text{best}} - \mathbf{X}_i^t) + c_2 \mathbf{R}_2 \odot (\mathbf{G}^{\text{best}} - \mathbf{X}_i^t), \quad (12)$$

where ω is the inertia weight, c_1 and c_2 are learning coefficients, $\mathbf{R}_1$ and $\mathbf{R}_2$ are random vectors whose components are sampled from $U(0, 1)$, and $\odot$ denotes element-wise multiplication.

For binary feature selection, the velocity is converted into a selection probability through

$$S(v_{ij}^{t+1}) = \frac{1}{1 + \exp(-v_{ij}^{t+1})}, \quad (13)$$

and the position is updated as

$$x_{ij}^{t+1} = \begin{cases} 1, & r_{ij} < S(v_{ij}^{t+1}), \\ 0, & \text{otherwise,} \end{cases} \quad r_{ij} \sim U(0, 1). \quad (14)$$

Standard PSO is attractive for feature selection because of its simple implementation and fast convergence. However, all particles mainly learn from $\mathbf{P}_i^{\text{best}}$ and a single $\mathbf{G}^{\text{best}}$. In a high-dimensional binary space, this limited learning structure can cause particles to become increasingly similar, reduce population diversity, and converge prematurely. The proposed method therefore retains the personal-best learning mechanism but replaces single global-best guidance with GWO-based multi-leader guidance. It further introduces intra-layer graph learning and upper-layer learning to provide particles with structurally diverse exemplars.

2.3. Grey wolf optimizer algorithm

The grey wolf optimizer (GWO), proposed by Mirjalili et al., models the leadership hierarchy and cooperative hunting mechanism of grey

wolves [32]. Candidate solutions are ranked as α , β , δ , and ω , where α , β , and δ denote the best three solutions and the remaining individuals are treated as ω wolves.

Recent hybrid methods have combined binary GWO with PSO to improve the balance between exploration and exploitation in wrapper feature selection [33]. Such methods demonstrate the benefit of complementary multi-leader and velocity-based search, but they generally perform direct wrapper optimization in the original feature space. In contrast, the proposed framework first removes weak feature granularities using GMNRS and then embeds the complete three-leader GWO hunting mechanism into a layered graph PSO search over the reduced candidate space.

Let $\mathbf{X}^\alpha, \mathbf{X}^\beta, \mathbf{X}^\delta$ denote the three best particles in the current population. For each leader $l \in \{\alpha, \beta, \delta\}$, the coefficient vectors are

$$\mathbf{A}_l = 2a(t)\mathbf{R}_{1,l} - a(t), \quad \mathbf{C}_l = 2\mathbf{R}_{2,l}, \quad (15)$$

where $a(t)$ is a decreasing convergence factor and $\mathbf{R}_{1,l}, \mathbf{R}_{2,l}$ are random vectors sampled from $U(0, 1)$.

The distance between particle $\mathbf{X}_i^l$ and leader $\mathbf{X}^l$ is

$$\mathbf{D}_{i,l} = \left| \mathbf{C}_l \odot \mathbf{X}^l - \mathbf{X}_i^l \right|. \quad (16)$$

The candidate position estimated from leader l is

$$\mathbf{Y}_{i,l} = \mathbf{X}^l - \mathbf{A}_l \odot \mathbf{D}_{i,l}. \quad (17)$$

The three leaders jointly estimate the search direction by

$$\mathbf{X}_i^{GWO} = \frac{\mathbf{Y}_{i,\alpha} + \mathbf{Y}_{i,\beta} + \mathbf{Y}_{i,\delta}}{3}. \quad (18)$$

Accordingly, the GWO guidance direction used in the proposed velocity update is

$$\mathbf{D}_i^{GWO} = \mathbf{X}_i^{GWO} - \mathbf{X}_i^l. \quad (19)$$

Compared with directly replacing the PSO velocity equation with a GWO position update, this formulation preserves the search memory and velocity accumulation of PSO while introducing three complementary global leaders. The α leader provides the strongest current search direction, whereas the β and δ leaders reduce the risk of the population being dominated by a single potentially suboptimal solution.

In summary, the three foundations play non-overlapping roles in the proposed framework. GMNRS supplies a classifier-independent dependency criterion for first-stage granularity screening; binary PSO provides the basic representation, velocity memory, and probabilistic feature selection mechanism; and GWO supplies a complete three-leader global guidance mechanism. Their integration forms a filter-wrapper framework rather than two repeated evolutionary searches: the first stage reduces the search space at the feature-granularity level, while the second stage searches for individual feature combinations within the retained candidate set.

3. A dual-stage feature selection algorithm based on layered graph particle swarm optimization and grey wolf optimizer

High-dimensional feature selection requires both an effective reduction of the search space and a comprehensive evaluation of feature combinations. Although filter methods can efficiently eliminate irrelevant features, they may overlook the collaborative effects among features. In contrast, wrapper methods can evaluate the predictive performance of feature subsets more directly, but searching the complete feature space usually incurs considerable computational costs. To combine the complementary advantages of these two strategies, a dual-stage feature selection framework integrating generalized multigranulation neighborhood rough sets with Layered Graph Particle Swarm Optimization and Grey Wolf Optimizer is developed, and termed DSLGPSO-GWO. In the

first stage, feature granularities are constructed and screened according to their decision-supporting capabilities, thereby generating a reduced candidate feature space. In the second stage, an improved LGPSO-GWO algorithm is employed to further optimize feature combinations within the retained candidate space. The overall workflow of the proposed DSLGPSO-GWO framework is illustrated in Fig. 1.

3.1. Stage I: GMNRS-based feature-granularity screening

Directly performing wrapper-based evolutionary search in the complete feature space may incur considerable computational costs, particularly when the original dataset contains numerous redundant or weakly relevant features. To reduce the search burden of the subsequent LGPSO-GWO optimizer, the first stage evaluates the original feature set at the granularity level. Specifically, adaptive feature granularities are first constructed using K-means. Their decision-supporting capabilities are then evaluated through the GMNRS dependency degree, after which the granularities are ranked according to their external importance. Finally, the highly ranked granularities are retained and expanded to generate the candidate feature set for the second-stage search.

3.1.1. Adaptive construction of feature granularities

Let $A = \{a_1, a_2, \dots, a_m\}$ denote the original conditional attribute set. For a given granularity number K , the K-means algorithm partitions A into a feature granularity family

$$\mathcal{G}_K = \{G_1, G_2, \dots, G_K\}, \quad (20)$$

which satisfies

$$G_r \cap G_s = \emptyset \quad (r \neq s), \quad \bigcup_{r=1}^K G_r = A. \quad (21)$$

Let μ_r denote the center of feature granularity G_r . For a given value of K , the feature granulation process minimizes the within-granularity dispersion:

$$J(K) = \sum_{r=1}^K \sum_{a_j \in G_r} \left\| \mathbf{a}_j - \mu_r \right\|_2^2, \quad (22)$$

where $\mathbf{a}_j$ denotes the data representation of feature a_j .

A fixed number of granularities may not be suitable for datasets with different dimensionalities and feature-distribution structures. Therefore, the silhouette coefficient is introduced to determine the number of granularities adaptively.

For $a_j \in G_r$, its average distance from the other features within the same granularity is defined as

$$\text{Intra}(a_j) = \frac{1}{|G_r| - 1} \sum_{\substack{a_q \in G_r \\ q \neq j}} \left\| \mathbf{a}_j - \mathbf{a}_q \right\|_2. \quad (23)$$

The minimum average distance between a_j and the other feature granularities is defined as

$$\text{Inter}(a_j) = \min_{s \neq r} \left\{ \frac{1}{|G_s|} \sum_{a_q \in G_s} \left\| \mathbf{a}_j - \mathbf{a}_q \right\|_2 \right\}. \quad (24)$$

Accordingly, the silhouette coefficient of a_j is

$$SC(a_j) = \frac{\text{Inter}(a_j) - \text{Intra}(a_j)}{\max \{ \text{Intra}(a_j), \text{Inter}(a_j) \}}. \quad (25)$$

The average silhouette coefficient corresponding to K feature granularities is

$$SC(K) = \frac{1}{m} \sum_{j=1}^m SC(a_j). \quad (26)$$

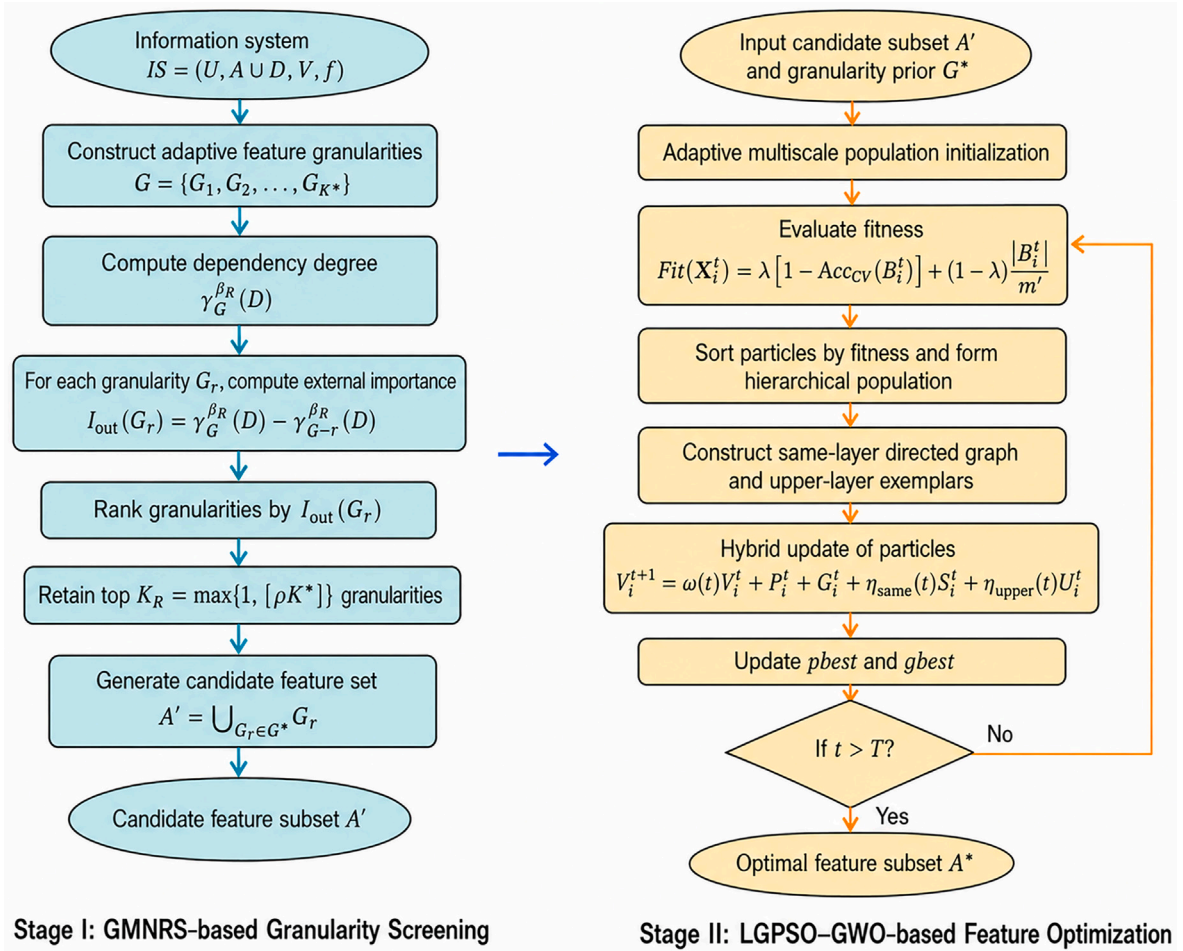


Fig. 1. The flowchart of DSLGPSO-GWO framework.

Given a candidate set of granularity numbers $\mathcal{K} = \{K_1, K_2, \dots, K_Q\}$, the optimal number of feature granularities is determined by

$$K^* = \arg \max_{K \in \mathcal{K}} SC(K). \quad (27)$$

The resulting adaptive feature granularity family is denoted by

$$\mathcal{G} = \{G_1, G_2, \dots, G_{K^*}\}. \quad (28)$$

The silhouette-based strategy evaluates both the compactness within each granularity and the separation among different granularities. It therefore avoids assigning the same predetermined granularity number to datasets with different feature structures.

