Expert Systems With Applications 332 (2027) 133647

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multi-objective feature selection in high-dimensional classification: An importance-guided and historical information-assisted approach

Kunjie Yu [a], Wen An [a], Ke Chen [a,*], Fangfang Zhang [b]

[a] School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, 450001, China
[b] School of Engineering and Computer Science, Victoria University of Wellington, Wellington, 6140, New Zealand

ARTICLE INFO

Keywords:
High-dimensional classification
Multi-objective optimization
Feature selection

ABSTRACT

Feature selection is a crucial preprocessing step for classification, inherently forming a bi-objective optimization problem aimed at minimizing feature subset size and classification error. However, most existing multi-objective feature selection methods fail to adequately account for potential feature interactions during the population initialization and underutilize historical information during the evolutionary process. Such issues may affect the breadth and depth of population search within high-dimensional feature spaces. To address these limitations, this paper proposes a multi-objective feature selection approach based on the importance-guided and historical information-assisted strategies (MOFS-IH). The main characteristics are twofold. First, the importance-guided population initialization strategy dynamically stratifies the feature space via information gain, thereby generating a well-distributed initial population that better captures feature interactions and facilitates rapid convergence. Second, the historical information-assisted population update mechanism leverages recorded feature-level information to guide the search toward promising feature subsets, thereby enabling more efficient utilization of historical evolutionary information. Extensive experiments on 15 benchmark datasets demonstrate that MOFS-IH outperforms six state-of-the-art multi-objective feature selection algorithms in terms of the quality of the obtained feature subsets.

1. Introduction

Classification is one of the central tasks in data analysis and pattern recognition, aiming to predict class labels of unseen instances using feature information (Chen et al., 2026a). However, with the quick development of sampling and information-processing technologies, the data volume in various real-world scenarios is increasing rapidly (Li et al., 2026a). Massive datasets of this kind tend to induce the curse of dimensionality, thereby compromising algorithmic performance. More critically, the majority of features in these datasets exhibit redundancy, irrelevance, or noise, which significantly impairs the accuracy of classification (Xia et al., 2026). As a data preprocessing and dimensionality reduction approach, feature selection has attracted wide attention for its ability to raise the quality of feature sets by eliminating redundant and irrelevant features and identifying valuable combinations among diverse features (Li et al., 2026b).

Generally, feature selection approaches can be categorized into filter-based, wrapper-based, and embedded-based according to different evaluation criteria (Li & Chai, 2024). Filter-based methods use various evaluation criteria to rank features and independently select the top ones, thus forming a selected feature subset (Wang et al., 2025), while wrapper-based methods utilize a learner to evaluate candidate feature subsets through diverse search methods (Yu et al., 2025). These operational differences lead to clear strengths and weaknesses. Filter-based approaches are computationally inexpensive but may ignore feature correlations. By contrast, wrapper-based methods often achieve higher accuracy but incur higher computational costs and a greater risk of overfitting (Li & Dang, 2026). Embedded-based methods, by contrast, utilize the strengths of both by integrating feature selection into model training, thus retaining the computational efficiency of filter methods while achieving adaptability to the learning model used in wrapper methods (Chen et al., 2026b). Nevertheless, embedded methods are highly sensitive to the choice and configuration of their base models, which limits their generality across different learning frameworks and data settings (Chandrashekar & Sahin, 2014).

Evolutionary computation (EC) is renowned for its global search ability and is widely applied to address multi-objective feature selection problems (Huang et al., 2025a; Li et al., 2016; Singh & Singh, 2017; Siswantoro et al., 2026; Xue et al., 2023). For instance, Xue et al. (2023) proposed a feature selection method based on the Non-dominated

* Corresponding author.
E-mail addresses: yukunjie@zzu.edu.cn (K. Yu), anwen_auto@163.com (W. An), chenkezixf@zzu.edu.cn (K. Chen), fangfang.zhang@ecs.vuw.ac.nz (F. Zhang).

https://doi.org/10.1016/j.eswa.2026.133647
Received 19 April 2026; Received in revised form 28 June 2026; Accepted 11 July 2026
Available online 21 July 2026

Sorting Genetic Algorithm II (NSGA-II). This method utilized the ReliefF algorithm to generate feature scores for population initialization and incorporated a crossover-mutation operator to guide the search direction, thereby enhancing convergence performance. However, by focusing solely on high-scoring features during initialization and ignoring the effectiveness of feature combinations, this method may easily lead the population to converge to local optima. In another work, Li et al. (2016) proposed a two-stage algorithm based on the Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D). The first stage employed freezing and activation operators for rapid dimensionality reduction, while the second stage refined the reduced feature space through redundancy removal and a preference-aware mechanism that prioritizes classification accuracy. Nevertheless, the single-feature screening logic in the dimensionality reduction phase overlooks feature interactions, which may hinder the identification of a better Pareto front. In fact, insufficient exploration of feature interaction information during the population initialization stage often causes most existing evolutionary algorithms to converge to local optima prematurely. Therefore, exploring approaches to enhance feature interactions at the initialization stage deserves further investigation.

Furthermore, feature selection is a combinatorial optimization problem featuring a solution space with significant complexity and discontinuity (Jiao et al., 2023). This characteristic frequently leads EC to overlook or fail to consistently retain effective feature subsets during iterative search, particularly in the case of scarce training samples or complex feature interactions (Azhagusundari et al., 2013). A promising strategy to alleviate this issue is to track, evaluate, and reuse the knowledge explored throughout the iterative process. This can integrate global information to compensate for deficiencies in local judgments, thus serving as a key mechanism for improving search stability (Chen et al., 2019; Wang et al., 2024, 2022). For example, Wang et al. (2024) quantify and dynamically update feature weights based on changes in classification accuracy in each iteration. This method reduces feature dimensionality while maintaining high classification accuracy through knowledge transfer between two subpopulations. However, its weight update mechanism tends to accumulate early evaluation biases, which impairs the global search ability, causes the loss of superior feature combinations, and limits its adaptability to the dynamically changing feature importance in practice. Accordingly, how to efficiently utilize valuable historical information during population iteration is both necessary and challenging.

To address these limitations, a multi-objective feature selection algorithm framework is proposed based on NSGA-II. Specifically, to balance the selection of individually important features and the enhancement of their interaction with other features, this paper proposes an importance-guided two-stage initialization strategy. The first stage prioritizes individually important features by boosting the selection probability of highly class-relevant features, ensuring core important features are prioritized in the initial population. The second stage focuses on strengthening feature interaction by coordinating combinations of features from different importance levels to increase the selection probability of potentially valuable features, thus promoting interaction between important features and other potentially effective ones. To further identify features better suited to the current iterative population and enhance the effectiveness of feature interactions, this paper proposes a historical information-assisted population update strategy. This strategy identifies feature combinations better suited to the current iterative population by distinguishing between high- and low-performing ones across different evolutionary states, thereby improving the efficiency of feature interactions and helping the algorithm avoid being trapped in local optima. The primary contributions of this work are outlined as follows:

1) A novel multi-objective feature selection method (MOFS-IH) is proposed. This method synergizes various feature-level combinations, fully integrates key historical information from different evolutionary states, significantly boosts the efficiency and relevance of inter-feature interactions, and ultimately yields a more desirable optimal feature subset.
2) An importance-guided initialization strategy that utilizes feature-level information and optimizes inter-level combinatorial relationships is devised. This approach not only guarantees the preservation of important features, but also ensures that the initial population has diversity and coverage in the feature space, laying a solid foundation for global search.
3) A historical information-assisted population update mechanism is proposed to systematically screen and dynamically prioritize the feature combinations that have demonstrated high effectiveness during the evolutionary process. In this mechanism, key historically effective information is harnessed to facilitate the population in breaking free from local optima.

2. Related work

Within this section, the advantages and disadvantages of existing research methods are first analyzed. Subsequently, our baseline algorithm and motivation are introduced.

2.1. Multi-objective feature selection optimization

Multi-objective optimization problems (MOPs) arise when optimal decisions necessitate tradeoffs among two or more conflicting objectives (Jiao et al., 2022a). Conflicting objectives are transformed into a unified form (i.e., all minimization or all maximization) to simplify the solution process. When objectives differ in optimization directions, those requiring conversion are adjusted based on specific needs (Jin & Sendhoff, 2008).

Such frameworks have been widely adopted in real-world applications, and feature selection stands out as a representative scenario. It inherently involves conflicting objectives that require tradeoffs, perfectly aligning with the definition of MOPs (Kamalov et al., 2025). In feature selection, minimizing classification error and minimizing the size of the feature subset are two conflicting objectives. Taking a dataset with D features as an example, a multi-objective feature selection problem with two objectives can be formulated as follows:

$$\begin{aligned} &\min F(X) = (f_1(X), f_2(X)) \\ &\quad \text{s.t. } X = (x_1, x_2, \dots, x_D) \in \Omega \\ &\qquad x_m \in \{0,1\} \quad \forall m \in \{1,2,\dots,D\} \end{aligned} \tag{1}$$

where $f_1(X)$ is the classification error of the feature subset represented by X, and $f_2(X)$ is the number of selected features in X, D denotes the dimensionality of a solution X within the decision space Ω, $x_m = 1$ means selecting the mth feature, and $x_m = 0$ means not selecting this feature.

2.2. EC for multi-objective feature selection

Feature selection in classification is fundamentally a multi-objective task with two core objectives: minimizing classification error and minimizing the size of the feature subset. Such tasks require balancing objective conflicts and global optimization, and EC, which features a population-driven search mechanism and powerful global search performance, has become an ideal tool for solving them (De La Iglesia, 2013).

Utilizing correlation-guided mechanisms enables focusing on high-potential feature regions and reducing blind search. In Chen et al. (2021), a feature selection algorithm integrating correlation guidance and a surrogate model was proposed. In this method, a guidance mechanism was constructed according to the correlation scores between features and classes to construct a candidate particle pool, and the surrogate model was combined to improve the computational efficiency of particle fitness. Meanwhile, particles with high fitness were preferentially

selected and their nearest neighbors were eliminated to preserve population diversity. Experimental results showed that the method could explore promising candidate solutions with efficiency,but its problem lies in that it overemphasized features highly correlated with classes and might neglect the high performance of weakly correlated features via combination.

Combining multiple strategies during population initialization represents a popular approach to improve the search performance of EC in feature selection. In Gupta and Gupta (2024), a hybrid multi-objective algorithm combining Random Forest, Information Gain, ReliefF, and NSGA-II was developed to integrate the advantages of filter and wrapper methodologies. This method evaluated the correlation between features and classes using these filter methods, and then obtained corresponding feature scores. Subsequently, the feature scores were used to guide the construction of the initial population and the execution of genetic operators, improving convergence efficiency. Experimental results demonstrated that this method attained lower classification accuracy than the other compared approaches. Its limitation lies in that this method largely overlooked inter-feature interactions during initialization due to its primary focus on feature-class correlation, and it also required manual adjustment of the weight vectors for its filter methods across different datasets.

Guiding evolution with historical information helps achieve a balance between exploration and exploitation in feature selection tasks. In Huang et al. (2025b), a binary particle swarm optimization algorithm was proposed. It constructed dual archives to store "acceleration coefficients" from historical iterations and dynamically adjusted learning parameters using a Gaussian distribution. Furthermore, it introduced a weighted center vector to replace the personal best position and employed an arctangent transfer function with greedy selection to enhance search directionality and convergence efficiency. The experimental results indicated that this method efficiently coordinated exploitation and exploration, outperforming various comparative algorithms on multiple datasets. However, its historical information came only from internal parameters, ignoring feature-level importance or correlation, which limited its insight into the feature space.

Obtaining multiple optimal feature subsets that share the same or similar objective function values has gained popularity for feature selection. For instance, a multi-objective differential evolution algorithm was designed to find multiple optimal feature subsets in Wang et al. (2021). In this algorithm, an initialization method based on feature relevance was presented to provide a starting point first. Then, a clustering strategy was adopted to split the population into multiple subpopulations. Finally, a dynamic crowding distance measure was utilized to preserve the diversity of each subpopulation. Experimental results indicated that this approach was able to acquire feature subsets with comparable or identical classification performance. Nevertheless, this method might fail to balance exploring optimal regions and identifying similar feature subsets, thereby hindering the algorithm's performance.