3.1.2. GMNRS-based granularity importance evaluation

After obtaining the feature granularity family $\mathcal{G}$, each granularity G_r induces a neighborhood relation in the sample space. Since different granularities may contain different numbers of features, the distance between samples x_p and x_q under G_r is defined as

$$d_{G_r}(x_p, x_q) = \sqrt{\frac{1}{|G_r|} \sum_{a_j \in G_r} (a_j(x_p) - a_j(x_q))^2}. \quad (29)$$

The normalization by $|G_r|$ reduces the direct influence of granularity size on the neighborhood distance. Given the neighborhood radius δ , the neighborhood relation induced by each feature granularity is obtained according to Eq. (2).

For each granularity G_r , its support for the decision assignment of an object is calculated using the multigranulation support function in Eq. (4). Based on the support results for all feature granularities, the generalized lower and upper approximations are determined according to Eqs. (5) and (6), respectively.

The positive region of the decision attribute is then obtained using Eq. (8), and the dependency degree of D on the complete feature granularity family $\mathcal{G}$ is calculated according to Eq. (9). The resulting dependency degree reflects the joint decision-supporting capability of all feature granularities.

However, the dependency degree of the complete family cannot directly distinguish the contribution of an individual granularity. Therefore, a leave-one-granularity-out strategy is introduced.

Definition 5. For $G_r \in \mathcal{G}$, define the reduced granularity family as

$$\mathcal{G}^- = \mathcal{G} \setminus \{G_r\}. \quad (30)$$

After removing G_r , the multigranulation approximation and dependency degree are recalculated using the remaining $K^* - 1$ granularities. The external importance of G_r is defined as

$$I_{out}(G_r) = \gamma_{\mathcal{G}}^{\beta_R}(D) - \gamma_{\mathcal{G}^-}^{\beta_R}(D). \quad (31)$$

Property 1. The external importance of a feature granularity can be interpreted as follows:

1. If $I_{out}(G_r) > 0$, removing G_r decreases the dependency degree, indicating that G_r provides useful decision-supporting information.

2. If $I_{\text{out}}(G_r) = 0$, removing G_r does not change the dependency degree, indicating that the contribution of G_r may be covered by the remaining granularities.
3. If $I_{\text{out}}(G_r) < 0$, removing G_r increases the dependency degree, indicating that G_r may introduce conflicting neighborhood information.

Proof. The three cases follow directly from Eq. (31) by comparing $\gamma_G^{\beta_R}(D)$ with $\gamma_{G-r}^{\beta_R}(D)$. $\square$

It should be noted that removing G_r changes the number of granularities from K^* to $K^* - 1$. Consequently, the multigranulation support proportion must be recalculated using the remaining granularities. Under this mechanism, the external importance may theoretically be negative. To preserve its original decision interpretation, no nonnegative truncation is applied.

After calculating the external importance of all feature granularities, they are ranked in descending order:

$$I_{\text{out}}(G_{(1)}) \geq I_{\text{out}}(G_{(2)}) \geq \dots \geq I_{\text{out}}(G_{(K^*)}), \quad (32)$$

where $G_{(r)}$ denotes the granularity ranked in the r -th position. When multiple granularities have the same external importance, their original order is retained.

The first-stage ranking is determined solely by the GMNRS-based external importance. Classification accuracy, the F-score, and the second-stage wrapper fitness do not participate in the granularity ranking process. Thus, the first stage remains independent of the classifier used in the subsequent wrapper search.

Algorithm 1: GMNRS-Based Feature-Granularity Screening.

Input: Information system $IS = (U, A \cup D, V, f)$; candidate granularity-number set $\mathcal{K}$; neighborhood radius δ ; GMNRS parameter β_R ; granularity retention ratio ρ .

Output: Retained feature-granularity family $\mathcal{G}^*$; external-importance set of the retained granularities I^* ; candidate feature set A' .

- 1 **for each** $K \in \mathcal{K}$ **do**
- 2 $\mathcal{G}_K \leftarrow \text{KMeans}(A, K)$;
- 3 Calculate $SC(K)$ according to Eq. (26);
- 4 $K^* \leftarrow \arg \max_{K \in \mathcal{K}} SC(K)$, $\mathcal{G} \leftarrow \mathcal{G}_{K^*}$;
- 5 Calculate $\gamma_G^{\beta_R}(D)$ according to Eq. (9);
- 6 **for each** $G_r \in \mathcal{G}$ **do**
- 7 $\mathcal{G}^- \leftarrow \mathcal{G} \setminus \{G_r\}$;
- 8 Calculate $\gamma_{\mathcal{G}^-}^{\beta_R}(D)$;
- 9 Calculate $I_{\text{out}}(G_r)$ according to Eq. (31);
- 10 Sort $\mathcal{G}$ by $I_{\text{out}}(G_r)$ in descending order;
- 11 $K_R \leftarrow \max\{1, \lceil \rho K^* \rceil\}$;
- 12 $\mathcal{G}^* \leftarrow \{G_{(1)}, G_{(2)}, \dots, G_{(K_R)}\}$;
- 13 $I^* \leftarrow \{I_{\text{out}}(G_r) \mid G_r \in \mathcal{G}^*\}$;
- 14 $A' \leftarrow \bigcup_{G_r \in \mathcal{G}^*} G_r$;
- 15 **return** $\mathcal{G}^*, I^*, A'$;

3.1.3. Granularity screening and candidate feature generation

After ranking the feature granularities, a retention ratio $\rho \in (0, 1]$ is introduced to control the first-stage screening strength. The number of retained feature granularities is calculated as

$$K_R = \max\{1, \lceil \rho K^* \rceil\}. \quad (33)$$

According to the ranking in Eq. (32), the retained feature granularity family is

$$\mathcal{G}^* = \{G_{(1)}, G_{(2)}, \dots, G_{(K_R)}\}. \quad (34)$$

The candidate feature set is generated by expanding all retained feature granularities:

$$A' = \bigcup_{G_r \in \mathcal{G}^*} G_r, \quad (35)$$

where $A' \subseteq A$. Let $m' = |A'|$ denote the number of candidate features.

The resulting candidate feature set A' constitutes the search space of the second-stage LGPSO-GWO optimizer. Therefore, the particle dimension is reduced from the original feature number m to the candidate feature number m' , where $m' \leq m$.

3.2. Stage II: LGPSO-GWO-based feature subset optimization

Although the candidate feature set generated in Stage I substantially reduces the original search space, it may still contain redundant features, and the interactions among the retained features have not been fully evaluated. Consequently, an improved LGPSO-GWO algorithm is employed to further search for an informative and compact feature subset within the candidate space.

Let $A' = \{a'_1, a'_2, \dots, a'_{m'}\}$ denote the candidate feature set generated in Stage I, where $m' = |A'|$. At iteration t , the position of the i -th particle is represented by a binary vector

$$\mathbf{X}_i^t = (x_{i1}^t, x_{i2}^t, \dots, x_{im'}^t), \quad x_{ij}^t \in \{0, 1\}.$$

If $x_{ij}^t = 1$, candidate feature a'_j is selected; otherwise, it is excluded. Accordingly, the feature subset represented by $\mathbf{X}_i^t$ is denoted by $B_i^t = \{a'_j \mid x_{ij}^t = 1\}$.

To simultaneously consider predictive performance and subset compactness, the fitness of particle $\mathbf{X}_i^t$ is defined as

$$\text{Fit}(\mathbf{X}_i^t) = \lambda [1 - \text{Acc}_{\text{CV}}(B_i^t)] + (1 - \lambda) \frac{|B_i^t|}{m'}, \quad (36)$$

where $\text{Acc}_{\text{CV}}(B_i^t)$ denotes the mean classification accuracy obtained through cross-validation using B_i^t , and $\lambda \in [0, 1]$ controls the trade-off between classification performance and the number of selected features. Since Eq. (36) is minimized, a particle with a smaller fitness value represents a more desirable feature subset.

The two stages therefore perform complementary functions. Stage I uses the classifier-independent GMNRS external importance to remove weak feature granularities and reduce the search dimension, whereas Stage II uses classifier-based evaluation to examine the joint predictive capability of candidate feature combinations.

3.2.1. Adaptive multiscale population initialization

Random initialization does not exploit the decision-support information obtained in Stage I and may generate many particles with similar subset sizes. To provide informative initial solutions while maintaining sufficient population diversity, a granularity-feature hierarchical probability is constructed and incorporated into an adaptive multiscale initialization strategy.

Definition 6. Let $\mathcal{G}^* = \{G_1, G_2, \dots, G_{K_R}\}$ denote the feature granularities retained in Stage I. Since the external importance may take negative values, define its shifted form as

$$\tilde{I}_r = I_{\text{out}}(G_r) - \min_{G_s \in \mathcal{G}^*} I_{\text{out}}(G_s) + \epsilon, \quad (37)$$

where $\epsilon > 0$. The granularity-level initialization probability is

$$P(G_r) = \frac{\tilde{I}_r}{\sum_{G_s \in \mathcal{G}^*} \tilde{I}_s}. \quad (38)$$

For $a'_j \in G_r$, the conditional feature probability based on the nonnegative ANOVA F -statistic is defined as

$$P(a'_j \mid G_r) = \frac{F_j + \epsilon}{\sum_{a'_q \in G_r} (F_q + \epsilon)}, \quad (39)$$

where F_j denotes the F -statistic of candidate feature a'_j . The unnormalized hierarchical prior and the final normalized feature prior are respectively defined as

$$\tilde{p}_j = P(G_r)P(a'_j | G_r), \quad a'_j \in G_r, \quad (40)$$

and

$$p_j = \frac{\tilde{p}_j + \varepsilon}{\sum_{q=1}^{m'} (\tilde{p}_q + \varepsilon)}. \quad (41)$$

Eq. (38) assigns larger probabilities to granularities with stronger decision-supporting capabilities, while Eq. (39) differentiates the features within each retained granularity.

Property 2. For every candidate feature a'_j , the hierarchical prior satisfies

$$p_j > 0, \quad \sum_{j=1}^{m'} p_j = 1. \quad (42)$$

Proof. Because $\varepsilon > 0$, Eq. (37) gives $\tilde{I}_r > 0$, and therefore $P(G_r) > 0$. Since the ANOVA F -statistic is nonnegative, Eq. (39) also yields $P(a'_j | G_r) > 0$. Hence $\tilde{p}_j > 0$. The numerator in Eq. (41) is therefore positive, and the denominator is the sum of all such positive numerators, which proves both claims. $\square$

The F -statistic in Eq. (39) is used only to differentiate features within a retained granularity during population initialization. It does not participate in the granularity ranking or retention procedure of Stage I. Therefore, the first-stage screening criterion remains exclusively determined by the GMNRS external importance.

Definition 7. Let N_p denote the population size. The target subset size of the i -th particle is defined as

$$k_i = 1 + \left\lfloor \frac{i-1}{N_p-1} (m'-1) \right\rfloor, \quad i = 1, 2, \dots, N_p. \quad (43)$$

To balance prior-guided exploitation and uniform exploration, define the particle-specific mixing coefficient as

$$\xi_i = \frac{i-1}{N_p-1}, \quad (44)$$

and the sampling probability of feature a'_j for particle i as

$$q_{ij} = (1 - \xi_i)p_j + \xi_i \frac{1}{m'}. \quad (45)$$

The i -th initial particle is generated by sampling k_i features without replacement according to $\{q_{ij}\}_{j=1}^{m'}$.

Property 3. The adaptive multiscale initialization satisfies the following conditions:

1. $k_1 = 1$;
2. $k_{N_p} = m'$;
3. $k_{i+1} \geq k_i$, for $i = 1, 2, \dots, N_p - 1$;
4. $q_{ij} > 0$ and $\sum_{j=1}^{m'} q_{ij} = 1$ for every particle i .

Proof. Eq. (43) is nondecreasing in i , and substituting $i = 1$ and $i = N_p$ gives the two endpoint values. Moreover, $\xi_i \in [0, 1]$. By the definition of the hierarchical feature prior, $\{p_j\}$ is a positive probability distribution, while $\{1/m'\}$ is the uniform distribution. Therefore, Eq. (45) is a convex combination of two probability distributions, which proves the fourth condition. $\square$

The target sizes of the initial particles therefore span the sparse and dense regions of the candidate solution space. The mixing mechanism also prevents excessive reliance on a limited number of high-prior features.

For particles with small ξ_i , initialization is primarily guided by the hierarchical prior, thereby increasing the probability of including features with strong decision-supporting and discriminative capabilities. As ξ_i increases, the distribution gradually approaches the uniform distribution, which expands the structural coverage of the initial population.