In summary, though various EC-based approaches have been presented to enhance feature selection performance through various strategies, they still suffer from two fundamental and interconnected challenges: the neglect of feature interactions during initialization and the underutilization of historical information across the evolution process.

2.3. Motivation

The above analysis indicates that most existing multi-objective feature selection methods initialize populations based on feature-class correlations while ignoring feature interactions, and utilize historical information only within a single iteration, failing to reuse effective combinations across generations. Such shortcomings risk compromising both the breadth and depth of the search. While highly correlated core features effectively simplify the feature space, insufficient utilization of information on potential efficient combinations in lowly correlated features often leads to a single search direction of the algorithm. Therefore, a novel two-stage population initialization strategy guided by feature importance level is proposed. In the first stage, the population is initialized from high-importance core features to ensure the desired quality and focus of initial solutions. In the second stage, the remaining initialization is completed through cross-level feature combination to endow the population with sufficient search diversity and avoid a single initial search direction. Furthermore, to better achieve a balance between exploration and exploitation during the search process by utilizing historical information, we reuse key information including feature screening and combination effectiveness from previous iterations to design a historical information-assisted update strategy, achieving dynamic balance between the two by adjusting the processing focus stage by stage to match varying iterative requirements.

3. Proposed approach

The framework of the MOFS-IH method is first introduced in this section, and its two key strategies are then elaborated in detail while its computational complexity is finally analyzed.

3.1. Framework of the proposed approach

The general framework of MOFS-IH is presented in Algorithm 1. Specifically, the method is divided into two core components: (1) An importance-guided population initialization strategy. (2) A historical information-assisted population update method. These two components are logically linked: the initialization strategy lays a high-quality foundation for subsequent evolution by including potential valuable features in the initial population, while the population update method further mines and retains valuable information from the initially accumulated evolutionary history.

Within the first component, the correlation between features and classes is evaluated using IG, then normalized, and sorted in descending order. After that, the feature space is dynamically partitioned into multiple levels by the proposed feature level division method. On this basis, the population is initialized using the probability values corresponding to each level and the properties (within the interval $[0, \pi/2]$, the rate at which y decreases accelerates as x increases) of the cosine function (Lines 1–5), which thereby significantly improves the quality of the initial population while ensuring potentially valuable features are not overlooked. As for the second part, a set of features that appear frequently and perform excellently during iterations is identified via the established elite archive and inferior individual archive, and after determining whether to merge and deduplicate the features contained in the current population, the population at the current iteration is updated via the proposed population update mechanism (Lines 10–21), which leverages valuable historical information while retaining the evolutionary states from previous iterations.

This study focuses on four key research issues: (1) How to design a novel method to effectively improve the selection probability of potentially valuable features? (2) What kind of strategy can be developed to enhance the quality of the initial population without adversely affecting the selection of potentially valuable features? (3) How can we mine valuable information for the evolutionary population from historical iterations? (4) By what approach can we design a strategy to efficiently utilize the mined valuable information while retaining the key evolutionary states from previous iterations? The details of how to tackle these four research issues are fully illuminated in the following sections.

3.2. Importance-guided initialization strategy

In feature selection tasks, initialization quality is critical as it directly determines the direction and effectiveness of the subsequent process. An effective initialization strategy can greatly enhance the efficiency of identifying high-quality solutions. To this end, we propose an importance-guided population initialization strategy, which balances

Algorithm 1 The framework of MOFS-IH.

Input: N: Population size; $T_{\max}$: Maximum number of iterations; T_{update}: The timing of population update;
Output: Non-dominated feature subsets;
1: $W \leftarrow$ Compute feature weights using IG and normalize them;
2: $W_s \leftarrow$ Sort W based on descending order;
3: $FL \leftarrow$ Dynamically divide feature levels using W_s;
4: $Pro \leftarrow$ Calculate the probability value of each FL using Eq. (6);
5: $P_t \leftarrow$ Utilize Pro and the properties of the cosine function to initialize the population with N individuals;
6: **for** $i = 1$ to $T_{\max}$ **do**
7: $gen \leftarrow i$;
8: $Q_t \leftarrow$ Perform the tournament selection, the uniform crossover, and the random position mutation on P_t to produce offspring;
9: $R_t \leftarrow P_t \cup Q_t$;
10: $Elite\_A \leftarrow$ individuals in the first front of R_t;
11: $Inferior\_A \leftarrow$ individuals in the last front of R_t;
12: $P_t \leftarrow$ Use the elite selection strategy to select parents from R_t;
13: **if** $gen \in T_{update}$ **then**
14: $merge\_pro \leftarrow$ Determine whether to save the features in the population at the gen based on probability values computed using Eq. (9);
15: $Changed\_F \leftarrow$ Obtain features with changed importance by updating feature importance using $Elite\_A$ and $Inferior\_A$;
16: **if** $rand < merge\_pro$ **then**
17: $contained\_F \leftarrow$ the features in the population at the gen;
18: $Renew\_F\_A \leftarrow$ Merge sets $contained\_F$ and $Changed\_F$, then remove duplicate features;
19: **else**
20: $Renew\_F\_A \leftarrow Changed\_F$;
21: **end if**
22: Update P_t using the $Renew\_F\_A$;
23: **end if**
24: **end for**
25: Obtain the non-dominated feature subsets;
26: **return** Non-dominated feature subsets.

the quality and exploratory diversity of the initial population via a synergy mechanism of feature importance stratification and cross-level combination, thereby generating a high-quality and diverse initial population. Specifically, the strategy involves three steps: acquisition of data characteristics, division of feature levels, and two-stage population initialization. Details of each component are elaborated below.

1) Acquisition of Data Characteristics: At the beginning, IG is employed to quantify the correlation between features and classes. The reason is that IG doesn't rely on parameter adjustment and outperforms other metrics in determining non-linear relationships between variables (Yin et al., 2023). Equation (2) calculates the IG of feature a as the difference between the entropy of the dataset and the conditional entropy given a, and the IG is then normalized to obtain the weight corresponding to feature a. The larger the weight, the more important the feature (Ramasamy & Meena Kowshalya, 2022). Notably, different approaches are employed to calculate IG for continuous and discrete features, as using different computing methods can produce distinct results.

For simplicity and uniformity, we first apply the CAIM (Kurgan & Cios, 2004) method to discretize the continuous features of the training set when computing IG. Unlike traditional discretization methods that require manual presetting of the number of intervals, CAIM requires no manual parameter setting. It maximizes global class-attribute interdependence through greedy iteration and automatically determines the optimal number of intervals and split boundaries. The original continuous data are retained for subsequent objective function evaluation and test set processing.

$$IG(D,a) = H(D) - H(D|a) \tag{2}$$

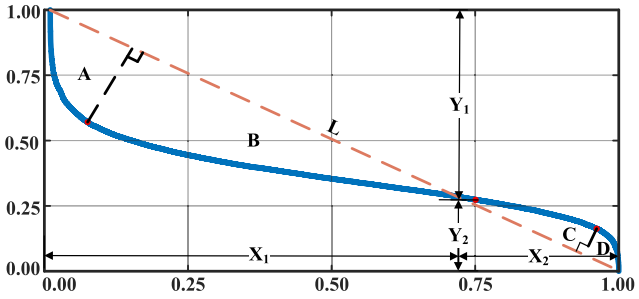

Fig. 1. An example is employed to illustrate the intersection.

where $H(D)$ represents the entropy of the dataset, and $H(D|a)$ is the conditional entropy given a.

After obtaining the normalized IG weights for all features, the features are sorted in descending order according to their weights to generate a curve reflecting the distribution of feature importance. Such curves fall into two categories, distinguished by whether they intersect a predefined reference line L (i.e., the line connecting the start and end points of the feature importance curve). Fig. 1 shows the case with intersections. For details of subsequent strategies, we take the curve with an intersection as an example.

2) Division of Feature Levels: Algorithm 2 presents the details of the dynamic division of feature levels. First, the point where the curve intersects the start-end connection line L is taken as a clear boundary point. Abscissa and ordinate interval variations for the two regions (before and after this point) are extracted and denoted as X_1, Y_1, X_2, and Y_2. Subsequently, the ratio R_1 of the number of levels in the two segments separated by the boundary point is given by Eq. (3) (Line 1). This is because when a slight change in the number of features corresponds to a marked change in importance, it indicates that small changes in these features may lead to significant differences. Therefore, levels need to be divided more finely to avoid overlooking key details.

$$R_1 = \left(\frac{X_1 \cdot Y_2}{X_2 \cdot Y_1}\right) \tag{3}$$

Furthermore, each divided segment can be regarded as a non-intersecting curve. Subsequently, for each segment, its knee point (Das, 1999) is identified, which is the point farthest from the line L. The rationale is that the point farthest from L corresponds to where the curve bends most sharply, i.e., the position with the most significant change in feature variation. Dividing the curve at this point yields intervals with distinct feature variation patterns. Within each interval, features change gently with a consistent trend and similar functional properties, while significant differences exist between intervals. This effectively addresses the issue of insufficient within-layer homogeneity in feature stratification. After this identification step, for the regions on both sides of the knee point, the ratio is calculated using Eq. (3) to clarify the relationship of levels between the two regions (Line 2). Based on this, taking the segment with the smallest ratio as the base (rounded up), the number of levels for each segment is deduced inversely through the ratio relationship (Lines 3–12). Lastly, for each segment, the positional information corresponding to each level is obtained by first selecting target regions from regions generated by each division based on Eq. (3) (where the front region is selected if the result ≥ 1, otherwise the back region) (Lines 13–17).

To further clarify the above process, we provide an example. Suppose the first knee point divides its segment into A and B in a ratio R_2 of 3.4 : 2.6; the second knee point divides its segment into parts C and D in a ratio R_3 of 0.6 : 0.2. Meanwhile, $R_1 = 3 : 1$. It is first determined that the number of levels contained in D is 1 (rounded up) because $R_1 \geq 1$ and $R_3 \geq 1$ (with D corresponding to the smallest ratio as the base); and the number of levels contained in C is derived as 3 from R_3. Following this, since the total number of levels contained in C and D is 4 and

Algorithm 2 Division of feature levels.

Input: W_S: Feature importance scores (sorted descending);
Output: L_s: Level start points; L_e: Level end points; k: Total level counts; Pro: Level probabilities;
1: $R_1 \leftarrow$ Compute level ratio of segments before/after intersection via Eq. (3) (based on W_S);
2: $R_2, R_3 \leftarrow$ Determine knee points of two segments and compute level ratios via Eq. (3) (left→right);
3: Name the four segments divided by two knee points as A, B, C, D (left→right);
4: **if** $R_1 \geq 1$ **then**
5: $C_L, D_L \leftarrow$ Compute level counts for C, D via R_3 (order: D first if $R_3 \geq 1$, else C);
6: $N_{AB} \leftarrow$ Determine total level counts of A, B via R_1;
7: $A_L, B_L \leftarrow$ Compute level counts for A, B via R_2 (priority: B first if $R_2 \geq 1$, else A);
8: **else**
9: $A_L, B_L \leftarrow$ Compute level counts for A, B via R_2;
10: $N_{CD} \leftarrow$ Determine total level counts of C, D via R_1;
11: $C_L, D_L \leftarrow$ Compute level counts for C, D via N_{CD}, R_3 (same priority logic as $R_1 \geq 1$);
12: **end if**
13: Determine knee points iteratively via Eq. (3) based on A_L, B_L, C_L, D_L (select front if ratio ≥ 1, else back) until all determined;
14: $L_c \leftarrow$ [curve start, all knee points, end] (1-indexed);
15: $L_s \leftarrow$ Extract points from L_c at odd indices (1st, 3rd, 5th, …) as level start points;
16: $L_e \leftarrow$ Extract points from L_c at even indices (2nd, 4th, 6th, …) as level end points;
17: $k \leftarrow$ Sum of A_L, B_L, C_L, D_L (total level counts);
18: $Pro \leftarrow$ Calculate level probabilities via Eq. (6);
19: Finish dividing feature levels;
20: **return** L_s, L_e, k, Pro

$R_1 = 3:1$, it can be known that the total number of levels contained in A and B is 12. Meanwhile, because $R_2 = 3.4:2.6$, it can be determined that the number of levels contained in B is 5 (rounded up), and the number of levels contained in A is 7 (rounded up). Finally, taking C as an example, to determine the positions of the 3 levels, the first knee point is identified in the region of C. Among the two segments divided by this knee point, the segment selected by following the same rule based on Eq. (3) as above (i.e., selecting the front region if the result ≥ 1, else the back region) is further searched for another knee point. At this point, the region of C has been divided into three levels. The coordinates are saved for subsequent initialization.

To convert structural information from feature-level division and level-priority differences into quantitative metrics, these differences are quantified as probability values Pro, calculated as in Eq. (6) (Line 18). Notably, the quantification of structural information considers changes in feature counts and their importance. Specifically, changes in feature counts F_n are quantified via Eq. (4) as the ratio of the level's feature count f_n^l to the total feature count f_n^t, and changes in feature importance E_c are characterized via Eq. (5) as the intra-level difference in feature importance. Moreover, since exponential functions decrease with increasing exponents, this difference is defined as the maximum value minus the original value. Furthermore, the higher a feature level's position, the less important the features it contains. Thus, we incorporate the level position l as a lag term to capture level priority variations.

$$F\_n = f_n^l / f_n^t \tag{4}$$

$$E\_c = 1 - (\max(f_i^l) - \min(f_i^l)) \tag{5}$$

$$Pro = exp(-(min(F\_n, E\_c) + 0.5 \cdot l) \cdot l) \tag{6}$$

where f_n is the feature count, f_i is the feature importance, and L is the number of levels.

3) Two-Stage Population Initialization: Building on the above preparations, a two-stage initialization strategy is designed. Algorithm 3 details this process. The first stage focuses on initializing each subpopulation's corresponding target level based on Pro to ensure utilization of important features. It is divided into two steps: First, we utilize the cosine function's characteristic in $[0, \pi/2]$, where the rate at which y decreases accelerates as x increases. This characteristic aligns with the idea that the number of assigned individuals gradually diminishes as probability values decrease. Guided by this characteristic, the entire population is first divided into k parts (Lines 1–2) (k is the number of levels). Second, each subpopulation's corresponding target level is initialized based on its associated Pro (Lines 5–11).

After first-stage target-level initialization, the second stage aims to broaden the initial population's distribution by configuring each individual's remaining levels. This stage consists of two steps. The first step adheres to the principle: more individuals should be allocated to low-probability features and fewer to high-probability ones, and this principle is employed to determine an individual's selection probability for each feature. This balances the selection opportunities of different features, avoiding resource redundancy or oversight. Under this principle, a preset threshold (hyperparameter β) is used to determine whether to initialize the remaining levels for each individual; if required, for each feature in the remaining levels, its three selection probabilities (0.25, 0.5, 0.75 Xu et al., 2020b) are assigned weights with a ratio of 3: 2: 1 (corresponding to 0.75, 0.5, 0.25, in inverse proportion to the probabilities). Specifically, 0.25, 0.5, and 0.75 are assigned weights of 3, 2, and 1 respectively (Lines 12–23). This inverse weighting scheme increases the selection probability of low-probability features and prevents individual invalidity that could arise when all levels adopt a uniformly high probability (e.g., 0.75). Following the pigeonhole principle, the differentiated probabilities treat limited level slots as pigeonholes, enabling features of different importance to co-occupy these slots. This promotes synergistic feature combinations, strengthens feature interactions, and facilitates the subsequent evolutionary search. Moreover, selected features are gradually eliminated for each level. When the feature space of a level becomes empty, the space is updated. This ensures all features in each level remain selectable, enhancing the efficiency of exploring effective feature combinations.

To further clarify the above process, a concrete example is provided. Specifically, suppose the feature space is divided into three levels, with probabilities of 0.8 ($Pro(1)$), 0.4 ($Pro(2)$), and 0.2 ($Pro(3)$) in ascending order of levels (i.e., the higher the level, the lower the probability). First, the arccosine function defines the x-range for y from 1 to $Pro(3)$, which is evenly discretized into N units (discretization units). The number of individuals per subpopulation is then determined by the discretization units within the interval between its Pro-corresponding x-value and the previous level's. For example, assuming the population size N is 100, each discretization unit has length $u = (\arccos(0.2) - \arccos(1))/N$. The first subpopulation contains $\arccos(0.8)/u$ individuals, with the remaining subpopulations calculated similarly. Second, take individual a_i in the first subpopulation (matching the first level) as an example. For features in its associated first level, Pro(1) is used to determine if the features are selected. For the remaining levels, there is a 70% (pre-set threshold) probability that the level's features are selected by the individual with a 50% chance of 0.25, a 37.5% chance of 0.5, and a 12.5% chance of 0.75; otherwise, no features in that level are selected.

3.3. Historical-assisted update strategy

In the later iterations, the population may get trapped in local optima or the algorithm may suffer from overfitting, essentially due to the algorithm's inability to explore better feature combinations. Briefly, enabling the population to explore more feature combinations in later

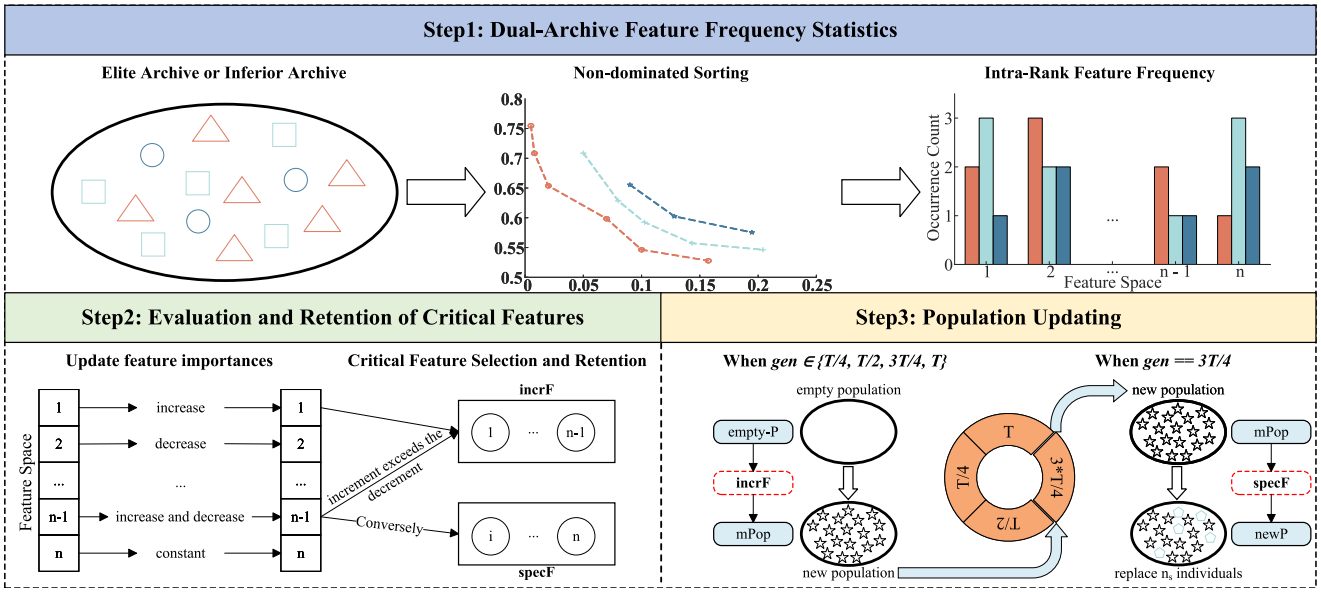

Fig. 2. A three-level example illustrating how historical information updates feature importance.

Algorithm 3 Population initialization.

Input: L_s: Level start points; L_e: Level end points; Pro: Level probability values; k: Total level count; N: Population size;
Output: P: initialization population;

$subp\_size \leftarrow$ Compute sizes of k subpopulations via arccos: x-range $[\arccos(Pro(k)), 0]$ split into N equal units, summing to N;
where Each $subp\_size$ element equals discretization units between current and previous level x-values;
$count \leftarrow 1$;
for $i = 1$ to k **do**
  for $j = 1$ to $subp\_size(i)$ **do**
    $ind \leftarrow$ Initialize individual with all solutions as 0;
    for $z = L_s(i)$ to $L_e(i)$ **do**
      if $rand() \leq Pro(i)$ **then**
        $ind.solution(z) \leftarrow 1$;
      end if
    end for
    for $m = 1$ to k **do**
      if $m \neq i$ **then**
        if $rand() \leq \beta$ **then**
          for $q = L_s(m)$ to $L_e(m)$ **do**
            $Pro_1 \leftarrow$ Select feature probability from 0.25, 0.5, 0.75 by ratio 3:2:1;
            $ind.solution(q) \leftarrow 1$ if $rand() \leq Pro_1$;
          end for
        else
          $ind.solution(L_s(m) : L_e(m)) \leftarrow 0$;
        end if
      end if
    end for
    P(count) $\leftarrow$ ind;
    $count \leftarrow count + 1$;
  end for
end for
return P

iterations helps achieve better solutions. To address this issue, a population update strategy is proposed to leverage historical information for identifying important features and exploring more additional optimal feature combinations. The strategy is divided into three stages across the total number of iterations T: the first stage $[T/4, T/2]$ aims to rapidly eliminate redundant and ineffective features; the second stage $[T/2, 3T/4]$ focuses on identifying higher-quality candidate features; the third stage $[3T/4, T]$ is dedicated to exploring feature combinations more compatible with the current population.

The equal division of iterations aligns with the coarse-to-fine progressive search pattern of feature selection, where the three stages are interlocked and equally important. Uniformly allocating the iteration budget thus becomes the most natural realization of this process, ensuring each stage receives sufficient search resources. It balances convergence speed and global exploration while eliminating extra parameter tuning, thereby enhancing the generality of the strategy. Fig. 2 illustrates the overall framework of the strategy, and specific details are elaborated as follows.

1) Dual-archive feature frequency statistics: The acquisition of effective historical information involves two key points. One is to distinguish important features from unimportant ones based on historical information. The other is to identify which features are more likely to combine with the effective features contained in the current population to help the algorithm escape local optima. For the first key point, the procedure is as follows. At the fixed iterations $(T/4, T/2, 3T/4)$, individuals from the top two fronts generated in each iteration are stored in the elite archive $Elite\_A$, and individuals from the bottom two fronts are stored in the inferior archive $Inferior\_A$. Individuals with completely identical feature selection vectors are removed from $Elite\_A$ and $Inferior\_A$ before each frequency update to prevent inflated feature frequencies that distort solution validity, while maintaining a diverse archive so that feature importance reflects true prevalence. Next, at these fixed iteration points, non-dominated sorting is applied to all individuals in $Elite\_A$ and $Inferior\_A$. Based on their Pareto ranks, the importance of the features in all individuals in these two archives is updated. For each feature, its frequency across archives and its front are comprehensively evaluated. Intuitively, a feature that appears more frequently in archive $Elite\_A$ and is situated at a lower front is deemed more important, thereby justifying an increase in its importance score. This evaluation logic applies similarly to other cases. Specifically, increases are calculated using Eq. (7) and decreases using Eq. (8).

$$dec(a) = \sum_{j=1}^{n_1} f_i^a \cdot ((d_n/(d_n + i_n)) \cdot (1/4)^{(n_1 - j)}) \tag{7}$$

$$inc(a) = \sum_{k=1}^{n_2} (1 - f_i^a) \cdot ((i_n/(d_n + i_n)) \cdot (1/4)^{(k-1)}) \tag{8}$$

where $dec(a)$ is the part of a importance that decreases, $inc(a)$ is the part of a importance that increases, n_1 denotes Pareto rank counts in archive $Inferior\_A$, n_2 represents Pareto rank counts in archive $Elite\_A$, f_i^a means the importance of a, d_n denotes the frequency of a in archive $Inferior\_A$, and i_n represents the frequency of a in archive $Elite\_A$.

2) Evaluation and retention of critical features: Features with increased importance are stored and split into two groups based on whether their importance increment exceeds their decrement. Specifically, features for which the increment exceeds the decrement (denoted as $incrF$) are likely to have general validity and are directly utilized in subsequent iterative points ($T/4, T/2, 3T/4$) to retain superior features. In contrast, those for which the increment does not exceed the decrement (denoted as $specF$) are only effective in specific feature combinations. Thus, the results of such features at each iterative point are stored separately and reserved exclusively for the final point $3T/4$ to help the algorithm escape local optima. For the second key point, whether to combine $incrF$ with the features contained in the population at the corresponding iterative point (denoted as $popF$) is determined by the probability Pro_1 obtained from Eq. (9). In theory, the later the iteration, the higher the retention value of $popF$. In other words, fewer remaining iteration points mean that $popF$ warrants greater retention. Thus, Pro_1 decays exponentially with the remaining iteration count rem, using a base of $1/2$. A larger base slows decay and risks premature convergence; a smaller one accelerates decay and may discard useful features. The value $1/2$ is a classic choice in evolutionary algorithms, providing a moderate decay rate.

$$Pro\_1 = (1/2)^{rem} \tag{9}$$

where rem is the number of remaining iterative points.