For each particle, k_i features are sampled without replacement according to Eq. (45). When a newly generated particle duplicates an existing one, its mixing level is progressively increased toward the uniform distribution and the particle is resampled. This retry mechanism improves the uniqueness of the initial solutions without discarding the prior information. Finally, the particle order is randomly permuted so that the particle index is not permanently associated with a particular subset size. The initial velocities are generated randomly within a bounded interval.

The proposed initialization strategy incorporates granularity importance, within-granularity feature discrimination, multiscale subset coverage, and random exploration. It therefore provides the subsequent evolutionary process with both informative starting points and sufficient structural diversity.

3.2.2. Layered graph-based multi-leader learning

After the initial population is evaluated, the particles are sorted in ascending order of fitness and assigned to a sequence of ordered layers $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_L$, where particles in an upper layer generally have better fitness values than those in a lower layer. The population is repartitioned at each iteration so that the layered structure can adapt to the continuously changing search states.

Definition 8. For particle $\mathbf{X}_i^t$, let $\mathcal{C}_i^{\text{same}}$ denote the other particles in its current layer. The valid same-layer learning candidates are defined as

$$\mathcal{B}_i^{\text{same}} = \left\{ j \in \mathcal{C}_i^{\text{same}} \mid \text{Fit}(\mathbf{X}_j^t) < \text{Fit}(\mathbf{X}_i^t) \right\}. \quad (46)$$

For $j \in \mathcal{B}_i^{\text{same}}$, define the fitness advantage and normalized Hamming distance as

$$\Delta_{ij}^t = \text{Fit}(\mathbf{X}_i^t) - \text{Fit}(\mathbf{X}_j^t), \quad (47)$$

and

$$d_H(\mathbf{X}_i^t, \mathbf{X}_j^t) = \frac{1}{m'} \sum_{h=1}^{m'} |x_{ih}^t - x_{jh}^t|. \quad (48)$$

The weight of the directed learning relationship from particle j to particle i is then defined as

$$w_{ij}^t = \frac{\Delta_{ij}^t}{\max_{k \in \mathcal{B}_i^{\text{same}}} \Delta_{ik}^t + \varepsilon} \left[1 + d_H(\mathbf{X}_i^t, \mathbf{X}_j^t) \right]. \quad (49)$$

Property 4. For every directed edge $j \rightarrow i$ induced by Eq. (46), its weight satisfies

$$0 < w_{ij}^t < 2. \quad (50)$$

Moreover, the directed learning graph within each layer contains no directed cycle.

Proof. For $j \in \mathcal{B}_i^{\text{same}}$, $\Delta_{ij}^t > 0$. Since $0 \leq d_H(\mathbf{X}_i^t, \mathbf{X}_j^t) \leq 1$ and

$$0 < \frac{\Delta_{ij}^t}{\max_{k \in \mathcal{B}_i^{\text{same}}} \Delta_{ik}^t + \varepsilon} < 1,$$

Eq. (49) gives $0 < w_{ij}^t < 2$. For acyclicity, observe that every directed edge $j \rightarrow i$ satisfies $\text{Fit}(\mathbf{X}_j^t) < \text{Fit}(\mathbf{X}_i^t)$. A directed cycle would therefore

imply a strict chain of fitness inequalities that eventually returns to the starting particle, which is impossible. $\square$

A candidate receives a larger graph weight when it has both a greater fitness advantage and a larger structural difference from the current particle. This encourages each particle to learn from solutions that are not only better but also capable of providing a distinct search direction.

At most two candidates with the largest graph weights are selected, and their weights are normalized:

$$\bar{w}_{ij}^t = \frac{w_{ij}^t}{\sum_{k \in \mathcal{E}_i^{\text{same}}} w_{ik}^t + \varepsilon}, \quad j \in \mathcal{E}_i^{\text{same}}, \quad (51)$$

where $\mathcal{E}_i^{\text{same}} \subseteq \mathcal{B}_i^{\text{same}}$ contains the selected intra-layer exemplars. The corresponding intra-layer direction is

$$\mathbf{d}_i^{\text{same},t} = \sum_{j \in \mathcal{E}_i^{\text{same}}} \bar{w}_{ij}^t (\mathbf{X}_j^t - \mathbf{X}_i^t). \quad (52)$$

If no particle in the current layer has a better fitness value, $\mathcal{B}_i^{\text{same}} = \emptyset$, and the individual historical best position is used as the fallback exemplar:

$$\mathbf{d}_i^{\text{same},t} = \mathbf{p}_i^{\text{best}} - \mathbf{X}_i^t. \quad (53)$$

The actual intra-layer learning term is

$$\mathbf{S}_i^t = c_{\text{same}} \mathbf{R}_i^{\text{same}} \odot \mathbf{d}_i^{\text{same},t}, \quad (54)$$

where c_{same} is the intra-layer learning factor, $\mathbf{R}_i^{\text{same}}$ is a random vector with elements uniformly distributed in $[0, 1]$, and $\odot$ denotes element-wise multiplication.

In addition to intra-layer competition, lower-layer particles receive guidance from superior layers. For a particle in layer $\mathcal{L}_j$, the algorithm searches successively from $\mathcal{L}_{j-1}$ toward the upper layers and collects at most two distinct particle positions. Let $\mathcal{E}_i^{\text{upper}}$ denote the resulting upper-layer exemplar set. If fewer than two distinct exemplars are available, the individual historical best position is used as a supplementary exemplar.

The upper-layer target position is defined as

$$\mathbf{T}_i^{\text{upper},t} = \frac{1}{|\mathcal{E}_i^{\text{upper}}|} \sum_{\mathbf{X}_j^t \in \mathcal{E}_i^{\text{upper}}} \mathbf{X}_j^t, \quad (55)$$

and the corresponding upper-layer learning term is

$$\mathbf{U}_i^t = c_{\text{upper}} \mathbf{R}_i^{\text{upper}} \odot (\mathbf{T}_i^{\text{upper},t} - \mathbf{X}_i^t). \quad (56)$$

For a particle in the top layer, no superior layer exists. Its historical best position is therefore used to construct the upper-layer term. The explicit fallback rules in Eqs. (53)–(56) ensure that the learning mechanism remains well-defined even when the current layer contains no superior competitor or the upper layers do not provide enough distinct exemplars.

The intra-layer graph emphasizes competitive collaboration among particles with comparable fitness, whereas upper-layer learning transmits high-quality information along the population hierarchy. Their combination provides multiple search directions and reduces dependence on a single global-best particle.

3.2.3. Hybrid LGPSO-GWO update and diversity maintenance

At each iteration, the three particles with the smallest fitness values are identified as the α , β , and δ leaders. Rather than directly using their position differences as a social term, the complete encircling and hunting mechanism of GWO is employed to generate the leader-guided target position.

To achieve a gradual transition from global exploration to local exploitation, the conventional convergence factor is replaced by the following nonlinear schedule:

$$a(t) = 2 \left[1 - \left(\frac{t}{T} \right)^2 \right], \quad (57)$$

where t is the current iteration and T is the maximum number of iterations. Compared with a linear schedule, this form maintains a relatively large exploration range during the early iterations and decreases more rapidly near the end of the search.

Using the GWO encircling and hunting equations introduced previously, the α , β , and δ leaders generate a GWO target position $\mathbf{T}_i^{\text{GWO},t}$ for particle i . The corresponding GWO guidance term is defined as

$$\mathbf{G}_i^t = c_{\text{GWO}} \mathbf{R}_i^{\text{GWO}} \odot (\mathbf{T}_i^{\text{GWO},t} - \mathbf{X}_i^t), \quad (58)$$

where c_{GWO} is the GWO guidance factor and $\mathbf{R}_i^{\text{GWO}}$ is a random vector. Consequently, the GWO component retains the distance estimation, encircling, and three-leader cooperative hunting mechanisms rather than merely adding the position differences of the three best particles.

A cosine-decreasing inertia weight is employed to further balance the search process:

$$\omega(t) = \omega_{\min} + \frac{\omega_{\max} - \omega_{\min}}{2} \left[1 + \cos \left(\frac{\pi t}{T} \right) \right]. \quad (59)$$

A large inertia weight at the beginning promotes broad exploration, whereas a smaller value near the end facilitates local refinement.

The relative contributions of intra-layer and upper-layer learning are also adjusted dynamically:

$$\eta_{\text{same}}(t) = 0.7 - 0.4 \frac{t}{T}, \quad (60)$$

$$\eta_{\text{upper}}(t) = 1 - \eta_{\text{same}}(t). \quad (61)$$

The intra-layer component receives a larger weight during the early search to exploit structural differences among comparable particles. As the iteration proceeds, the contribution of upper-layer guidance increases, encouraging the population to converge toward more competitive regions.

By integrating individual memory, GWO guidance, intra-layer competition, and upper-layer learning, the velocity of a non-elite particle is updated as

$$\mathbf{V}_i^{t+1} = \omega(t) \mathbf{V}_i^t + c_1 \mathbf{R}_i^1 \odot (\mathbf{p}_i^{\text{best}} - \mathbf{X}_i^t) + \mathbf{G}_i^t + \eta_{\text{same}}(t) \mathbf{S}_i^t + \eta_{\text{upper}}(t) \mathbf{U}_i^t, \quad (62)$$

where c_1 is the cognitive learning factor and $\mathbf{R}_i^1$ is a random vector with elements in $[0, 1]$.

The updated velocity is restricted to $[-V_{\max}, V_{\max}]$ to prevent the binary transition probabilities from becoming prematurely saturated. The standard sigmoid-based binary mapping is subsequently applied to obtain $\mathbf{X}_i^{t+1}$. If the resulting particle contains no selected feature, the feature with the largest hierarchical prior probability is activated, ensuring that every particle represents a nonempty feature subset.

The α , β , and δ particles are treated as elite solutions during each iteration. Their positions, velocities, and fitness values are preserved while the remaining particles are updated. This mechanism prevents the best feature subsets discovered by the current population from being lost because of stochastic position changes.

Although elite preservation improves convergence stability, excessive concentration around a few elite solutions may reduce population diversity. The average pairwise diversity at iteration t is therefore measured by

$$\text{Div}(t) = \frac{2}{N_p(N_p - 1)} \sum_{i=1}^{N_p-1} \sum_{j=i+1}^{N_p} d_H(\mathbf{X}_i^t, \mathbf{X}_j^t). \quad (63)$$

A perturbation operation is activated when the global best fitness has not improved for a predefined number of consecutive iterations or

when $\text{Div}(t)$ falls below a specified threshold. The three elite particles are excluded from perturbation. A proportion of the particles with the worst fitness values is selected, and a small number of their dimensions are flipped.

The dimension-selection probability for perturbing particle i is constructed from the hierarchical feature prior:

$$\tilde{q}_{ij}^{\text{mut}} = \begin{cases} p_j, & x_{ij}^t = 0, \\ 1 - p_j, & x_{ij}^t = 1. \end{cases} \quad (64)$$

After normalization, features with high prior probabilities are more likely to be activated, whereas currently selected features with low prior probabilities are more likely to be removed. The velocities of the perturbed dimensions are regenerated within $[-V_{\max}, V_{\max}]$, and the affected particles are reevaluated. Their historical best positions and the global best solution are then updated whenever an improvement occurs.

This perturbation mechanism introduces new structural information into low-quality particles while preserving the current elites. It therefore helps the population recover from search stagnation and reduces the risk of premature convergence.

After the maximum number of iterations is reached, the global best particle is returned. The positions with a value of one identify the selected features in A' . By mapping these positions back to the original conditional attribute set, the final feature subset is obtained as $A^* \subseteq A' \subseteq A$. The overall iterative structure of the proposed LGPSO–GWO optimizer is illustrated in Fig. 2, while the complete optimization procedure of Stage II is summarized in Algorithm 2.

3.3. Computational complexity analysis

The computational complexity of the proposed DSLGPSO–GWO method is analyzed according to its two-stage procedure. Let n and m denote the numbers of samples and original features, respectively, and let $m' = |A'|$ denote the number of candidate features retained after Stage I, where $m' \leq m$. Moreover, N_p and T denote the population size and the maximum number of iterations, respectively. Since the candidate granularity numbers and the control parameters of K-means are fixed in the implementation, they are treated as constants in the following asymptotic analysis.

In Stage I, each feature is regarded as an n -dimensional object and adaptive K-means is used to construct the feature granularities. Under

Algorithm 2: LGPSO–GWO-Based Feature Subset Optimization.