3) Population update: The population update method is divided into two parts based on the use of historical information. One part is that at each iteration point ($T/4, T/2, 3T/4, T$), a new population is generated using the $incrF$ corresponding to that iteration point. This population then replaces the original population to form the new population $mPop$ for subsequent evolution. This method enriches the population with high-quality features, thereby enabling the population to meet the specific objectives of each iteration point. Notably, while this stage-wise population update effectively targets current objectives, it discards previous iteration results. To fully leverage the advantages of this stage-wise update, we optimize the individual retention strategy by retaining the top two fronts of individuals from each iteration and the complete population at each iteration point, and the retained data is ultimately used for test set validation.

The other part is that at the $3T/4$th iteration point, n_s individuals are randomly selected from $mPop$, where n_s is determined by Eq. (10). The value of n_s depends on the relationship between remaining iterations and population size. Subsequently, three selection probabilities $(0.25, 0.5, 0.75)$ at a ratio of $1:2:3$ are assigned to each feature in $specF$ for these n_s individuals. This is because retaining the $specF$ features in the current population as much as possible facilitates more thorough exploration of more population-compatible feature subsets. Note that, to ensure the feature combinations between $specF$ and the current population are fully explored, these n_s individuals must be retained in the population without elimination in subsequent iterations. Alternatively, not updating their objective function values achieves the same effect.

$$n_s = ((T - gen)/T) \cdot N \tag{10}$$

where gen is the current number of iterations.

3.4. Computational complexity

In terms of computational complexity, the primary consumption of MOFS-IH arises from the nondominated sorting procedure. First, computing the IG between each feature and the class has a complexity of $O(n)$ in feature-level division, where n denotes the dimensionality of the feature space. Second, the dataset is dynamically divided into multiple levels based on its structure. For this step, computing the first inflection point requires the distance from the normalized IG of the entire feature space to L (i.e., the line connecting the first and last points of the normalized IG distribution), while computing the remaining inflection points requires the distance from the normalized IG of each feature subset to L. Hence, the total complexity of this step is $O(2 \cdot n)$. Third, maintaining and updating the dual archives for storing historical feature information costs $O(N)$ (N being the population size). The feature space update guided by historical information has a complexity of $O(4 \cdot n)$, and that of the nondominated sorting operation is $O(2 \cdot N^2)$. Thus, the total complexity of MOFS-IH is $O(6 \cdot n + N + 2 \cdot N^2) = O(2 \cdot N^2 + n)$. This indicates that when $n > N^2$, the computational complexity of the algorithm is bounded by the dimensionality of the feature space, whereas when $N^2 > n$, its complexity is comparable to that of existing NSGA-II-based multi-objective algorithms.

4. Experiment design

A set of experiments was performed to confirm the effectiveness of MOFS-IH by comparing it with baseline algorithms. This section provides an overview of the investigated datasets, compared algorithms, parameter settings, and sensitivity analysis of key parameters.

4.1. Datasets

To assess the performance of the proposed feature selection algorithm, 15 real-world datasets are used for experimental analysis. These datasets are open-source and can be downloaded from two public repositories: https://archive.ics.uci.edu/ml/datasets.php and http://featureselection.asu.edu. Table 1 presents their basic information. The datasets cover both low and high dimensions, with the number of features ranging from 1024 to 12,533, and involve both binary and multi-class classification tasks. Moreover, most datasets have small sample sizes and high-dimensional feature spaces, which greatly increase the difficulty of classification. Thus, the datasets used in our experiments are sufficient to verify the generality and efficacy of each competing method.

4.2. Comparative methods and parameter settings

To assess the proposed approach, six state-of-the-art comparative algorithms are selected to compare with MOFS-IH. The details are as follows:

1) MOEA/D-based feature selection combining dominance and decomposition search forms (DDFS) (Liang et al., 2024).
2) MOEA/D-based for multi-objective feature selection with single-objective tasks (MFMOEA) (Liang et al., 2023).
3) A variable granularity search-based improved algorithm for feature selection (VGS-MOEA) (Cheng et al., 2022).
4) An information gain- and clustering-based feature selection algorithm (IGEA) (Zhang et al., 2024).
5) A hybrid algorithm integrating spectral clustering-based filtering and group evolution (FG-HFS) (Xu et al., 2024).
6) MOEA/D-based with optimization and knowledge transfer for unbalanced classification (MFUBFS) (Liang et al., 2025).

Table 2 presents the parameter settings for the six comparative methods and the proposed approach MOFS-IH. Notably, the parameter settings of the comparative algorithms follow those in their original papers. To ensure fairness, the common parameters need to be kept consistent. First, the maximum number of iterations (T) and population size (N) are both set to 100. Second, to ensure stable and statistically significant results, all algorithms are run independently 30 times on each dataset, and the final results are averaged. Third, each dataset is stratified by

Table 1
Basic information Of 15 datasets.

N0.	Datasets	#Features	#Classes	#Instances	#Instance/Feature
1	Yale	1024	15	165	0.16
2	toxicity	1203	2	171	0.14
3	SRBCT	2308	4	83	0.04
4	GLIOMA	4434	4	50	0.01
5	Breast3	4869	3	95	0.02
6	Leukemia1	5327	3	72	0.01
7	DLBCL	5469	2	77	0.01
8	Brain1	5920	5	90	0.02
9	Prostate_GE	5966	2	102	0.02
10	Prostate6033	6033	2	102	0.02
11	ALLAML	7129	2	72	0.01
12	Adenocarcinoma	9672	2	76	0.01
13	Prostate_Tumor	10,509	2	102	0.01
14	CLL_SUB_111	11,340	3	111	0.01
15	11_Tumors	12,533	11	174	0.01

Table 2
Parameter setting.

Algorithms	Parameter values
DDFS (Liang et al., 2024)	The number of nearest neighbors $K = 5$.
MFMOEA (Liang et al., 2023)	The neighborhood size $a = 0.1 * N$, the neighborhood selection probability $CR = 0.8$, the angle value $Angle = 5$, the mutation probability $p_m = 1/k_1$ (k_1 is the number of selected features).
VGS-MOEA (Cheng et al., 2022)	$p_c = 0.9$, $p_m = 1/n$, $\alpha = 10$, the number of groups is $k_1 = min\{10 \cdot \log_2(n), n\}$.
IGEA (Zhang et al., 2024)	$K = 5$.
FG-HFS (Xu et al., 2024)	The parameter is adaptively adjusted.
MFUBFS (Liang et al., 2025)	The parameters are the same as those of MFMOEA.
MOFS-IH	$\beta = 0.7$, $p_c = 0.9$, the number of bits in random position mutation $u2 = 10$.

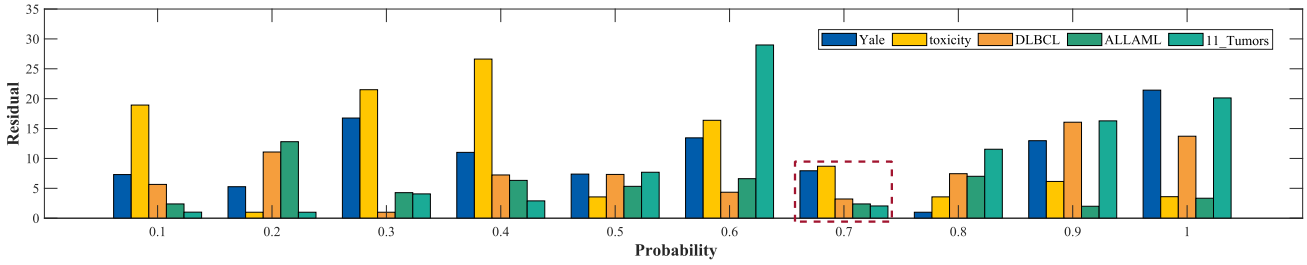

Fig. 3. Determining the β using datasets of varying dimensions.

class labels and split into a training set (80%) and a test set (20%) to validate classification performance. Additionally, ten-fold cross-validation is only performed within the training set, with the mean result across the ten folds taken as the objective function value of each individual, to obtain the optimal feature subset and avoid training deviation. The screened feature subset is then utilized on the independent test set for final performance evaluation. Finally, to achieve a trade-off between accuracy and efficiency, the K-value of the KNN classifier is set to 5 (Xu et al., 2020a), because its lazy learning nature requires no model training and incurs minimal evaluation time, which effectively reduces the computational overhead of repeated fitness assessments during evolutionary search.

In this study, we evaluate seven multi-objective algorithms using hypervolume (HV) (While et al., 2006) and inverse generation distance (IGD) (Zitzler et al., 2003) as performance metrics. For the calculation of HV, the reference point is set to $(1, 1)$ (Bosman & Thierens, 2003). For the calculation of IGD, the true Pareto front for high-dimensional feature selection is nearly unobtainable. Following the method described in Jiao et al. (2022b), the Pareto fronts from 30 runs of each algorithm are merged into a combined population. The non-dominated candidate solutions from this combined population are treated as the true Pareto front. In general, higher HV values indicate better algorithm performance, whereas smaller IGD values represent superior algorithm performance.

4.3. Sensitivity analysis of parameters

In the initialization phase, a threshold β is introduced to help the algorithm achieve a more widely distributed population, and its value can significantly affect the algorithm's performance. To this end, we determined the optimal β value through experiments on different datasets. Specifically, we first ran the MOFS-IH algorithm 30 times for each β value to obtain the average HV for each dataset, and then identified the optimal HV value among these averages. Finally, we calculated the residuals by subtracting the HV values corresponding to different β values from the optimal HV value of each dataset, and then selected the β value with consistently small residuals across all datasets as the final parameter setting. Fig. 3 shows residuals for different β values across datasets of varying dimensions, where the optimal β value is marked with a box. To illustrate the residual differences more clearly, we magnified the residuals 100-fold.

5. Results and analysis

In this portion, the presented MOFS-IH is compared to the six state-of-the-art algorithms introduced in the previous section in terms of HV and IGD. These metrics are widely adopted to evaluate multi-objective optimization algorithms. To verify significant performance differences between algorithms, the Wilcoxon test at a significance level of 0.05 is

Table 3
Mean HV results on test data.