Input: Information system $IS = (U, A \cup D, V, f)$; candidate feature set A' ; retained feature-granularity family $\mathcal{G}^*$; corresponding external-importance set $\mathcal{I}^*$; wrapper classifier C ; Stage-II control parameter set Θ_2 .

Output: Optimized feature subset A^* .

- 1 Calculate the hierarchical feature prior $\{p_j\}_{j=1}^{m'}$ according to Eqs. (37)–(41);
- 2 Initialize $\{X_i^0, V_i^0\}_{i=1}^{N_p}$ using the adaptive multiscale strategy;
- 3 Initialize $\mathbf{P}^{\text{best}}$, $\mathbf{G}^{\text{best}}$, and $s \leftarrow 0$;
- 4 **for** $t = 1$ **to** T **do**
- 5 Sort particles by fitness and construct $\mathcal{L}_1, \dots, \mathcal{L}_L$;
- 6 Select α, β, δ as elites $\mathcal{E}^t$;
- 7 Update $a(t)$, $\omega(t)$, $\eta_{\text{same}}(t)$, and $\eta_{\text{upper}}(t)$;
- 8 **for each** $i \notin \mathcal{E}^t$ **do**
- 9 Select same-layer and upper-layer exemplars;
- 10 Calculate S_i^t, U_i^t , and G_i^t ;
- 11 Update V_i^{t+1} according to Eq. (62);
- 12 Update X_i^{t+1} by binary mapping;
- 13 Preserve elites and evaluate the population;
- 14 Update $\mathbf{P}^{\text{best}}$, $\mathbf{G}^{\text{best}}$, and s ;
- 15 Calculate $\text{Div}(t)$;
- 16 **if** $s \geq \tau_s$ **or** $\text{Div}(t) < \tau_d$ **then**
- 17 Perturb the worst non-elite particles according to Eq. (64);
- 18 Reevaluate the perturbed particles and update $\mathbf{P}^{\text{best}}$ and $\mathbf{G}^{\text{best}}$;
- 19 $s \leftarrow 0$;
- 20 $A^* \leftarrow \text{Decode}(\mathbf{G}^{\text{best}})$;
- 21 **return** A^* ;

the fixed K-means settings, this process requires $\mathcal{O}(mn)$. The optimal granularity number is determined using the silhouette coefficient. Let $s \leq m$ denote the number of feature objects involved in silhouette evaluation. Its computational cost is approximately $\mathcal{O}(s^2n)$; when all features are used, $s = m$.

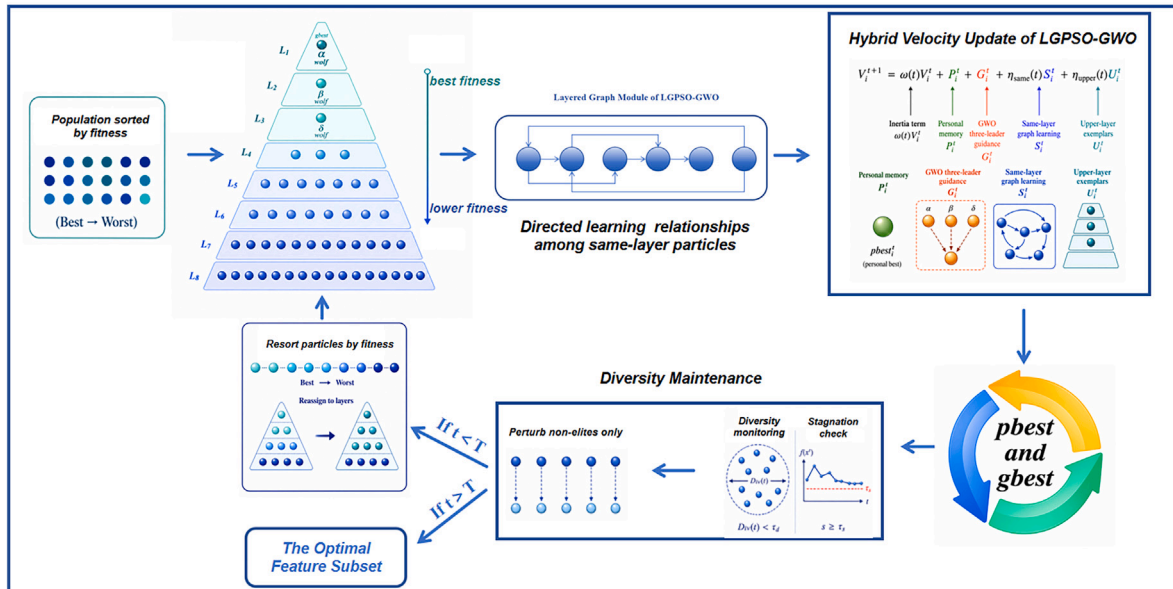


Fig. 2. Overall iterative structure of the LGPSO–GWO algorithm.

After the feature granularities are obtained, the neighborhood relations among samples are constructed for GMNRS evaluation. For a feature granularity G_r containing m_r features, calculating the pairwise distances among n samples requires $\mathcal{O}(n^2 m_r)$. Since the feature granularities form a partition of the original feature set and $\sum_r m_r = m$, constructing the support information for all granularities requires

$$\mathcal{O}\left(\sum_r n^2 m_r\right) = \mathcal{O}(n^2 m).$$

The blockwise distance calculation adopted in the implementation reduces the peak memory consumption without changing this asymptotic time complexity.

Once the support information has been obtained, the GMNRS dependency degrees and the external importance of the feature granularities can be calculated by reusing the existing support results. Their costs, together with granularity ranking and candidate feature generation, are lower than the pairwise-distance calculation and can therefore be absorbed into the preceding terms. Consequently, the overall complexity of Stage I can be expressed as

$$\mathcal{T}_I = \mathcal{O}(mn + s^2 n + n^2 m). \quad (65)$$

In Stage II, the F -statistics and the granularity–feature hierarchical prior are first calculated, followed by adaptive multiscale population initialization. These operations require at most $\mathcal{O}(nm' + N_p m')$, which is lower than the subsequent iterative optimization cost.

During each iteration, the particles are ranked and dynamically organized into multiple layers. The construction of same-layer learning relationships involves pairwise comparisons among particles, while upper-layer guidance, three-leader GWO learning, velocity updating, binary position mapping, elite preservation, and diversity maintenance are also performed in the candidate feature space. Therefore, the non-classifier evolutionary operations over T iterations are bounded by $\mathcal{O}(TN_p^2 m')$.

The dominant computational cost of Stage II generally arises from wrapper-based fitness evaluation. Let $C_{CV}(n, d)$ denote the cost of evaluating a feature subset containing d features through cross-validation, including classifier training and prediction. At most N_p particles are evaluated during initialization and in each iteration. The additional evaluations caused by the diversity-maintenance mechanism involves only a fixed proportion of the population and can therefore be absorbed into the same asymptotic term. Accordingly, the complexity of Stage II is

$$\mathcal{T}_{II} = \mathcal{O}\left(TN_p^2 m' + (T+1)N_p C_{CV}(n, m')\right), \quad (66)$$

where the lower-order initialization terms are omitted.

For the KNN classifier used in the wrapper search, cross-validation-based evaluation requires approximately $C_{CV}(n, m') = \mathcal{O}(n^2 m')$, because each test object is compared with the training objects in the selected feature subspace. Therefore, the complexity of Stage II becomes

$$\mathcal{T}_{II} = \mathcal{O}\left(TN_p^2 m' + (T+1)N_p n^2 m'\right). \quad (67)$$

Combining the two stages, the overall computational complexity of DSLGPSO–GWO is

$$\mathcal{T}_{\text{total}} = \mathcal{O}\left(mn + s^2 n + n^2 m + TN_p^2 m' + (T+1)N_p C_{CV}(n, m')\right). \quad (68)$$

Although Stage I introduces additional costs for adaptive feature granulation and GMNRS-based screening, it reduces the wrapper-search dimension from m to m' . A conventional single-stage evolutionary feature selection method operating directly on the original feature space requires approximately $\mathcal{O}(TN_p C_{CV}(n, m))$ for repeated fitness evaluations. In contrast, DSLGPSO–GWO performs the dominant wrapper search in the reduced candidate space A' . Since $m' \leq m$, and typically $m' \ll m$ for high-dimensional data, Stage I can substantially reduce the repeated classifier-evaluation cost of the subsequent evolutionary search.

4. Experiments and analysis

This section systematically evaluates the proposed DSLGPSO–GWO method from the perspectives of component effectiveness, parameter robustness, classification performance, feature reduction capability, convergence behavior, and statistical significance. All experiments were implemented in Python 3.8 and conducted on a computer running Windows 11 (version 25H2), equipped with an AMD Ryzen 9 7940HX processor with 16 CPU cores and 16 GB of RAM. No parallel computing or GPU-based hardware acceleration was employed during the experiments. The experimental study is organized as follows. Section 4.1 introduces the benchmark datasets and experimental settings. Section 4.2 describes the comparison algorithms used to evaluate DSLGPSO–GWO from different methodological perspectives. Section 4.3 conducts component-wise ablation experiments to investigate the individual contributions of the first-stage screening, adaptive multiscale initialization, layered graph learning, and GWO guidance mechanisms. Section 4.4 analyzes the influence of the granularity retention ratio ρ on the classification performance. Section 4.5 compares DSLGPSO–GWO with representative feature selection and swarm-optimization methods in terms of classification accuracy and the number of selected features. Section 4.6 examines the convergence behavior of the proposed optimizer against several recent PSO variants. Finally, Section 4.7 employs non-parametric statistical tests and critical difference diagrams to further verify whether the observed performance differences are statistically reliable.

4.1. Datasets and experimental settings

To comprehensively evaluate the proposed DSLGPSO–GWO on datasets with different sample sizes, feature dimensions, and class distributions, twelve publicly available benchmark datasets are employed in the experiments. Most of the datasets are obtained from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/>). The Texture dataset is collected from the KEEL Dataset Repository (<https://sci2s.ugr.es/keel/dataset.php?cod=72>) and is also publicly available through OpenML (<https://www.openml.org/d/40499>). In addition, Lungcancer-1 is a high-dimensional microarray benchmark originating from Brigham and Women's Hospital and Harvard Medical School, and its public benchmark information is available from a microarray dataset repository (<https://csse.szu.edu.cn/staff/zhuzx/datasets.html>). The detailed characteristics of all datasets are summarized in Table 1.

As shown in Table 1, the twelve datasets contain 181 to 13,611 samples, 7 to 12,533 conditional features, and 2 to 11 decision classes. The datasets therefore cover substantially different data characteristics, ranging from conventional low- and medium-dimensional classification problems to high-dimensional small-sample scenarios. In particular, Lungcancer-1 contains more than twelve thousand features but only a

Table 1
Datasets.

Nos.	Datasets	Abbreviation	Samples	Features	Class
1	Acoustic	Acoustic	400	50	4
2	Climate Model Simulation	Climate	540	18	2
3	Crashes				
3	Breast Cancer	Breast	683	10	2
4	Wine Quality (Red)	Wine-red	1599	11	6
5	AV	AV	2043	7	2
6	Cardiotocography	Card	2126	20	3
7	Statlog (Image Segmentation)	Segment	2310	19	7
8	Spambase	Spambase	4601	57	2
9	Texture	Texture	5500	40	11
10	Dry Bean	Bean	13,611	16	7
11	Electrical Grid Stability	EGSS	10,000	13	2
12	Simulated Data				
12	LungCancer-Brigham And Women Hospital	Lungcancer-1	181	12,533	2

small number of samples, providing a challenging case for evaluating the search-space reduction capability of the proposed dual-stage framework.

To reduce the evaluation bias caused by a single learning model, three classifiers with different learning mechanisms are employed to evaluate the feature subsets, including K-Nearest Neighbor (KNN), Support Vector Machine with a radial basis function kernel (SVM-RBF), and Multi-Layer Perceptron (MLP). KNN evaluates the local neighborhood discrimination of the selected features, SVM-RBF examines their ability to support nonlinear decision boundaries, and MLP further evaluates their representation capability under a neural-network-based classifier. The classification results are reported in the form of mean accuracy $\pm$ standard deviation.

For the proposed DSLGPSO-GWO, the neighborhood radius and the generalized multigranulation parameter in Stage I are set to $\delta = 0.20$ and $\beta_R = 0.70$, respectively, and the default granularity retention ratio is set to $\rho = 0.70$. The candidate numbers of feature granularities are selected from $\{2, 3, 4, 5, 6, 8, 10, 12\}$, and the optimal granularity number is adaptively determined according to the silhouette coefficient. For K-means clustering, the number of initializations and the maximum number of iterations are set to 20 and 500, respectively.