Datasets	DDFS	MFMOEA	VGS-MOEA	IGEA	FG-HFS	MFUBFS	MOFS-IH
Yale	7.32E-01±4.52E-02(+)	7.10E-01±2.82E-02(+)	7.60E-01±2.59E-02(+)	7.70E-01±3.55E-02(+)	7.09E-01±4.28E-02(+)	7.35E-01±4.92E-02(+)	**8.20E-01±3.11E-02**
toxicity	6.41E-01±8.12E-02(+)	7.50E-01±2.44E-02(+)	7.59E-01±2.92E-02(+)	7.01E-01±7.32E-02(+)	6.43E-01±6.43E-02(+)	7.78E-01±2.90E-02(=)	**7.80E-01±2.44E-02**
SRBCT	9.27E-01±7.07E-02(+)	9.84E-01±1.15E-02(+)	9.88E-01±3.33E-03(+)	9.52E-01±4.39E-02(+)	9.20E-01±4.18E-02(+)	9.56E-01±3.44E-02(+)	**9.99E-01±3.39E-04**
GLIOMA	8.76E-01±5.63E-02(+)	9.16E-01±4.26E-02(+)	9.09E-01±2.28E-02(+)	9.39E-01±4.34E-02(+)	8.92E-01±6.08E-02(+)	9.24E-01±4.18E-02(+)	**9.91E-01±2.33E-02**
Breast3	6.68E-01±8.40E-02(+)	7.27E-01±3.14E-02(+)	7.49E-01±3.60E-02(+)	7.31E-01±4.98E-02(+)	6.46E-01±8.08E-02(+)	7.40E-01±4.97E-02(+)	**8.13E-01±4.17E-02**
Leukemia1	8.95E-01±4.79E-02(+)	9.10E-01±3.00E-02(+)	9.53E-01±3.25E-02(+)	9.62E-01±3.79E-02(+)	8.70E-01±8.57E-02(+)	9.62E-01±3.35E-02(+)	**9.97E-01±1.21E-02**
DLBCL	9.52E-01±5.64E-02(+)	9.56E-01±3.58E-02(+)	9.92E-01±2.91E-03(+)	9.89E-01±2.37E-02(+)	9.17E-01±7.78E-02(+)	9.89E-01±2.37E-02(+)	**1.00E+00±9.06E-05**
Brain1	8.12E-01±4.48E-02(+)	8.13E-01±2.42E-02(+)	8.35E-01±1.97E-02(+)	8.40E-01±4.47E-02(+)	7.84E-01±3.68E-02(+)	8.42E-01±3.64E-02(+)	**8.95E-01±2.45E-02**
Prostate_GE	9.01E-01±4.64E-02(+)	8.96E-01±3.11E-02(+)	9.35E-01±2.51E-02(+)	9.66E-01±3.10E-02(+)	9.32E-01±2.92E-02(+)	9.70E-01±2.64E-02(+)	**9.70E-01±2.46E-02**
Prostate6033	9.12E-01±4.98E-02(+)	9.26E-01±1.37E-02(+)	9.32E-01±2.24E-02(+)	9.66E-01±2.55E-02(+)	9.34E-01±4.34E-02(+)	9.62E-01±2.30E-02(+)	**9.84E-01±2.28E-02**
ALLAML	9.49E-01±3.83E-02(+)	9.27E-01±2.46E-02(+)	9.66E-01±3.29E-02(+)	9.60E-01±3.32E-02(+)	9.21E-01±3.97E-02(+)	9.62E-01±3.35E-02(+)	**1.00E+00±1.51E-05**
Adenocarcinoma	7.96E-01±3.18E-02(+)	8.05E-01±1.51E-02(+)	8.63E-01±4.79E-02(+)	8.91E-01±4.90E-02(=)	8.12E-01±2.21E-02(+)	8.96E-01±4.13E-02(=)	**9.16E-01±3.41E-02**
Prostate_Tumor	8.32E-01±4.56E-02(+)	8.86E-01±2.72E-02(+)	9.17E-01±3.61E-02(+)	9.16E-01±6.21E-02(=)	8.70E-01±5.74E-02(+)	9.24E-01±4.07E-02(=)	**9.47E-01±2.29E-02**
CLL_SUB_111	6.82E-01±9.80E-02(+)	6.88E-01±6.56E-02(+)	7.72E-01±7.63E-02(+)	7.93E-01±5.41E-02(+)	7.45E-01±7.11E-02(+)	8.12E-01±5.55E-02(+)	**8.61E-01±4.85E-02**
11_Tumors	7.55E-01±3.79E-02(+)	7.82E-01±1.80E-02(+)	7.77E-01±3.51E-02(+)	7.56E-01±4.76E-02(+)	6.98E-01±5.60E-02(+)	7.68E-01±3.11E-02(+)	**8.30E-01±2.02E-02**
+/=/-	15/0/0	15/0/0	15/0/0	13/2/0	15/0/0	12/3/0	–
Friedman's rank	7	5	3	4	6	2	**1**

Table 4
Mean IGD results on test data.

Datasets	DDFS	MFMOEA	VGS-MOEA	IGEA	FG-HFS	MFUBFS	MOFS-IH
Yale	8.91E-02±2.48E-02(+)	1.24E-01±2.24E-02(+)	7.96E-02±1.17E-02(+)	5.35E-02±1.89E-02(+)	1.07E-01±2.50E-02(+)	7.21E-02±2.28E-02(+)	**3.70E-02±1.17E-02**
toxicity	1.98E-01±7.75E-02(+)	1.14E-01±3.44E-02(+)	9.76E-02±1.97E-02(+)	1.38E-01±7.11E-02(+)	1.84E-01±6.32E-02(+)	**6.19E-02±2.38E-02**(=)	6.44E-02±2.22E-02
SRBCT	5.23E-02±4.31E-02(+)	6.00E-02±1.88E-02(+)	3.20E-02±8.64E-03(+)	3.20E-02±2.13E-02(+)	7.02E-02±1.70E-02(+)	3.05E-02±1.92E-02(+)	**1.63E-02±1.38E-02**
GLIOMA	7.87E-02±3.62E-02(+)	7.36E-02±2.69E-02(+)	5.86E-02±1.96E-02(+)	3.98E-02±2.82E-02(=)	7.48E-02±2.52E-02(+)	5.05E-02±2.46E-02(+)	**1.95E-02±2.40E-02**
Breast3	1.37E-01±6.85E-02(+)	9.63E-02±3.05E-02(+)	7.39E-02±2.07E-02(+)	8.21E-02±3.49E-02(+)	1.52E-01±7.53E-02(+)	7.73E-02±3.37E-02(+)	**3.57E-02±1.83E-02**
Leukemia1	7.04E-02±4.12E-02(+)	6.13E-02±3.15E-02(+)	3.28E-02±1.32E-02(+)	2.81E-02±1.95E-02(+)	9.64E-02±7.09E-02(+)	2.56E-02±1.48E-02(+)	**6.84E-03±1.16E-02**
DLBCL	4.04E-02±3.62E-02(+)	5.01E-02±2.14E-02(+)	1.91E-02±1.14E-02(+)	1.50E-02±1.56E-02(+)	6.14E-02±6.76E-02(+)	1.22E-02±1.50E-02(+)	**5.71E-03±1.16E-02**
Brain1	7.21E-02±2.51E-02(+)	8.14E-02±1.82E-02(+)	5.52E-02±1.22E-02(+)	4.82E-02±2.76E-02(+)	9.06E-02±1.69E-02(+)	4.97E-02±2.23E-02(+)	**2.89E-02±1.50E-02**
Prostate_GE	6.61E-02±1.89E-02(+)	7.49E-02±1.64E-02(+)	5.01E-02±1.46E-02(+)	4.06E-02±2.17E-02(=)	5.60E-02±1.82E-02(+)	3.35E-02±1.82E-02(=)	**3.28E-02±1.92E-02**
Prostate6033	6.11E-02±4.51E-02(+)	5.18E-02±2.24E-02(+)	4.19E-02±1.96E-02(+)	1.94E-02±1.44E-02(+)	4.35E-02±3.63E-02(+)	2.34E-02±1.16E-02(+)	**9.21E-03±1.14E-02**
ALLAML	4.73E-02±3.32E-02(+)	6.68E-02±1.81E-02(+)	3.97E-02±2.51E-02(+)	4.03E-02±3.28E-02(+)	7.82E-02±3.87E-02(+)	3.87E-02±3.26E-02(+)	**2.34E-04±1.49E-04**
Adenocarcinoma	1.34E-01±3.14E-02(+)	1.24E-01±1.14E-02(+)	8.63E-02±3.16E-02(+)	6.57E-02±3.00E-02(=)	1.21E-01±2.16E-02(+)	5.63E-02±2.51E-02(=)	**4.73E-02±2.13E-02**
Prostate_Tumor	1.13E-01±4.30E-02(+)	7.89E-02±2.85E-02(+)	4.91E-02±2.53E-02(+)	5.35E-02±4.45E-02(=)	8.65E-02±4.87E-02(+)	4.05E-02±2.90E-02(=)	**2.28E-02±1.25E-02**
CLL_SUB_111	1.83E-01±8.39E-02(+)	1.77E-01±5.41E-02(+)	1.10E-01±5.90E-02(+)	9.23E-02±3.50E-02(+)	1.29E-01±5.06E-02(+)	8.03E-02±3.53E-02(+)	**4.56E-02±2.37E-02**
11_Tumors	7.98E-02±2.07E-02(+)	1.03E-01±2.38E-02(+)	7.86E-02±1.42E-02(+)	4.94E-02±1.98E-02(-)	1.71E-01±3.42E-02(+)	**4.59E-02±1.31E-02**(-)	6.92E-02±2.30E-02
+/=/-	15/0/0	15/0/0	15/0/0	10/4/1	15/0/0	10/4/1	–
Friedman's rank	6	5	4	3	7	2	**1**

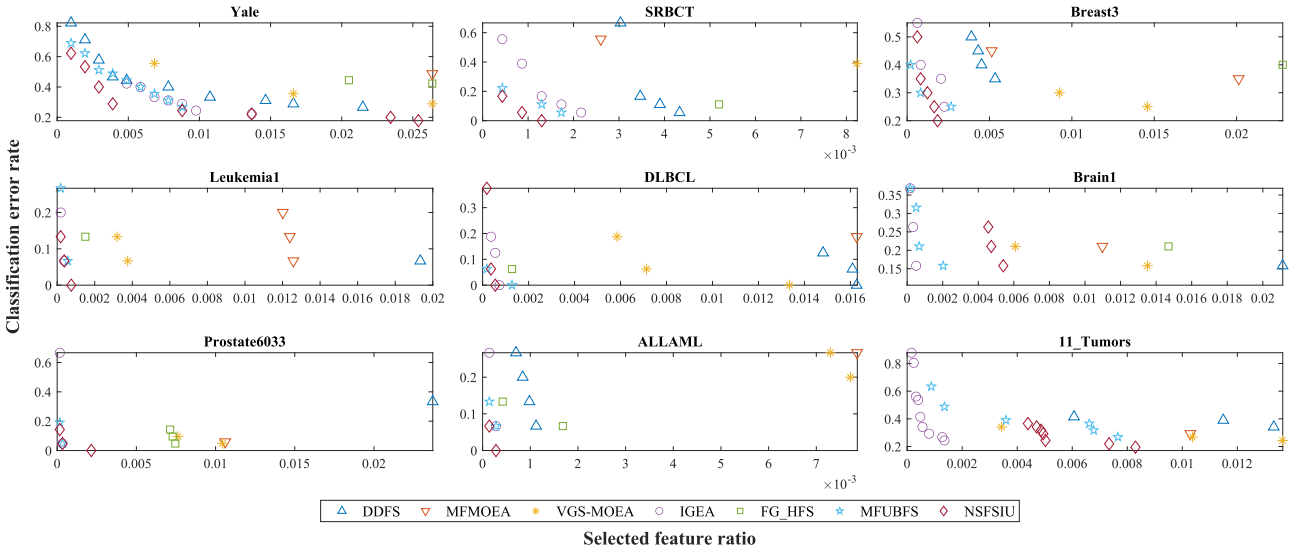

Fig. 4. Distributions of nondominated solutions obtained by each algorithm on test sets in terms of median HV value.

employed to examine the differences between MOFS-IH and the other evaluated algorithms, while the Friedman test is used to offer a comprehensive ranking. In the results, "+", "=", "-" are used to represent whether the proposed algorithm is significantly superior to, similar to, or significantly inferior to the comparison algorithms, with a higher number of "+" symbols indicating greater superiority. Moreover, boldface in the tables is used to indicate optimal results.

5.1. Nondominated front distribution analysis

To highlight MOFS-IH's strengths in convergence and diversity, nine representative low- and high-dimensional datasets were selected. The resultant Pareto fronts, corresponding to the median HV metrics of all algorithms across 30 independent runs, are illustrated in Fig. 4. This result demonstrates the superior balance MOFS-IH achieves between both objectives.

As clearly presented in the figure, the proposed MOFS-IH algorithm outperforms other comparative algorithms by achieving a superior Pareto front. Even on high-dimensional datasets such as 11_Tumors (with 11 classes and a sample-to-feature ratio of just 0.01), MOFS-IH still achieves the highest classification performance while using the fewest features. This is because the population update strategy can utilize historical information generated during iterations to strengthen the algorithm's performance advantages in the late iterative stages and accelerate convergence. Furthermore, the Pareto front identified by MOFS-IH is the most comprehensive compared with those of other comparative algorithms, especially on low-sample datasets such as Breast3 and DLBCL (with fewer than 100 samples). This is attributed to the supplementary selection of low-level features for individuals during the initialization phase, which effectively enhances the overall diversity of the population and thereby enables the algorithm to obtain a more comprehensive Pareto front even on low-sample datasets. Overall, MOFS-IH surpasses the other comparative algorithms by filtering out numerous redundant features and selecting only a small number of key features to achieve high classification accuracy.

5.2. MOFS-IH vs other comparative algorithms

HV and IGD are metrics for evaluating multi-objective feature selection methods. In this experiment, higher HV and lower IGD values indicate that an algorithm has an advantage in addressing such feature selection problems.

Table 3 presents the mean, standard deviation, Wilcoxon and Friedman rank-sum test results for HV, obtained from MOFS-IH and the 6 other algorithms after 30 independent runs on 15 classification datasets. As observed from the table, MOFS-IH significantly surpasses the other 6 algorithms in mean HV. For instance, MOFS-IH achieves a mean HV of 0.86 on the CLL_SUB_111 dataset. In contrast, DDFS (the worst performer) only achieves a mean HV of 0.68, which means an 18% improvement of MOFS-IH over DDFS. Even the top-performing algorithm, MFUBFS, reaches a mean HV of only 0.81 with a 5% improvement of MOFS-IH over it. In addition, MOFS-IH wins 85, draws 5, and loses 0 out of 90 comparisons in the significance tests. Furthermore, MOFS-IH achieves the first rank in the Friedman rank-sum test. These outcomes verify that our population update strategy can fully utilize the valuable information obtained at different iteration points, thereby fulfilling the objectives of each iteration point. Specifically, within the population update strategy, the first iteration point aims to quickly eliminate redundant and ineffective features. The second focuses on selecting higher-quality features, and the third is dedicated to screening features that are more compatible with the current population. These processes complement each other, fostering elite feature combinations in promising regions, and thus enabling the exploration of high-quality feature subsets.

Table 4 compares the average IGD of MOFS-IH and the other 6 algorithms after 30 independent runs. Smaller IGD values indicate that the Pareto front is closer to the true Pareto front. The Friedman rank-sum test further demonstrates that MOFS-IH exhibits significantly better performance. Specifically, MOFS-IH achieves the optimal mean IGD across all datasets when compared with DDFS, MFMOEA, VGS-MOEA, and FG-HFS. In terms of significance tests, MOFS-IH also wins 80, draws 8, and loses 2 across 90 comparisons. Although the mean IGD of MOFS-IH is slightly higher (i.e., worse) than that of IGEA or MFUBFS on two datasets (toxicity and 11_Tumors), this is mainly because both the initialization and population update strategies of MOFS-IH focus on enhancing feature interactions and exploration breadth. It tends to retain slightly more features on datasets with complex feature relationships to maintain combination diversity, which leads to a marginally larger number of selected features. Nevertheless, this does not compromise the overall effectiveness, and the significance tests indicate that these differences are not statistically significant. Overall, among the seven algorithms, the proposed MOFS-IH consistently achieves the most competitive performance in terms of the Pareto front's convergence and diversity. This superiority is attributed to the fact that, after automatic level division, the method not only increases the selection probability

Table 5
Mean time consumed by 7 algorithms on 15 datasets (Unit: s).

DataSet	DDFS	MFMOEA	VGS-MOEA	IGEA	FG-HFS	MFUBFS	MOFS-IH
Yale	97.09	824.37	**29.24**	2913.49	41.06	53.78	46.22
toxicity	82.78	706.77	75.58	437.85	95.95	**50.82**	84.82
SRBCT	**55.25**	198.83	63.21	400.55	75.21	67.61	81.09
GLIOMA	58.15	113.88	64.58	423.01	122.49	**54.46**	88.79
Breast3	**105.11**	382.39	206.47	370.21	194.12	115.94	140.44
Leukemia1	**102.29**	284.35	102.59	397.49	194.41	114.03	124.95
DLBCL	106.49	291.40	**103.56**	640.32	217.76	110.31	133.70
Brain1	138.42	477.97	**124.27**	336.17	290.69	316.67	161.67
Prostate_GE	**158.76**	422.22	174.14	340.43	169.28	175.32	193.54
Prostate6033	167.63	1437.45	143.49	1231.73	222.33	**129.05**	167.03
ALLAML	138.91	430.25	186.03	370.82	334.43	**122.35**	162.70
Adenocarcinoma	191.12	732.03	**178.17**	380.83	515.24	206.72	224.37
Prostate_Tumor	324.89	1325.62	305.28	400.62	829.73	**230.15**	335.17
CLL_SUB_111	**239.57**	1056.83	655.31	369.17	877.72	311.47	294.15
11_Tumors	665.81	3861.20	594.31	**375.64**	1789.39	1242.06	526.11
Friedman's rank	**1**	7	2	6	5	3	4

of high-quality top-level features during initialization but also considers cross-level feature combinations, even those with very low corresponding cross-level probabilities, thereby ensuring both the convergence and diversity of the population.

5.3. Computation time

Training time is a critical metric for evaluating algorithm efficiency. Therefore, to assess the time performance of the proposed elimination method, we compared the average computation time of all algorithms across 15 datasets.

It can be seen from Table 5 that MOFS-IH lacks an advantage in training time, especially compared to VGS-MOEA. This is because VGS-MOEA initializes via grouping, where a single element of an individual solution may correspond to multiple features. This property gives the method an inherent advantage in terms of runtime, especially as the dimension of the feature space increases. However, this approach causes the algorithm to overlook combinations of effective features, preventing it from escaping local optima. As can also be seen from Tables 3 and 4, the results of VGS-MOEA are significantly inferior to the results achieved by MOFS-IH. Notably, the algorithms corresponding to the first and third positions in terms of runtime are decomposition-based algorithms (DDFS and MFUBFS). The main reason is that MFMOEA and MFUBFS do not compute the dominance relationship in their environmental selection mechanisms. In contrast, MOFS-IH needs to calculate the dominance relationship between multiple objective values in its elitist selection strategy. As analyzed in the computational complexity section, the complexity of dominance relationship calculation grows quadratically with population size, which results in longer computation time for MOFS-IH on low-dimensional datasets. Moreover, as the dimension of the feature space increases, this further limits the efficiency of feature level division. However, dominance relationship-based algorithms can better balance multiple conflicting objectives. Therefore, MOFS-IH still chooses NSGA-II as its base algorithm, trading minimal computational time for enhanced performance.

While MOFS-IH lags behind grouping- or decomposition-based algorithms in runtime, it typically finds high-accuracy, low-feature subsets relatively quickly compared to other methods. Taking IGEA (a dominance relationship-based algorithm) as an example, MOFS-IH has a lower runtime than IGEA on most datasets. This is because we have significantly reduced the probability of invalid and redundant features being selected through feature level division. This also indicates that our method still has some advantages in terms of runtime.

5.4. Major component contribution analysis

This section analyzes the contributions of MOFS-IH's two components. First, to verify the uniformity of distribution and convergence of the initialization strategy, we compared the initial population of the variant MOFS-IH^{u-} (without the update strategy) with that generated by probability-based (NSGA-II^{p}) and importance-based initialization (NSGA-II^{i}) methods. Since verifying the quality of initial populations requires focusing not only on their uniformity of distribution and convergence but also on their support for subsequent evolution, we compared, under the same experimental conditions, not only the distributive characteristics and convergence performance of initial populations from a single run of each method in the objective space but also the mean and standard deviation of HV values from 30 independent runs of the three methods on 15 datasets. These analyses were performed to comprehensively verify the effectiveness of MOFS-IH^{u-}. Second, we compared MOFS-IH^{u-} with the full MOFS-IH to verify the effectiveness of the historical information-assisted population update strategy. Furthermore, to verify the optimal selection of the iteration point (i.e., applying the specF library at 3/4 of the total iterations yields the best performance) for the $specF$ library within this strategy, we compared its effects when applied at different iteration points. To ensure the sufficiency of the experiments, all algorithm variants were run for 30 independent runs on 15 datasets. Subsequently, the corresponding performance metrics were computed. The details are elaborated in the following sections.

1) Verification of importance-guided population initialization strategy: Filter-derived feature importance (or its derived values)-based initialization is a conventional practice in the field of feature selection. To verify whether the proposed importance-guided population initialization strategy can generate an initial population with uniform distribution and convergence toward the Pareto front, we compared its generated initial population with those generated by NSGA-II^{p} and NSGA-II^{i} initialization methods. We selected these methods for two reasons. On the one hand, they represent the typical logics of "importance-guided heuristic guidance". On the other hand, by comparing them with these mature yet improvable traditional methods, such comparisons can more directly illustrate the improvement in population quality (in terms of distribution uniformity and convergence) achieved by the proposed strategy. Notably, we choose probability-based and importance-based initialization methods as representatives of initialization methods based on filter-derived feature importance (or its derivatives) because they are consistent with the probability mechanism in our proposed method. Fig. 5 illustrates the differences in convergence of the initial populations generated by the three methods. As can be clearly evident from the figure, the Pareto-optimal solution set of the initial population generated by MOFS-IH^{u-} is more concentrated in the lower-left corner area, showing that this method overall achieves better convergence performance than the other two initialization methods.

To clearly characterize the overall population distribution at the initialization stage, the Mean Nearest Neighbor Distance (MNN) metric is employed for quantitative analysis of initial population distributions

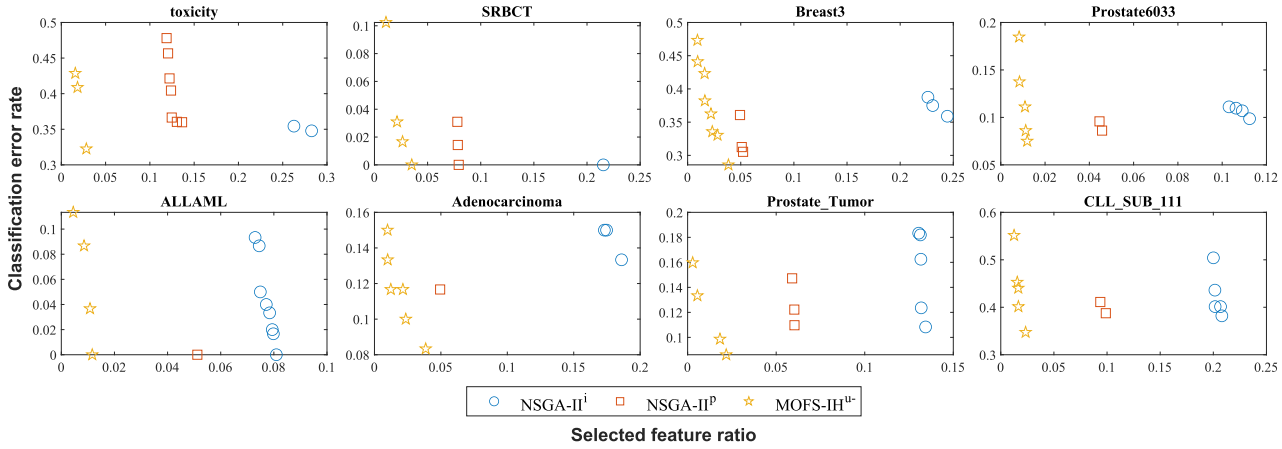

Fig. 5. An example illustrating differences in initial populations: MOFS-IH^{u-} vs. other initialization methods.

Table 6
Mean MNN results of MOFS-IH^{u-} and others' initializations.

Dataset	NSGA-II^{p}	NSGA-II^{i}	MOFS-IH^{u-}
Yale	4.55E-03±6.31E-04(+)	7.69E-03±9.85E-04(+)	**3.13E-02±3.03E-03**
toxicity	7.17E-03±7.64E-04(+)	1.05E-02±1.21E-03(+)	**3.26E-02±2.73E-03**
SRBCT	5.51E-04±1.08E-04(+)	3.05E-03±3.65E-04(+)	**2.80E-02±2.36E-03**
GLIOMA	3.39E-03±8.30E-04(+)	5.13E-03±8.72E-04(+)	**3.00E-02±4.06E-03**
Breast3	4.71E-03±6.64E-04(+)	6.38E-03±8.42E-04(+)	**2.80E-02±2.26E-03**
Leukemia1	1.10E-03±3.26E-04(+)	2.85E-03±4.72E-04(+)	**3.13E-02±4.50E-03**
DLBCL	1.68E-03±5.57E-04(+)	3.34E-03±7.11E-04(+)	**3.03E-02±3.64E-03**
Brain1	2.01E-03±4.30E-04(+)	4.71E-03±8.38E-04(+)	**3.03E-02±3.14E-03**
Prostate_GE	1.69E-03±2.98E-04(+)	2.87E-03±4.85E-04(+)	**2.55E-02±2.95E-03**
Prostate6033	1.18E-03±2.78E-04(+)	3.05E-03±6.80E-04(+)	**2.60E-02±4.07E-03**
ALLAML	1.96E-04±1.70E-04(+)	2.59E-03±5.50E-04(+)	**2.70E-02±2.93E-03**
Adenocarcinoma	8.28E-04±7.02E-04(+)	7.35E-04±6.84E-04(+)	**7.94E-03±2.17E-03**
Prostate_Tumor	1.53E-03±3.20E-04(+)	3.05E-03±4.31E-04(+)	**2.96E-02±3.17E-03**
CLL_SUB_111	4.23E-03±7.05E-04(+)	6.34E-03±8.31E-04(+)	**3.39E-02±2.42E-03**
11_Tumors	1.83E-03±2.28E-04(+)	3.48E-03±4.40E-04(+)	**2.84E-02±2.58E-03**
+/=/-	15/0/0	15/0/0	–
Friedman's rank	3	2	**1**

Table 7
Mean HV results of MOFS-IH^{u-} and others' initializations.