In Stage II, the population size and the maximum number of iterations are set to $N_p = 50$ and $T = 50$, respectively. Five-fold stratified cross-validation is employed during the feature subset evaluation, and KNN with $k = 3$ is used as the search classifier. The population is divided into eight layers with sizes $(1, 1, 1, 3, 7, 9, 12, 16)$ from the top layer to the bottom layer. The inertia weight is dynamically adjusted between $w_{\max} = 0.90$ and $w_{\min} = 0.40$. The learning coefficients are set to $c_1 = 1.49$, $c_{\text{GWO}} = 1.49$, $c_{\text{same}} = 0.30$, and $c_{\text{upper}} = 0.80$, while the maximum velocity is set to $V_{\max} = 6.0$. The adaptive multiscale initialization strategy is adopted for the initial population. Furthermore, the stagnation threshold is set to 5 iterations, the mutation rate is set to 0.01, the proportion of perturbed particles is set to 0.30, and the diversity threshold is set to 0.05. The random seed is fixed at 42 to improve experimental reproducibility.

4.2. Comparison algorithms

To comprehensively evaluate DSLGPSO-GWO, seven representative algorithms are selected for comparison, including three rough-set- or graph-based feature selection methods, one ant-colony-based feature selection method, and three recent PSO variants. Specifically, GLSFS, KND-UFS, and FNGCE are representative filter-based feature selection approaches, while QACO is a recently developed swarm-intelligence method specifically designed for feature selection. In addition, PSO-KCAD, QLEPSO, and APLPSO are recently improved PSO algorithms originally proposed for general or application-oriented optimization problems. For the latter three algorithms, their core search and learning mechanisms are retained and adapted to binary feature selection under a unified evaluation framework.

- (1) *Graph-based Local Search Feature Selection (GLSFS)* [34]: GLSFS is a graph-driven feature selection method based on granular-rectangular neighborhood rough sets for interval-valued data. It models relationships among features through a weighted graph and evaluates their importance from a global structural perspective.
- (2) *k-Nearest Neighborhood Dependency-based Unsupervised Feature Selection (KND-UFS)* [35]: KND-UFS is an unsupervised feature selection method integrating Fuzzy C-Means clustering with k-nearest-neighbor rough sets. It evaluates feature importance according to clustering information and subsequently removes redundant features through neighborhood-based dependency evaluation.
- (3) *Fuzzy Neighborhood Generalized Composite Entropy (FNGCE)* [8]: FNGCE is constructed within a generalized multigranulation fuzzy neighborhood rough-set framework to quantify feature

uncertainty. A forward greedy search strategy is then employed to select informative features according to the proposed uncertainty measure.

- (4) *Queen-inspired Adaptive Ant Colony Optimization (QACO)* [36]: QACO is an adaptive multi-phase ant colony optimization approach developed for robust feature selection. Its adaptive population regulation and multi-phase search mechanisms enhance search diversity and reduce premature convergence.
- (5) *Particle Swarm Optimization based on K-means Clustering and Adaptive Dual-Group Strategy (PSO-KCAD)* [37]: PSO-KCAD combines K-means clustering with an adaptive dual-group mechanism to coordinate global exploration and local exploitation. Its original optimization mechanism is adapted to the binary feature selection problem in the present experiments.
- (6) *Q-learning Enhanced Particle Swarm Optimization (QLEPSO)* [38]: QLEPSO integrates a position-update strategy pool, Q-learning-based strategy selection, and adaptive parameter control into PSO. Since the original algorithm was developed for a combinatorial scheduling problem, its search mechanism is adapted to binary feature selection while preserving its original learning strategy.
- (7) *Adaptive Pattern Learning Particle Swarm Optimization (APLPSO)* [39]: APLPSO constructs multiple learning patterns with different numbers of exemplars and enables each particle to adaptively select a suitable learning pattern. Originally developed for large-scale optimization, APLPSO's learning framework is adapted to the binary feature selection task in this study.

For a fair comparison, all population-based optimization algorithms are evaluated under a unified experimental protocol. Specifically, the same data preprocessing procedure and five-fold stratified cross-validation partitions are used for all methods. KNN with $k = 3$ is employed as the search classifier, and the population size and maximum number of iterations are uniformly set to $N_p = 50$ and $T = 50$, respectively. Thus, the compared population-based algorithms adopt the same stopping criterion and a comparable evolutionary search budget. The random seed is fixed at 42 for reproducibility.

The fitness evaluation criterion is also kept consistent. Specifically, the accuracy weight in Eq. (36) is set to $\lambda = 0.9$, while the feature-number weight is $1 - \lambda = 0.1$. The candidate feature dimension is denoted by $m' = |A'|$. Except for algorithm-specific parameters required by their original search mechanisms, the above experimental settings are kept unchanged. Therefore, the comparative results mainly reflect differences among the search and learning mechanisms rather than differences in the evaluation protocol.

4.3. Ablation study

To investigate the contribution of the main components of DSLGPSO-GWO, a component-wise ablation study is conducted. Instead of cumulatively adding modules, each ablation variant removes or replaces only one major component while keeping the remaining framework unchanged, thereby allowing the effect of each mechanism to be evaluated more directly. Four ablation variants are considered.

(1) *w/o Stage-I*: The GMNRS-based first-stage granularity screening is removed, and the second-stage optimizer directly searches the original feature space. Since no first-stage granularity importance is available, the initialization retains the multiscale subset-size mechanism but does not use the granularity-derived prior information.

(2) *w/o AMI*: The first-stage screening is retained, whereas the adaptive multiscale initialization strategy is replaced by conventional random population initialization. This variant is used to evaluate whether the proposed initialization mechanism can provide more informative and diverse initial solutions.

(3) *w/o Layered Graph*: The layered graph learning mechanism is removed while the remaining components are retained. Therefore,

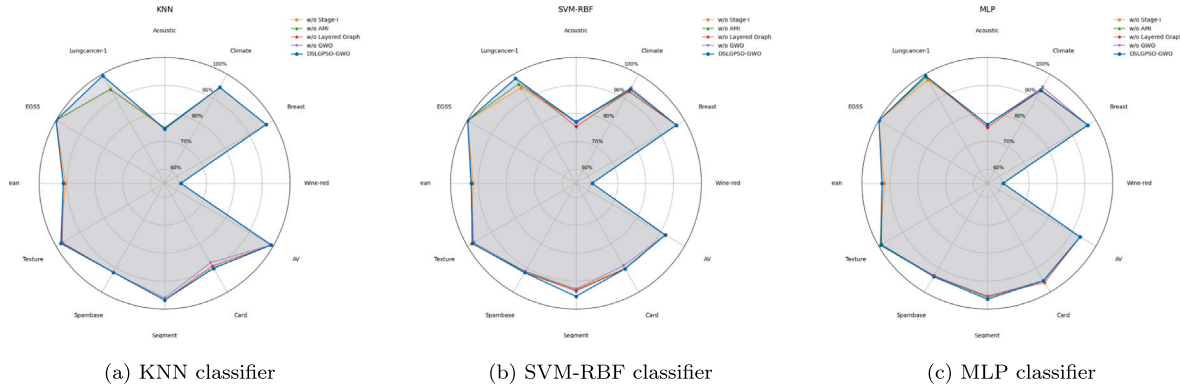


Fig. 3. Classification accuracy of different ablation variants under KNN, SVM-RBF, and MLP classifiers.

particles no longer obtain the structured intra-layer and upper-layer learning information introduced by the proposed graph-based population organization.

(4) *w/o GWO*: The three-leader GWO guidance term is removed from the hybrid update mechanism, while the layered graph PSO learning structure is preserved. This variant is used to examine the contribution of multi-leader global guidance.

Fig. 3 presents the classification accuracy of the four ablation variants and the complete DSLGPSO-GWO with KNN, SVM-RBF, and MLP classifiers.

As shown in Fig. 3, the complete DSLGPSO-GWO generally lies on or close to the outer envelope of the radar plots under the three classifiers, indicating that the complete combination of the four mechanisms provides the most stable overall performance. Removing Stage I causes performance degradation on several datasets, demonstrating that eliminating weak granularities before evolutionary search is beneficial not only for reducing the search space but also for preserving useful decision-related information. Replacing AMI with random initialization also leads to performance loss on some of the datasets, which indicates that jointly considering feature priority and subset-scale diversity can provide a better starting population.

The variants without layered graph learning or GWO guidance also show decreases on several datasets. This suggests that the two mechanisms play complementary roles: layered graph learning provides multiple structured exemplars for particles at different quality levels, whereas GWO introduces additional global guidance through multiple elite leaders. Nevertheless, the differences among the variants are relatively small on some datasets with already high classification accuracy, such as those approaching the performance ceiling. This phenomenon indicates that the contribution of an individual module is data-dependent rather than uniformly dominant. Overall, the consistent advantage of the complete model across different classifiers supports the effectiveness of integrating Stage-I screening, adaptive initialization, hierarchical graph learning, and GWO guidance within a unified framework.

4.4. Parameter sensitivity analysis

The granularity retention ratio ρ is an important parameter in Stage I because it directly controls the number of feature granularities transferred to the second-stage evolutionary search. A relatively small ρ produces stronger filtering and a more compact candidate space, whereas an excessively large ρ retains more granularities and weakens the dimensionality-reduction effect of Stage I. Therefore, the sensitivity of DSLGPSO-GWO to ρ is investigated over $\rho \in \{0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80\}$. The resulting classification accuracies with KNN, SVM-RBF, and MLP are shown in Fig. 4.

It can be observed from Fig. 4 that the influence of ρ varies across datasets. On several datasets, including AV, EGSS, and Lungcancer-1,

the classification accuracy remains nearly unchanged over a relatively wide range of ρ , indicating that the proposed framework is robust to moderate changes in the retention ratio. In contrast, datasets such as Acoustic, Wine-red, Card, and Texture exhibit more noticeable variations, suggesting that the amount of information retained after Stage I is more important for these datasets.

An important characteristic of the curves is the presence of multiple horizontal segments. According to Eq. (33), the number of retained granularities is determined by $K_R = \max\{1, \lceil \rho K^* \rceil\}$. Consequently, two adjacent values of ρ may correspond to the same K_R and therefore produce the same retained granularity family and candidate feature set. Under such conditions, identical or highly similar classification results are expected rather than indicating insensitivity of the optimization procedure itself.

When ρ is too small, aggressive granularity removal may discard useful discriminative information, which explains the relatively low accuracy observed for some datasets at $\rho = 0.50$. Increasing ρ generally alleviates this information loss, although retaining too many granularities also reduces the search-space compression achieved by Stage I and does not necessarily improve classification performance. Overall, the results indicate that the proposed framework maintains competitive performance over a relatively broad range of ρ . Therefore, $\rho = 0.70$ is used as a unified default setting in all comparison experiments. It should be noted that the parameters are not tuned separately for individual datasets. A unified parameter setting is adopted across all benchmark datasets, and the sensitivity analysis is mainly conducted to examine the robustness of DSLGPSO-GWO to parameter variations rather than to determine dataset-specific optimal parameters.

4.5. Comparison experiments

To evaluate the effectiveness and generalization capability of DSLGPSO-GWO, comparative experiments are conducted against the seven baseline algorithms introduced in Section 4.2. Classification performance is evaluated using KNN, SVM-RBF, and MLP classifiers. In addition to classification accuracy, the number of selected features is reported to assess the compactness of the obtained feature subsets. Therefore, the comparison considers both predictive performance and dimensionality reduction rather than relying on a single evaluation criterion.

Tables 2–4 report the classification accuracy obtained by the eight algorithms. Under the KNN classifier, DSLGPSO-GWO achieves the best or tied-best accuracy on eight of the twelve datasets. In particular, it shows clear competitive advantages on Acoustic, Breast, AV, Card, Spambase, EGSS, and Lungcancer-1. Although PSO-KCAD, QLEPSO, or APLPSO obtains slightly better results on Wine-red, Segment, Texture, and Bean, the proposed method remains close to the best result on these datasets.

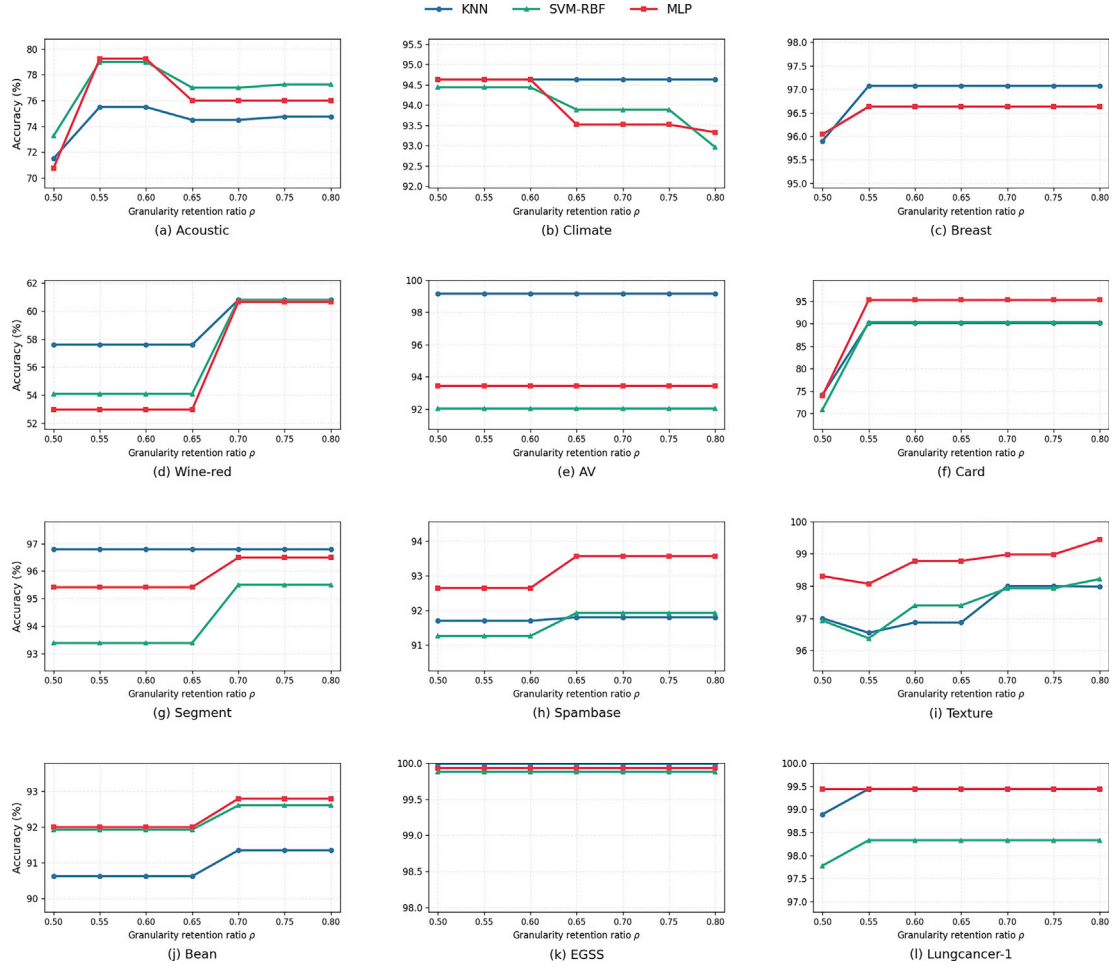
Fig. 4. Sensitivity analysis of the granularity retention ratio ρ under different classifiers.

Table 2

Classification accuracy of different algorithms under KNN classifier (%).

Datasets	GLSFS	KND-UFS	FNGCE	QACO	PSO-KCAD	QLEPSO	APLPSO	DSLGPSO-GWO
Acoustic	64.00 $\pm$ 2.15	67.50 $\pm$ 2.85	58.00 $\pm$ 6.64	67.50 $\pm$ 4.18	72.75 $\pm$ 1.66	70.50 $\pm$ 4.65	73.00 $\pm$ 4.15	74.50 $\pm$ 2.81
Climate	89.81 $\pm$ 2.55	91.26 $\pm$ 0.45	91.11 $\pm$ 2.46	91.85 $\pm$ 2.51	94.63 $\pm$ 2.22	91.11 $\pm$ 2.39	90.93 $\pm$ 2.58	94.63 $\pm$ 1.48
Breast	95.31 $\pm$ 1.58	96.33 $\pm$ 2.63	96.19 $\pm$ 2.85	95.17 $\pm$ 2.22	96.78 $\pm$ 1.18	96.78 $\pm$ 1.18	94.73 $\pm$ 1.42	97.07 $\pm$ 1.23
Wine-red	59.60 $\pm$ 2.73	49.84 $\pm$ 2.07	52.47 $\pm$ 2.74	58.85 $\pm$ 1.94	61.54 $\pm$ 1.55	61.54 $\pm$ 1.55	59.04 $\pm$ 2.82	60.79 $\pm$ 2.29
AV	91.02 $\pm$ 0.68	94.86 $\pm$ 0.89	85.27 $\pm$ 0.57	96.57 $\pm$ 0.27	98.48 $\pm$ 0.84	98.48 $\pm$ 0.84	94.47 $\pm$ 1.32	99.17 $\pm$ 0.25
Card	80.72 $\pm$ 1.48	52.35 $\pm$ 1.96	79.22 $\pm$ 1.43	81.24 $\pm$ 2.68	89.51 $\pm$ 1.11	89.60 $\pm$ 1.09	89.46 $\pm$ 1.64	90.12 $\pm$ 1.01
Segment	91.60 $\pm$ 0.98	95.23 $\pm$ 0.47	94.03 $\pm$ 0.81	93.98 $\pm$ 0.84	96.62 $\pm$ 0.45	96.10 $\pm$ 0.61	96.84 $\pm$ 0.59	96.80 $\pm$ 0.63
Spambase	88.50 $\pm$ 1.25	88.79 $\pm$ 0.94	90.41 $\pm$ 1.14	90.42 $\pm$ 0.56	91.52 $\pm$ 0.51	91.09 $\pm$ 0.48	89.70 $\pm$ 0.33	91.80 $\pm$ 0.67
Texture	97.15 $\pm$ 0.38	97.73 $\pm$ 0.30	96.89 $\pm$ 3.12	97.64 $\pm$ 0.29	98.44 $\pm$ 0.40	97.96 $\pm$ 0.29	98.11 $\pm$ 0.20	98.00 $\pm$ 0.43
Bean	91.30 $\pm$ 0.20	86.33 $\pm$ 0.23	86.26 $\pm$ 1.52	87.33 $\pm$ 0.34	91.68 $\pm$ 0.44	91.46 $\pm$ 0.21	90.71 $\pm$ 0.38	91.35 $\pm$ 0.40
EGSS	–	93.88 $\pm$ 0.38	90.19 $\pm$ 0.93	97.39 $\pm$ 0.25	99.96 $\pm$ 0.04	99.96 $\pm$ 0.04	99.96 $\pm$ 0.04	99.99 $\pm$ 0.02
Lungcancer-1	–	98.33 $\pm$ 1.36	98.36 $\pm$ 2.50	96.11 $\pm$ 3.33	96.11 $\pm$ 3.33	97.78 $\pm$ 2.08	95.56 $\pm$ 4.16	99.44 $\pm$ 1.11

Table 3

Classification accuracy of different algorithms under SVM-RBF classifier (%).

Datasets	GLSFS	KND-UFS	FNGCE	QACO	PSO-KCAD	QLEPSO	APLPSO	DSLGPSO-GWO
Acoustic	73.75 $\pm$ 5.65	75.25 $\pm$ 5.02	73.50 $\pm$ 5.67	73.50 $\pm$ 4.64	76.75 $\pm$ 3.76	74.75 $\pm$ 5.56	78.25 $\pm$ 3.32	77.00 $\pm$ 3.12
Climate	91.48 $\pm$ 2.58	90.26 $\pm$ 1.37	91.48 $\pm$ 2.58	92.22 $\pm$ 2.24	93.70 $\pm$ 1.80	91.48 $\pm$ 2.58	91.48 $\pm$ 2.58	93.89 $\pm$ 1.50
Breast	95.32 $\pm$ 1.64	96.63 $\pm$ 1.51	96.63 $\pm$ 2.35	96.05 $\pm$ 1.78	96.20 $\pm$ 0.96	96.20 $\pm$ 0.96	94.88 $\pm$ 1.85	96.63 $\pm$ 1.20
Wine-red	60.36 $\pm$ 4.67	55.60 $\pm$ 3.30	55.79 $\pm$ 3.02	58.85 $\pm$ 3.76	61.42 $\pm$ 4.94	61.42 $\pm$ 4.94	60.29 $\pm$ 4.53	60.79 $\pm$ 2.17
AV	88.45 $\pm$ 1.09	91.68 $\pm$ 0.61	87.18 $\pm$ 0.84	91.39 $\pm$ 0.39	90.31 $\pm$ 0.66	90.31 $\pm$ 0.66	91.68 $\pm$ 0.61	92.02 $\pm$ 1.05
Card	86.83 $\pm$ 1.79	56.61 $\pm$ 2.03	84.39 $\pm$ 1.39	83.26 $\pm$ 2.63	85.00 $\pm$ 1.22	88.57 $\pm$ 1.35	87.54 $\pm$ 1.40	90.31 $\pm$ 1.27
Segment	86.28 $\pm$ 1.26	93.11 $\pm$ 1.90	93.05 $\pm$ 1.18	91.21 $\pm$ 0.89	95.19 $\pm$ 0.50	90.43 $\pm$ 1.96	93.29 $\pm$ 0.84	95.50 $\pm$ 0.84
Spambase	92.02 $\pm$ 0.49	89.79 $\pm$ 0.94	88.62 $\pm$ 0.14	90.28 $\pm$ 0.53	91.15 $\pm$ 0.46	90.96 $\pm$ 0.26	88.72 $\pm$ 0.72	91.93 $\pm$ 0.77
Texture	97.02 $\pm$ 0.38	97.55 $\pm$ 0.46	94.78 $\pm$ 4.37	95.64 $\pm$ 0.97	98.56 $\pm$ 0.56	98.42 $\pm$ 0.23	98.33 $\pm$ 0.29	97.93 $\pm$ 0.55
Bean	92.42 $\pm$ 0.35	88.24 $\pm$ 0.49	86.73 $\pm$ 1.20	90.07 $\pm$ 0.22	92.44 $\pm$ 0.64	92.32 $\pm$ 0.51	91.72 $\pm$ 0.19	92.61 $\pm$ 0.56
EGSS	–	98.80 $\pm$ 0.18	97.60 $\pm$ 0.56	99.39 $\pm$ 0.11	99.90 $\pm$ 0.09	99.90 $\pm$ 0.09	99.90 $\pm$ 0.09	99.88 $\pm$ 0.07
Lungcancer-1	–	97.78 $\pm$ 1.11	97.45 $\pm$ 0.32	97.78 $\pm$ 2.08	97.78 $\pm$ 2.08	97.22 $\pm$ 2.48	97.22 $\pm$ 2.48	98.33 $\pm$ 1.36

Table 4
Classification accuracy of different algorithms under MLP classifier (%).

Datasets	GLSFS	KND-UFS	FNGCE	QACO	PSO-KCAD	QLEPSO	APLPSO	DSLGPSO-GWO
Acoustic	74.50 ± 3.32	78.50 ± 4.14	75.50 ± 4.91	76.50 ± 6.63	74.75 ± 3.66	73.25 ± 2.32	75.75 ± 1.27	76.00 ± 2.67
Climate	91.48 ± 2.58	90.26 ± 1.37	91.48 ± 2.58	92.22 ± 2.24	93.52 ± 2.03	91.48 ± 2.58	91.48 ± 2.58	93.52 ± 1.31
Breast	96.05 ± 1.10	96.63 ± 1.51	96.63 ± 2.35	96.05 ± 1.78	96.20 ± 1.25	96.20 ± 1.25	94.88 ± 2.12	96.63 ± 1.52
Wine-red	60.86 ± 5.02	55.66 ± 3.24	55.04 ± 2.97	58.73 ± 3.47	60.92 ± 5.83	60.92 ± 5.83	59.86 ± 4.93	60.66 ± 1.66
AV	89.48 ± 1.13	91.68 ± 0.65	86.98 ± 0.62	91.04 ± 1.30	91.73 ± 0.53	91.73 ± 0.53	91.78 ± 0.71	93.44 ± 0.74
Card	86.50 ± 1.04	56.61 ± 2.03	84.39 ± 1.39	83.26 ± 2.63	93.98 ± 1.38	93.37 ± 1.63	91.25 ± 3.04	95.25 ± 0.70
Segment	86.28 ± 1.26	93.11 ± 1.90	93.05 ± 1.18	91.21 ± 0.89	95.67 ± 0.85	95.06 ± 0.86	95.84 ± 0.46	96.49 ± 0.60
Spambase	93.63 ± 0.67	91.76 ± 0.27	91.66 ± 0.81	91.04 ± 1.30	92.13 ± 0.51	92.31 ± 0.76	91.02 ± 0.68	93.57 ± 0.53
Texture	97.02 ± 0.38	97.55 ± 0.46	94.78 ± 4.37	95.64 ± 0.97	98.58 ± 0.36	98.58 ± 0.35	98.09 ± 0.32	98.98 ± 0.20
Bean	92.42 ± 0.35	88.24 ± 0.49	86.73 ± 1.20	90.07 ± 0.22	91.92 ± 0.23	91.96 ± 0.37	91.10 ± 0.29	92.80 ± 0.64
EGSS	–	99.79 ± 0.05	98.34 ± 1.16	99.72 ± 0.13	99.85 ± 0.13	99.85 ± 0.13	99.85 ± 0.13	99.93 ± 0.02
Lungcancer-1	–	98.89 ± 1.36	98.55 ± 1.37	99.44 ± 1.11	98.89 ± 1.36	98.89 ± 2.22	99.44 ± 1.11	99.44 ± 1.11

Table 5
Number of selected features by different algorithms.