Dataset	NSGA-II^{p}	NSGA-II^{i}	MOFS-IH^{u-}
Yale	6.89E-01±4.47E-02(+)	6.25E-01±3.33E-02(+)	**7.44E-01±3.51E-02**
toxicity	6.86E-01±2.73E-02(+)	5.86E-01±3.80E-02(+)	**7.13E-01±5.40E-02**
SRBCT	**9.73E-01±2.10E-02(=)**	9.33E-01±4.40E-02(+)	9.60E-01±3.89E-02
GLIOMA	8.77E-01±5.76E-02(+)	6.86E-01±8.77E-03(+)	**9.03E-01±6.72E-02**
Breast3	6.46E-01±5.05E-02(+)	5.99E-01±4.60E-02(+)	**7.06E-01±6.53E-02**
Leukemia1	8.65E-01±3.85E-02(+)	8.65E-01±4.32E-02(+)	**9.42E-01±4.19E-02**
DLBCL	9.06E-01±1.90E-02(+)	9.04E-01±3.11E-02(+)	**9.58E-01±5.02E-02**
Brain1	7.80E-01±2.54E-02(+)	6.44E-01±2.02E-02(+)	**8.52E-01±3.99E-02**
Prostate_GE	8.70E-01±2.37E-02(+)	8.65E-01±4.69E-02(+)	**9.48E-01±2.78E-02**
Prostate6033	8.99E-01±2.51E-02(+)	9.21E-01±1.06E-02(+)	**9.54E-01±2.34E-02**
ALLAML	9.18E-01±1.26E-03(+)	9.28E-01±3.72E-02(+)	**9.87E-01±2.71E-02**
Adenocarcinoma	7.92E-01±1.90E-02(+)	7.53E-01±1.02E-02(+)	**8.21E-01±6.30E-02**
Prostate_Tumor	8.29E-01±1.22E-02(+)	8.05E-01±3.53E-02(+)	**8.78E-01±7.04E-02**
CLL_SUB_111	4.56E-01±2.56E-02(+)	4.23E-01±4.26E-02(+)	**7.41E-01±9.43E-02**
11_Tumors	7.40E-01±2.42E-02(+)	6.04E-01±1.99E-02(+)	**7.83E-01±3.84E-02**
+/=/-	14/1/0	15/0/0	–
Friedman's rank	2	3	**1**

generated by three initialization methods across 30 independent runs. This metric refers to the arithmetic mean of distances from each individual in the population to its nearest neighbor in the objective space, with both the horizontal and vertical coordinates in the objective space normalized prior to distance calculation. It is evident from Eq. (11) that the larger the MNN value, the more scattered the population in the objective space. This metric intuitively reflects the global population distribution, with its core idea derived from the distance-based distribution evaluation framework and consistent with the density estimation principle in classical multi-objective algorithms (Chen et al., 2023). Statistics based on nearest neighbor distances are also effectively applied to diversity preservation mechanisms in algorithms (Li et al., 2015), acting as a validated and widely recognized distribution analysis tool.

$$\mathbf{MNN} = \frac{1}{N}\sum_{i=1}^{N}\min_{j=1 j\neq i}^{N}\left\|\boldsymbol{F}(\boldsymbol{x}_i)-\boldsymbol{F}(\boldsymbol{x}_j)\right\|_2 \tag{11}$$

Table 8
Mean HV results of ablation experiments across 15 datasets.

Dataset	NSGA-II	MOFS-IH^{u-}	MOFS-IH
Yale	6.66E-01±3.76E-02(+)	7.44E-01±3.51E-02(+)	**8.20E-01±3.11E-02**
toxicity	5.29E-01±3.46E-02(+)	7.13E-01±5.40E-02(+)	**7.80E-01±2.44E-02**
SRBCT	8.18E-01±1.81E-02(+)	9.60E-01±3.89E-02(+)	**9.99E-01±3.39E-04**
GLIOMA	6.43E-01±1.31E-02(+)	9.03E-01±6.72E-02(+)	**9.91E-01±2.33E-02**
Breast3	4.56E-01±3.26E-02(+)	7.06E-01±6.53E-02(+)	**8.13E-01±4.17E-02**
Leukemia1	6.53E-01±3.88E-02(+)	9.42E-01±4.19E-02(+)	**9.97E-01±1.21E-02**
DLBCL	6.38E-01±3.58E-02(+)	9.58E-01±5.02E-02(+)	**1.00E+00±9.06E-05**
Brain1	5.22E-01±1.33E-02(+)	8.52E-01±3.99E-02(+)	**8.95E-01±2.45E-02**
Prostate_GE	5.12E-01±2.14E-02(+)	9.48E-01±2.78E-02(=)	**9.70E-01±2.46E-02**
Prostate6033	5.18E-01±2.81E-02(+)	9.54E-01±2.34E-02(+)	**9.84E-01±2.28E-02**
ALLAML	5.93E-01±2.43E-02(+)	9.87E-01±2.71E-02(=)	**1.00E+00±1.51E-05**
Adenocarcinoma	5.48E-01±1.00E-02(+)	8.21E-01±6.30E-02(+)	**9.16E-01±3.41E-02**
Prostate_Tumor	5.25E-01±1.99E-02(+)	8.78E-01±7.04E-02(+)	**9.47E-01±2.29E-02**
CLL_SUB_111	3.06E-01±3.49E-02(+)	7.41E-01±9.43E-02(+)	**8.61E-01±4.85E-02**
11_Tumors	4.91E-01±1.60E-02(+)	7.83E-01±3.84E-02(+)	**8.30E-01±2.02E-02**
+/=/-	15/0/0	13/2/0	–
Friedman's rank	3	2	**1**

Table 9
Mean IGD results of ablation experiments across 15 datasets.

Dataset	NSGA-II	MOFS-IH^{u-}	MOFS-IH
Yale	1.13E-01±2.30E-02(+)	6.89E-02±1.94E-02(+)	**3.70E-02±1.17E-02**
toxicity	1.22E-01±2.73E-02(+)	1.15E-01±5.40E-02(+)	**5.57E-02±2.12E-02**
SRBCT	6.27E-02±1.65E-02(+)	2.27E-02±2.00E-02(+)	**1.63E-02±1.38E-02**
GLIOMA	7.30E-02±5.03E-02(+)	6.40E-02±5.27E-02(+)	**1.08E-02±1.52E-02**
Breast3	1.67E-01±4.99E-02(+)	1.20E-01±6.29E-02(+)	**3.96E-02±2.19E-02**
Leukemia1	8.74E-02±3.44E-02(+)	3.46E-02±3.09E-02(+)	**6.84E-03±1.16E-02**
DLBCL	6.51E-02±8.77E-03(+)	4.39E-02±2.54E-02(+)	**2.33E-02±1.81E-02**
Brain1	9.03E-02±8.02E-03(+)	4.78E-02±1.97E-02(+)	**3.60E-02±1.80E-02**
Prostate_GE	8.52E-02±1.16E-02(+)	4.64E-02±2.04E-02(+)	**3.55E-02±1.75E-02**
Prostate6033	6.66E-02±2.28E-02(+)	2.47E-02±1.92E-02(+)	**9.21E-03±1.14E-02**
ALLAML	6.85E-02±3.08E-04(+)	1.35E-02±2.70E-02(+)	**2.34E-04±1.49E-04**
Adenocarcinoma	7.53E-02±1.75E-02(+)	7.02E-02±4.39E-02(+)	**1.40E-02±1.37E-02**
Prostate_Tumor	1.02E-01±1.10E-02(+)	8.52E-02±5.46E-02(+)	**2.28E-02±1.25E-02**
CLL_SUB_111	3.09E-01±2.57E-02(+)	9.64E-02±6.43E-02(+)	**3.80E-02±1.65E-02**
11_Tumors	7.83E-02±9.32E-03(+)	4.93E-02±1.33E-02(+)	**2.65E-02±1.10E-02**
+/=/-	15/0/0	15/0/0	–
Friedman's rank	3	2	**1**

Table 10
Mean HV values of $specF$ at different iteration points.

Dataset	MOFS-IH$^{0.25T}$	MOFS-IH$^{0.5T}$	MOFS-IH
Yale	7.72E-01±2.95E-02(+)	7.87E-01±3.06E-02(+)	**8.20E-01±3.11E-02**
toxicity	7.34E-01±3.34E-02(+)	7.62E-01±3.48E-02(+)	**7.80E-01±2.44E-02**
SRBCT	9.80E-01±2.93E-02(+)	9.87E-01±2.22E-02(+)	**9.99E-01±3.39E-04**
GLIOMA	9.13E-01±5.74E-02(+)	9.58E-01±4.56E-02(+)	**9.91E-01±2.33E-02**
Breast3	7.46E-01±3.95E-02(+)	7.75E-01±3.89E-02(+)	**8.13E-01±4.17E-02**
Leukemia1	9.49E-01±4.67E-02(+)	9.74E-01±3.68E-02(+)	**9.97E-01±1.21E-02**
DLBCL	9.92E-01±1.89E-02(+)	9.99E-01±5.91E-04(+)	**1.00E+00±9.06E-05**
Brain1	8.37E-01±2.83E-02(+)	8.72E-01±2.91E-02(=)	**8.95E-01±2.45E-02**
Prostate_GE	9.45E-01±2.29E-02(+)	9.58E-01±2.09E-02(+)	**9.70E-01±2.46E-02**
Prostate6033	9.58E-01±2.82E-02(+)	9.74E-01±2.41E-02(+)	**9.84E-01±2.28E-02**
ALLAML	9.80E-01±4.38E-02(+)	9.89E-01±2.52E-02(+)	**1.00E+00±1.51E-05**
Adenocarcinoma	8.68E-01±5.02E-02(+)	8.72E-01±4.17E-02(+)	**9.16E-01±3.41E-02**
Prostate_Tumor	8.85E-01±3.63E-02(+)	9.33E-01±3.44E-02(=)	**9.47E-01±2.29E-02**
CLL_SUB_111	7.94E-01±4.73E-02(+)	8.54E-01±4.22E-02(=)	**8.61E-01±4.85E-02**
11_Tumors	7.61E-01±2.77E-02(+)	8.23E-01±2.20E-02(=)	**8.30E-01±2.02E-02**
+/=/-	15/0/0	11/4/0	–
Friedman's rank	3	2	**1**

where N denotes the population size, $\boldsymbol{F}(\boldsymbol{x}_i)$ denotes the mapping value of the ith individual $\boldsymbol{x}_i$ in the objective space.

As seen from Table 6, the distribution of MOFS-IH^{u-} in the objective space is significantly superior to that of the other two initialization methods. The above results indicate that MOFS-IH^{u-} ensures comprehensive coverage of the solution space during the construction of the initial population and directs more individuals to converge toward the target optimization region (i.e., the dominant region denoted by the lower-left corner), thereby laying a better foundation for subsequent algorithm iterations.

Table 7 summarizes the performance of three distinct initialization methods across 15 datasets. The data in the table clearly shows that MOFS-IH^{u-} exhibits outstanding comprehensive performance: on 14 datasets, its performance is significantly superior to that of the other two comparative methods. Only on the SRBCT dataset, its performance is slightly inferior to probability-based initialization, but

the difference between them is not statistically significant. This result fully indicates that, relying on its importance-guided population construction mechanism, MOFS-IH^{u-} can more efficiently explore the feature solution space and accurately mine potential effective feature combinations.