Datasets	GLSFS	KND-UFS	FNGCE	QACO	PSO-KCAD	QLEPSO	APLPSO	DSLGPSO-GWO
Acoustic	32	14	26	17	20	22	15	14
Climate	11	8	11	6	4	2	1	4
Breast	6	5	8	3	3	3	2	3
Wine-red	7	1	6	4	5	5	3	5
AV	4	3	6	3	4	4	3	4
Card	13	1	18	7	4	3	5	4
Segment	12	6	10	6	4	5	3	4
Spambase	37	22	45	22	23	18	14	18
Texture	26	20	21	13	11	9	9	10
Bean	10	2	4	4	4	4	3	4
Average	15.8	8.2	15.5	8.5	8.2	7.5	5.8	7.0
EGSS	–	6	8	4	1	1	1	1
Lungcancer-1	–	6	5164	4978	6184	867	6218	512

Under the SVM-RBF classifier, DSLGPSO–GWO obtains the best or tied-best performance on seven datasets, including Climate, Breast, AV, Card, Segment, Bean, and Lungcancer-1. Its strong performance under a nonlinear kernel classifier indicates that the selected feature subsets are not specifically tailored to the local distance mechanism of KNN. On several datasets, such as Acoustic, Wine-red, Spambase, Texture, and EGSS, other algorithms achieve slightly higher accuracies, showing that no single feature selection method dominates every data distribution.

The advantage of DSLGPSO–GWO becomes particularly evident under the MLP classifier, where it achieves the best or tied-best accuracy on nine of the twelve datasets. Notable results are observed on AV, Card, Segment, Texture, Bean, EGSS, and Lungcancer-1. The consistent performance across KNN, SVM-RBF, and MLP indicates that the selected subsets retain discriminative information that can be effectively utilized by classifiers with substantially different learning mechanisms. Overall, DSLGPSO–GWO obtains the best or tied-best result in 24 of the 36 dataset–classifier combinations, demonstrating strong cross-classifier robustness without claiming uniform superiority on every individual task.

Table 5 further reports the number of selected features. Since GLSFS does not provide valid results for EGSS and Lungcancer-1, the average number of selected features is calculated over the ten datasets for which all algorithms have complete results. DSLGPSO–GWO retains an average of 7.0 features, which is the second smallest value among the compared algorithms, following APLPSO with 5.8 features. This indicates that the proposed method can substantially reduce feature dimensionality while avoiding excessively aggressive feature elimination.

It should be emphasized that selecting the fewest features does not necessarily produce the highest classification accuracy. For example, some competing methods retain extremely compact subsets but exhibit lower classification performance. On the high-dimensional Lungcancer-1 dataset, which contains 12,533 original features, DSLGPSO–GWO retains 512 features. Although this number is not the smallest among

all methods, the corresponding classification accuracy remains superior or competitive across all three classifiers. Similarly, on EGSS, DSLGPSO–GWO retains only one feature while maintaining nearly perfect classification performance. These results demonstrate that DSLGPSO–GWO achieves a favorable balance between feature subset compactness and discriminative capability rather than pursuing dimensionality reduction alone.

The overall advantage can be attributed to the complementary roles of the two stages. Stage I eliminates weak feature granularities according to classifier-independent GMNRS dependency information, thereby reducing the search space while preserving decision-relevant structures. Stage II subsequently performs a finer wrapper search within the retained candidate space, allowing interactions among individual features to be further optimized. As a result, DSLGPSO–GWO combines the efficiency of granularity-level filtering with the predictive capability of wrapper-based evolutionary search.

4.6. Convergence analysis

To further investigate the search behavior of the proposed evolutionary optimizer, DSLGPSO–GWO is compared with PSO-KCAD, QLEPSO, and APLPSO in terms of convergence performance. Since the purpose of this experiment is to compare the optimization mechanisms rather than the first-stage filtering capability, the first-stage screening result is fixed for each dataset, and the compared optimizers search within the same candidate feature space. The horizontal axis of Fig. 5 denotes the number of iterations, while the vertical axis represents the best fitness value obtained so far. Since the fitness function is formulated as a minimization objective, a lower value indicates a better feature subset.

As illustrated in Fig. 5, DSLGPSO–GWO exhibits competitive convergence behavior on the twelve datasets. On Acoustic, Breast, AV, Card, Segment, EGSS, and LungCancer-1, the proposed method eventually reaches the lowest or nearly lowest fitness level among the compared

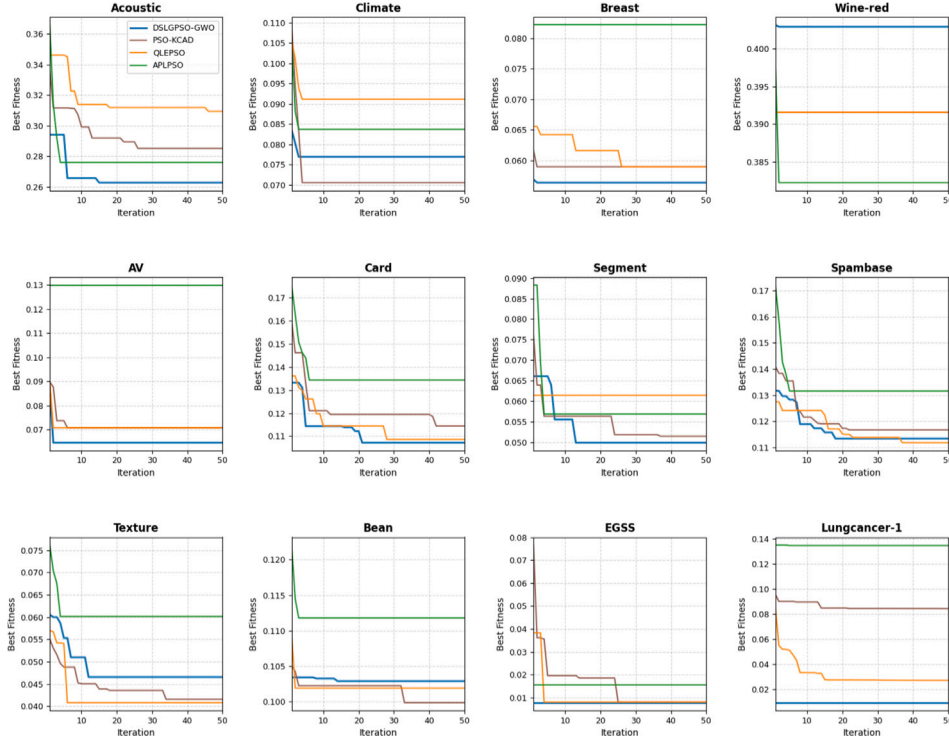


Fig. 5. Convergence curves of different optimization algorithms on twelve datasets.

algorithms. In particular, relatively rapid fitness reduction can be observed during the early search stage on several datasets, indicating that the proposed initialization and multi-source learning mechanisms can guide the population toward promising regions within a limited number of iterations.

The convergence advantage, however, is not uniform across all datasets. On several datasets, including Climate, Wine-red, Spambase, Texture, and Bean, one of the comparison algorithms reaches a lower final fitness value or exhibits a similar convergence trajectory. This observation is consistent with the classification experiments and suggests that different search mechanisms may have different advantages depending on the feature-space structure. Nevertheless, DSLGPSO-GWO maintains competitive final fitness across most datasets and avoids severe stagnation in the majority of cases.

The observed convergence behavior can be explained by the interaction of several mechanisms. The adaptive multiscale initialization provides diverse candidate subsets at the beginning of the search. During evolution, intra-layer competition and upper-layer learning introduce multiple structurally different exemplars, reducing excessive dependence on a single global optimum. Meanwhile, the three-leader GWO guidance supplies complementary global search directions. Their combination enables the population to exploit high-quality regions while retaining sufficient exploration capability, which contributes to the overall convergence stability of DSLGPSO-GWO.

4.7. Statistical testing

Although the comparative experiments demonstrate the competitive performance of DSLGPSO-GWO across different classifiers, differences in classification accuracy alone are insufficient to determine whether the observed performance differences are statistically reliable. Therefore, the Friedman test and the Nemenyi post-hoc test are further employed to provide a non-parametric statistical comparison among the considered algorithms.

It should be noted that GLSFS does not provide valid experimental results on the EGSS and Lungcancer-1 datasets. To ensure a complete-block comparison in which all algorithms are evaluated on exactly the same datasets, these two datasets are excluded from the statistical analysis for all methods. Consequently, the Friedman and Nemenyi tests are conducted on the remaining ten common datasets. For each dataset, the algorithm with the highest classification accuracy is assigned rank 1, while the remaining algorithms are ranked in descending order of classification accuracy. When multiple algorithms obtain the same accuracy, their corresponding ranks are averaged.

Table 6 presents the average ranks of the eight algorithms, together with the Friedman statistics, corresponding p -values, and critical difference values under the KNN, SVM-RBF, and MLP classifiers.

As shown in Table 6, DSLGPSO-GWO achieves the best average rank under all three classifiers, with average ranks of 1.75, 1.80, and 1.75 under KNN, SVM-RBF, and MLP, respectively. Under the KNN classifier, the Friedman statistic is 45.16 with a p -value of 1.28×10^{-7} . Under SVM-RBF, the Friedman statistic is 28.61 with a p -value of 1.70×10^{-4} , while under MLP, the corresponding values are 26.01 and 5.01×10^{-4} , respectively. Since all three p -values are considerably smaller than the significance level of 0.05, the null hypothesis that all algorithms exhibit equivalent classification performance is rejected. These results confirm that statistically significant overall differences exist among the compared algorithms under all three classifiers.

Since the Friedman test only determines whether an overall difference exists among the algorithms, the Nemenyi post-hoc test is subsequently employed to identify significant pairwise differences. The critical difference is calculated as

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N_d}}, \quad (69)$$

where $k = 8$ denotes the number of compared algorithms and $N_d = 10$ denotes the number of common datasets included in the statistical

Table 6
Friedman and Nemenyi statistical test results under different classifiers.

	DSLGPSPSO-GWO	GLSFS	KND-UFS	FNGCE	QACO	PSO-KCAD	QLEPSO	APLPSO	χ^2	p -Value	CD
KNN	1.75	6.50	5.85	6.75	5.35	2.10	3.10	4.60	45.16	1.28×10^{-7}	3.32
SVM-RBF	1.80	5.15	5.45	6.50	5.75	2.95	4.10	4.30	28.61	1.70×10^{-4}	3.32
MLP	1.75	5.10	5.30	6.25	5.75	3.25	3.85	4.75	26.01	5.01×10^{-4}	3.32

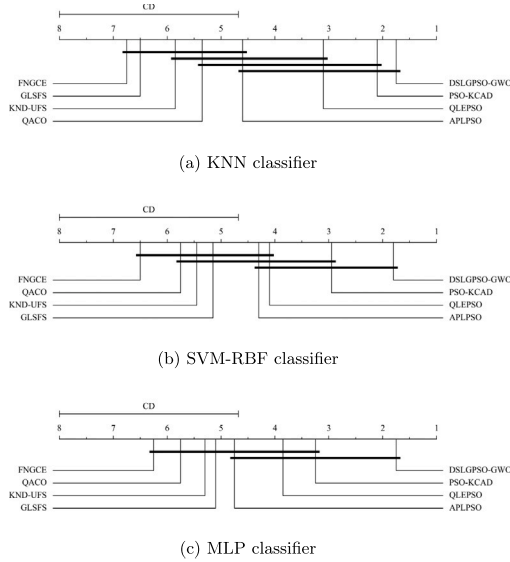


Fig. 6. Nemenyi critical difference diagrams under KNN, SVM-RBF, and MLP classifiers.

analysis. At the significance level $\alpha = 0.05$, the resulting critical difference is $CD = 3.32$. Therefore, two algorithms are regarded as significantly different when the absolute difference between their average ranks exceeds 3.32.

The corresponding Nemenyi critical difference diagrams for the three classifiers are shown in Fig. 6.

As illustrated in Fig. 6, DSLGPSPSO-GWO is consistently located at the best-ranking side of the critical difference diagrams under the three classifiers. According to the Nemenyi criterion, the differences between DSLGPSPSO-GWO and GLSFS, KND-UFS, FNGCE, and QACO exceed the critical difference of 3.32 under KNN, SVM-RBF, and MLP, indicating statistically significant differences between the proposed method and these algorithms. In contrast, the rank differences between DSLGPSPSO-GWO and PSO-KCAD, QLEPSO, and APLPSO do not exceed the critical difference, and therefore their pairwise differences cannot be regarded as statistically significant at the 0.05 significance level.

Overall, the statistical results are consistent with the comparative classification results. DSLGPSPSO-GWO achieves the best average rank under all three classifiers, while the Friedman test confirms significant overall performance differences among the compared algorithms. Furthermore, the Nemenyi post-hoc analysis demonstrates that the proposed method has statistically significant advantages over several representative baselines, while remaining statistically comparable to the strongest recent PSO-based competitors. These results provide further statistical evidence for the effectiveness and robustness of DSLGPSPSO-GWO across different classification models.

5. Conclusion and outlook

This section summarizes the main contributions and experimental findings of the proposed DSLGPSPSO-GWO framework and briefly discusses several directions for its further extension.

5.1. Conclusion

This paper proposes a dual-stage feature selection framework, termed DSLGPSPSO-GWO, for high-dimensional data. In Stage I, adaptive feature granularities are constructed using K-means, and their decision-supporting capabilities are evaluated through GMNRS-based external importance. Highly ranked granularities are then retained to generate a reduced candidate feature space. In Stage II, a granularity-feature hierarchical prior and an adaptive multiscale initialization strategy are introduced to provide informative and diverse initial solutions. Furthermore, layered graph learning, upper-layer guidance, and the three-leader GWO mechanism are integrated into the particle update process to provide multiple complementary learning directions, while elite preservation and diversity-oriented perturbation help maintain search stability. Through the cooperation of granularity-level screening and fine-grained feature subset optimization, the proposed framework reduces the wrapper search space while preserving discriminative feature information. Experimental results on twelve benchmark datasets show that DSLGPSPSO-GWO achieves competitive classification performance and feature reduction capability with KNN, SVM-RBF, and MLP classifiers. Ablation, convergence, and statistical analyses further support the effectiveness and robustness of the proposed framework.

5.2. Limitations and future work

The proposed framework still provides several opportunities for further extension. First, some parameters, such as the neighborhood radius δ , the generalized multigranulation parameter β_R , and the granularity retention ratio ρ , are currently predefined. Future studies may investigate data-driven or adaptive parameter adjustment strategies to further improve the flexibility of DSLGPSPSO-GWO across different data distributions. Second, more efficient neighborhood computation and incremental updating mechanisms can be explored to enhance the scalability of the framework for larger-scale data. In addition, the current method can be further extended from static supervised classification to dynamic or multi-source information systems, where feature structures and granulation patterns may change over time or differ across data sources. Finally, multi-objective optimization may be incorporated to provide a more flexible balance among classification performance, feature subset size, and other selection criteria, thereby further extending the applicability of the proposed framework.

CRediT authorship contribution statement

Qinglong Feng: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Weihua Xu:** Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62376229 and in part by the Natural Science Foundation of Chongqing under Grant CSTB2023NSCQ-LZX0027.

Data availability

The link to the data used for the research described has been shared in the article.

References

- [1] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, *ACM Comput. Surv. (CSUR)* 50 (2017) 1–45.
- [2] A. Moslemi, A tutorial-based survey on feature selection: recent advancements on feature selection, *Eng. Appl. Artif. Intell.* 126 (2023) 107136.
- [3] Z. Pawlak, Rough sets, *Int. J. Comput. & Inf. Sci.* 11 (1982) 341–356.
- [4] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, *Inf. Sci.* 178 (2008) 3577–3594.
- [5] Y. Qian, J. Liang, Y. Yao, C. Dang, MGRS: a multi-granulation rough set, *Inf. Sci.* 180 (2010) 949–970.
- [6] L. Sun, L. Wang, W. Ding, Y. Qian, J. Xu, Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets, *IEEE Trans. Fuzzy Syst.* 29 (2020) 19–33.
- [7] W. Xu, K. Yuan, W. Li, Dynamic updating approximations of local generalized multigranulation neighborhood rough set, *Appl. Intell.* 52 (2022) 9148–9173.
- [8] X. Zhang, W. Zhao, Uncertainty measures and feature selection based on composite entropy for generalized multigranulation fuzzy neighborhood rough set, *Fuzzy Sets Syst.* 486 (2024) 108971.
- [9] W. Xu, Q. Bu, Matrix-based incremental feature selection method using weight-partitioned multigranulation rough set, *Inf. Sci.* 681 (2024) 121219.
- [10] C. Wang, Y. Zhang, S. An, T. Deng, Adaptive feature selection based on fuzzy rough set fusion model with class variance, *Pattern Recognit.* 170 (2026) 112014.
- [11] J. Gao, Z. Wang, T. Jin, J. Cheng, Z. Lei, S. Gao, Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection, *Knowl.-Based Syst.* 286 (2024) 111380.
- [12] Y. Xue, C. Zhang, A novel importance-guided particle swarm optimization based on MLP for solving large-scale feature selection problems, *Swarm Evol. Comput.* 91 (2024) 101760.
- [13] L. Qu, W. He, J. Li, H. Zhang, C. Yang, B. Xie, Explicit and size-adaptive PSO-based feature selection for classification, *Swarm Evol. Comput.* 77 (2023) 101249.
- [14] J. He, L. Qu, P. Wang, Z. Li, An oscillatory particle swarm optimization feature selection algorithm for hybrid data based on mutual information entropy, *Appl. Soft Comput.* 152 (2024) 111261.
- [15] X. Jin, B. Wei, L. Deng, S. Yang, J. Zheng, F. Wang, An adaptive pyramid PSO for high-dimensional feature selection, *Expert Syst. Appl.* 257 (2024) 125084.
- [16] S. Yang, B. Wei, L. Deng, X. Jin, M. Jiang, Y. Huang, F. Wang, A leader-adaptive particle swarm optimization with dimensionality reduction strategy for feature selection, *Swarm Evol. Comput.* 91 (2024) 101743.
- [17] B. Wei, S. Yang, W. Zha, L. Deng, J. Huang, X. Su, F. Wang, Particle swarm optimization algorithm based on comprehensive scoring framework for high-dimensional feature selection, *Swarm Evol. Comput.* 95 (2025) 101915.
- [18] B. Wei, J. Huang, L. Deng, S. Yang, J. Zheng, Y. Huang, Reinforcement learning-based particle swarm optimization with adaptive scoring mechanism for high-dimensional feature selection, *Swarm Evol. Comput.* 98 (2025) 102104.
- [19] L. Zhang, Adaptive diversity-driven multi-strategy particle swarm optimization for global optimization and feature selection, *Appl. Soft Comput.* (2026) 115525.
- [20] T. Thaher, H. Chantar, J. Too, M. Mafarja, H. Turabieh, E.H. Houssein, Boolean particle swarm optimization with various evolutionary population dynamics approaches for feature selection problems, *Expert Syst. Appl.* 195 (2022) 116550.
- [21] J. Osei-Kwakye, F. Han, A.A. Amponsah, Q.-H. Ling, T.A. Abeo, A diversity enhanced hybrid particle swarm optimization and crow search algorithm for feature selection, *Appl. Intell.* 53 (2023) 20535–20560.
- [22] B. Zhou, J. Wang, P. Feng, X. Zhang, D. Yu, J. Zhang, A genetic particle swarm optimization algorithm for feature fusion and hyperparameter optimization for tool wear monitoring, *Expert Syst. Appl.* 285 (2025) 127975.
- [23] P. Dhal, C. Azad, A multi-objective feature selection method using Newton's law based PSO with GWO, *Appl. Soft Comput.* 107 (2021) 107394.
- [24] L. Sun, X. Wang, W. Ding, J. Xu, TSFNFR: two-stage fuzzy neighborhood-based feature reduction with binary whale optimization algorithm for imbalanced data classification, *Knowl.-Based Syst.* 256 (2022) 109849.
- [25] L. Sun, S. Si, W. Ding, X. Wang, J. Xu, TFSFB: two-stage feature selection via fusing fuzzy multi-neighborhood rough set with binary whale optimization for imbalanced data, *Inf. Fusion* 95 (2023) 91–108.
- [26] X. Guo, J. Hu, H. Yu, M. Wang, B. Yang, A new population initialization of meta-heuristic algorithms based on hybrid fuzzy rough set for high-dimensional gene data feature selection, *Comput. Biol. Med.* 166 (2023) 107538.
- [27] Z. Li, X. Cai, Q. Huang, Y. Lin, A novel cooperative swarm intelligence feature selection method for hybrid data based on fuzzy β covering and fuzzy self-information, *Inf. Sci.* 675 (2024) 120757.
- [28] Y. Chen, X. Cai, Z. Li, An ant colony optimization attribute reduction algorithm for hybrid data using fuzzy β covering and fuzzy mutual information, *Appl. Soft Comput.* 167 (2024) 112373.
- [29] J. Zhang, Z. Zhao, Opposition-based particle swarm optimization algorithm integrating neighborhood rough set for feature selection in heterogeneous information system, *J. Supercomput.* 81 (2025) 866.
- [30] T. Xie, W. Ding, H. Ju, J. Huang, Y. Chen, An innovative feature selection approach with two-stage fusion using fuzzy neighborhood strategies, *Appl. Soft Comput.* (2026) 114871.
- [31] J. Kennedy, R. Eberhart, Particle swarm optimization (PSO), in: *Proc. IEEE International Conference on Neural Networks*, Perth, Australia, vol. 4, 1995, pp. 1942–1948.
- [32] S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey wolf optimizer, *Adv. Eng. Softw.* 69 (2014) 46–61.
- [33] M. Othman, K.R. Ku-Mahamud, Adaptive binary grey wolf optimization and particle swarm optimization algorithm for heart disease classification, in: *AIP Conference Proceedings*, vol. 3407, AIP Publishing LLC, 2026, p. 020005.
- [34] X. Zhang, X. Shen, Graph-driven feature selection via granular-rectangular neighborhood rough sets for interval-valued data sets, *Appl. Soft Comput.* 170 (2025) 112716.
- [35] W. Xu, Y. Zhang, Y. Qian, A novel unsupervised feature selection for high-dimensional data based on FCM and k -nearest neighbor rough sets, *IEEE Trans. Neural Netw. Learn. Syst.* 36 (2025) 10889–10998.
- [36] Z. Wang, Y. Liang, W. Sun, J. Liu, QACO: an adaptive multi-phase ant colony optimization approach for robust feature selection, *Swarm Evol. Comput.* 100 (2026) 102217.
- [37] Y. Fan, D. Tian, Q. Xu, J. Sun, Q. Xu, Z. Shi, Particle swarm optimization based on K-means clustering and adaptive dual-groups strategy, *Swarm Evol. Comput.* 100 (2026) 102226.
- [38] X.-Y. Li, Z.-Y. Yang, M.-W. Li, W.-C. Hong, Integrated scheduling of cargo vessels, research vessels, and marine experiments in multifunctional ports using Q-learning enhanced PSO, *Swarm Evol. Comput.* 102 (2026) 102315.
- [39] Z.-T. Lai, Z.-J. Wang, S. Liu, Z.-G. Chen, Z.-H. Zhan, S. Kwong, J. Zhang, Adaptive pattern learning particle swarm optimization for large-scale optimization, *Swarm Evol. Comput.* 101 (2026) 102268.