2) Verification of historical information-assisted population update strategy: The historical information-assisted population update strategy aims to enhance effective interaction among potential features from different stages, thereby helping the algorithm accelerate convergence and escape local optima. We compared the proposed historical information-assisted population update mechanism with NSGA-II and MOFS-IH^{u-}, with the experimental results shown in Tables 8 and 9. It is obvious from the tables that after incorporating the population update mechanism, MOFS-IH achieves the best overall performance on all 15 datasets. Specifically, MOFS-IH significantly outperforms NSGA-II and MOFS-IH^{u-} in terms of HV on 13 of the 15 datasets, and in terms of IGD on all 15 datasets. This result indicates that the proposed historical information-assisted population update strategy significantly improves the algorithm's performance. This is because the strategy can acquire effective features from historical iteration processes to help the population escape local optima. Furthermore, it adopts different objectives and processing methods according to different iteration stages, thereby enhancing the potential for feature interaction.

Table 10 verifies the optimality of the iteration point selection for the specF library within this strategy. It can be observed that applying the specF library at $3/4$ of the total iterations (i.e., $3/4T$) yields significantly better performance than at $1/4T$ or $1/2T$. This indicates that dividing the entire iterative process into three interconnected equal-length stages allows the population to fully leverage historical information in the later stage to explore richer feature combinations; meanwhile, the stage-specific strategies effectively fulfill our original design of distinct functionalities for each stage.

6. Conclusion

This paper proposes a multi-objective feature selection method integrating importance-guided population initialization and historical information-assisted update mechanisms. First, an automatic feature level division method is designed to extract potential valid dataset information, and a cross-level feature combination-based initialization method is developed to enhance initial population quality. Second, a historical information acquisition method based on feature actual performance is proposed to capture effective and potential features during iteration. To utilize these features effectively, the whole iteration process is split into distinct stages, based on which a population update strategy is developed to accelerate convergence and escape local optima.

Performance is validated by comparison with six advanced multi-objective feature selection algorithms on 15 benchmark datasets with HV and IGD indicators. The experimental results show that MOFS-IH performs better than the competing algorithms in these two metrics on most datasets. This is attributed to the importance-guided initialization method, which generates a uniformly distributed initial population converging to the Pareto front, laying a solid foundation for subsequent iterations. Additionally, the historical information-assisted update strategy leverages advantages of different iterative stages to form elite feature combinations (with few features) in promising directions, facilitating superior subset discovery.

Despite its advantages, MOFS-IH still has potential for further exploration. First, domain expertise can be introduced to prescreen features, narrowing the search space and accelerating convergence. Second, the level probability information derived in this work can serve as intermediate knowledge and be transferred to feature selection tasks in similar data domains, enhancing the cross-task adaptability of the algorithm.

CRediT authorship contribution statement

Kunjie Yu: Conceptualization, Methodology, Software, Formal analysis, Validation, Writing – review & editing; **Wen An:** Methodology, Software, Validation, Writing – original draft, Writing – review & editing; **Ke Chen:** Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – review & editing; **Fangfang Zhang:** Supervision, Writing – review & editing.

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.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62476254, 62576321, and 62206255), the Key Research and Development Program of Henan (Grant No. 251111113900), and the Natural Science Foundation of Henan Province (Grant No. 252300421501).

Data availability

Data will be made available on request.

References

Azhagusundari, B., Thanamani, A. S. et al. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *2*(2), 18–21.

Bosman, P. A. N., & Thierens, D. (2003). The balance between proximity and diversity in multiobjective evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, *7*(2), 174–188.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chen, K., Liu, K., Wang, P., Liang, J., Yang, Z., & Yu, K. (2026a). An efficient evolutionary feature selection algorithm with divide-and-conquer strategy for classification. *Expert Systems with Applications*, *302*, 130416.

Chen, K., Xue, B., Zhang, M., & Zhou, F. (2021). Correlation-guided updating strategy for feature selection in classification with surrogate-assisted particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, *26*(5), 1015–1029.

Chen, K., Zhou, F.-Y., & Yuan, X.-F. (2019). Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. *Expert Systems with Applications*, *128*, 140–156.

Chen, X., Huang, Z., Tian, F., Zhang, R., Zhou, X., Gao, Q., & Zhang, Y. (2026b). A knowledge-guided feature selection method for lung cancer immunotherapy response prediction via pathology image analysis. *Expert Systems with Applications*, *307*, 131084.

Chen, X., Tian, Z., & Rao, M. (2023). Block feature selection based on NSGA-II applied to fault diagnosis of gearboxes. *Advanced Engineering Informatics*, *57*, 102095.

Cheng, F., Cui, J., Wang, Q., & Zhang, L. (2022). A variable granularity search-based multiobjective feature selection algorithm for high-dimensional data classification. *IEEE Transactions on Evolutionary Computation*, *27*(2), 266–280.

Das, I. (1999). On characterizing the "knee" of the pareto curve based on normal-boundary intersection. *Structural Optimization*, *18*(2), 107–115.

De La Iglesia, B. (2013). Evolutionary computation for feature selection in classification problems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(6), 381–407.

Gupta, S., & Gupta, S. (2024). Fitness and historical success information-assisted binary particle swarm optimization for feature selection. *Knowledge-Based Systems*, *306*, 112699.

Huang, J., Chen, Y., Heidari, A. A., Liu, L., Chen, H., & Liang, G. (2025a). DPDEPSO: A particle swarm optimization for balancing different objectives in multi-objective feature selection. *Expert Systems with Applications*, *293*, 128709.

Huang, J., Chen, Y., Heidari, A. A., Liu, L., Chen, H., & Liang, G. (2025b). DPDEPSO: A particle swarm optimization for balancing different objectives in multi-objective feature selection. *Expert Systems with Applications*, *293*, 128709.

Jiao, R., Nguyen, B. H., Xue, B., & Zhang, M. (2023). A survey on evolutionary multiobjective feature selection in classification: Approaches, applications, and challenges. *IEEE Transactions on Evolutionary Computation*, *28*(4), 1156–1176.

Jiao, R., Xue, B., & Zhang, M. (2022a). Benefiting from single-objective feature selection to multiobjective feature selection: A multiform approach. *IEEE Transactions on Cybernetics*, *53*(12), 7773–7786.

Jiao, R., Xue, B., & Zhang, M. (2022b). Solving multiobjective feature selection problems in classification via problem reformulation and duplication handling. *IEEE Transactions on Evolutionary Computation*, *28*(4), 846–860.

Jin, Y., & Sendhoff, B. (2008). Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(3), 397–415.

Kamalov, F., Sulieman, H., Alzaatreh, A., Emarly, M., Chamlal, H., & Safaraliev, M. (2025). Mathematical methods in feature selection: A review. *Mathematics*, *13*(6), 996.

Kurgan, L. A., & Cios, K. J. (2004). Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, *16*(2), 145–153.

Li, A.-D., He, Z., & Zhang, Y. (2016). Bi-objective variable selection for key quality characteristics selection based on a modified NSGA-II and the ideal point method. *Computers in Industry*, *82*, 95–103.

Li, G., Chen, Z., Zhou, Y., Huang, Z., & Xia, X. (2026a). Towards large-scale multi-objective feature selection: A two-stage evolutionary algorithm guided by dual feature weightings. *Expert Systems with Applications*, *298*, 129823.

Li, H., & Dang, Q. (2026). Multimodal multi-objective evolutionary algorithm assisted by graph neural networks based selection mechanism. *Expert Systems with Applications*, *310*, 131328.

Li, L., Zhang, Y., Xiao, Z., Lin, Q., Wang, X., He, X., & Zhong, M. (2026b). An adaptive length-variation based evolutionary multitasking algorithm for feature selection of high-dimensional classification. *Expert Systems with Applications*, *295*, 128874.

Li, M., Yang, S., & Liu, X. (2015). Pareto or non-pareto: Bi-criterion evolution in multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, *20*(5), 645–665.

Li, W., & Chai, Z. (2024). MPEA-FS: A decomposition-based multi-population evolutionary algorithm for high-dimensional feature selection. *Expert Systems with Applications*, *247*, 123296.

Liang, J., Zhang, Y., Chen, K., Qu, B., Yu, K., Yue, C., & Suganthan, P. N. (2024). An evolutionary multiobjective method based on dominance and decomposition for feature selection in classification. *Science China Information Sciences*, *67*(2), 120101.

Liang, J., Zhang, Y., Qu, B., Chen, K., Yu, K., & Yue, C. (2023). A multiform optimization framework for multiobjective feature selection in classification. *IEEE Transactions on Evolutionary Computation*, *28*(4), 1024–1038.

Liang, J., Zhang, Y.-Y., Qu, B., Chen, K., Yu, K., & Yue, C. (2025). A multiform framework for multiobjective feature selection in unbalanced classification: Combining oversampling and cost-sensitive learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *55*(8), 5717–5729.

Ramasamy, M., & Meena Kowshalya, A. (2022). Information gain based feature selection for improved textual sentiment analysis. *Wireless Personal Communications*, *125*(2), 1203–1219.

Singh, U., & Singh, S. N. (2017). Optimal feature selection via NSGA-II for power quality disturbances classification. *IEEE Transactions on Industrial Informatics*, *14*(7), 2994–3002.

Siswantoro, J., Wijaya, J. N., Prasetyo, V. R., & Arwoko, H. (2026). A hybrid binary grey wolf and whale optimization method for feature selection in classification tasks. *Expert Systems with Applications*, *321*, 132277.

Wang, P., Xue, B., Liang, J., & Zhang, M. (2021). Multiobjective differential evolution for feature selection in classification. *IEEE Transactions on Cybernetics*, *53*(7), 4579–4593.

Wang, X., Shangguan, H., Huang, F., Wu, S., & Jia, W. (2024). MEL: Efficient multi-task evolutionary learning for high-dimensional feature selection. *IEEE Transactions on Knowledge and Data Engineering*, *36*(8), 4020–4033.

Wang, X., Wang, Y., Wong, K.-C., & Li, X. (2022). A self-adaptive weighted differential evolution approach for large-scale feature selection. *Knowledge-Based Systems*, *235*, 107633.

Wang, Z., Shao, L., Heidari, A. A., Wang, M., & Chen, H. (2025). Prior knowledge evaluation and emphasis sampling-based evolutionary algorithm for high-dimensional medical data feature selection. *Expert Systems with Applications*, *273*, 126737.

While, L., Hingston, P., Barone, L., & Huband, S. (2006). A faster algorithm for calculating hypervolume. *IEEE Transactions on Evolutionary Computation*, *10*(1), 29–38.

Xia, M., Zhang, L., Li, K., & Cheng, F. (2026). ODS-EA: An objective to decision space-based evolutionary algorithm for high-dimensional feature selection. *Expert Systems with Applications*, *303*, 130560.

Xu, H., Xue, B., & Zhang, M. (2020a). A duplication analysis-based evolutionary algorithm for biobjective feature selection. *IEEE Transactions on Evolutionary Computation*, *25*(2), 205–218.

Xu, H., Xue, B., & Zhang, M. (2020b). Segmented initialization and offspring modification in evolutionary algorithms for bi-objective feature selection. In *Proceedings of the 2020 genetic and evolutionary computation conference* (pp. 444–452).

Xu, Z., Yang, F., Tang, C., Wang, H., Wang, S., Sun, J., & Zhang, Y. (2024). FG-HFS: A feature filter and group evolution hybrid feature selection algorithm for high-dimensional gene expression data. *Expert Systems with Applications*, *245*, 123069.

Xue, Y., Zhu, H., & Neri, F. (2023). A feature selection approach based on NSGA-II with relieff. *Applied Soft Computing*, *134*, 109987.

Yin, Y., Jang-Jaccard, J., Xu, W., Singh, A., Zhu, J., Sabrina, F., & Kwak, J. (2023). IGRF-RFE: A hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. *Journal of Big Data*, *10*(1), 15.

Yu, F., Guan, J., Wu, H., Wang, H., & Ma, B. (2025). Multi-population differential evolution approach for feature selection with mutual information ranking. *Expert Systems with Applications*, *260*, 125404.

Zhang, B., Wang, Z., Li, H., Lei, Z., Cheng, J., & Gao, S. (2024). Information gain-based multi-objective evolutionary algorithm for feature selection. *Information Sciences*, *677*, 120901.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., & Da Fonseca, V. G. (2003). Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, *7*(2), 117–132.