Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085308

ARTICLE

Five-Region Rough Isolation Forest with Multi-Strategy Feature Optimization for High-Dimensional Anomaly Detection

Dong-Fang Wu[1,2], Jiaojiao Deng[1,2], Zhiwei Ye[1,2,*], Rong Gao[1,2], Fan Ma[1,2] and Dingfeng Song[1,2]

[1]School of Computer Science and Artificial Intelligence, Hubei University of Technology, Wuhan, China
[2]Hubei Provincial Key Laboratory of Green Intelligent Computing Power Network, Hubei University of Technology, Wuhan, China
*Corresponding Author: Zhiwei Ye. Email: hgcsyzw@hbut.edu.cn

Received: 08 May 2026; Accepted: 10 July 2026; Published: 13 August 2026

ABSTRACT: Anomaly detection is used to identify data points deviating from normal patterns and plays a significant role in fields such as fault diagnosis, biomedicine, and cybersecurity. However, anomaly detection tasks in practical applications often involve high-dimensional data which presents challenges such as severe feature redundancy, hidden anomaly patterns, and difficulty in capturing uncertainty. These issues make it difficult to effectively identify anomalies, thereby limiting the model's discriminative power and stability. To address these challenges, we propose a high-dimensional anomaly detection model, MSFO-EIF, integrating multi-strategy feature optimization with rough set modeling. First, in the feature space optimization stage, we incorporate label information and employ a Maximum Relevance Minimum Redundancy (mRMR) pre-screening and an uncertainty clustering mechanism to filter and structurally organize the original features, thereby reducing redundancy and retaining key discriminative information. Second, in the feature selection phase, we construct a multi-strategy co-evolution mechanism to optimize feature subsets within a label-guided search space, thereby mitigating the risk of local optima. Finally, in the anomaly detection stage, multi-region partitioning and rough set concepts are introduced to characterize the sample distribution structure at a fine granularity, thereby enhancing the ability to identify boundary samples and weak anomalies. Experimental results show that the proposed model outperforms other anomaly detection models, including KNN, LOF, IBBA-EIF, and RRSM, on multiple high-dimensional datasets. Specifically, MSFO-EIF achieves the improvements of 1.68%, 1.33%, 1.69%, 2.67% and 1.87% in accuracy, precision, F1-score, AUC and AUC-PR, respectively, highlighting its superior detection performance and robustness.

KEYWORDS: High-dimensional anomaly detection; isolated forest; hybrid breeding optimization algorithm; feature selection; rough set

1 Introduction

With the rapid development of next-generation information technologies, the demand for anomaly detection in fields such as biomedicine and text analysis is growing steadily [1]. Anomaly detection is critical for identifying abnormal system behavior and potential risks. The data involved in anomaly detection tasks is predominantly high-dimensional, and high-dimensional data typically suffers from severe feature redundancy, strong noise interference, and sparse distribution. These issues can lead to the curse of dimensionality, reducing the separability of anomalous patterns and increasing model complexity [2]. How to mitigate the adverse effects of high-dimensional data through effective feature processing and modeling has become a critical factor in improving anomaly detection performance [3]. Consequently, constructing

effective feature representations and enhancing anomaly detection performance in high-dimensional data environments has emerged as one of the core research directions [4].

Anomaly detection aims to identify outlier samples in data that deviate significantly from normal patterns [5]. Early methods were primarily based on statistical modeling and distance metrics [6–8], achieving anomaly identification by calculating the distance or similarity between samples. Subsequently, density-based methods [9] enhanced the ability of local anomaly detection by analyzing data distributions and classifying samples in low-density regions as anomalies. Furthermore, the method based on fuzzy information granules [10,11] converts raw data into information granules by constructing fuzzy relationships and uses uncertainty metrics to describe the degree of anomaly, offering certain advantages in handling complex data. In recent years, methods based on isolation mechanisms [12,13] have gradually garnered attention. Isolation forests achieve anomaly isolation by randomly partitioning the feature space, and various improvement strategies have been derived from this approach. These include introducing virtual nodes to enhance structural dynamism [14], combining deep neural networks for nonlinear mapping, and improving model expressiveness through functional partitioning. These methods have further expanded the application of isolation mechanisms in complex data scenarios. Although existing methods continue to evolve, anomaly detection still faces several challenges in high-dimensional, complex data environments.

(1) In feature selection, the high-dimensional feature space grows exponentially, resulting in a vast search space. Traditional optimization methods struggle to efficiently identify optimal feature subset [15]. In high-dimensional data, there is actually a large amount of redundant and irrelevant features. This not only increases the computational complexity but also weakens the ability to capture key discriminative information. Different features exhibit varying importance across different subspaces, which further complicates feature modeling [16]. Relying solely on simple evaluation criteria or local search strategies makes it difficult to comprehensively capture feature dependencies. And it often leads to the omission of important features or the repeated selection of redundant ones, thereby compromising the discriminative performance of the feature subset. Furthermore, in the feature selection process, coordinating global search with local optimization remains challenging [17]. Global exploration helps identify potential high-quality solutions but incurs high computational costs. While local search accelerates convergence but is prone to getting trapped in local optima, resulting in unstable search results. Additionally, the optimal feature subsets often vary across different data distributions, making feature selection sensitive to initialization and search strategies, which increases algorithm design complexity.

(2) In the context of anomaly detection, anomalous samples in high-dimensional data are typically scarce and unevenly distributed. Anomalous patterns often manifest only in specific feature subspaces, making the differences between anomalous and normal samples more subtle. Modeling directly in the original feature space is prone to interference from noise and redundant features, which weakens the model's sensitivity to anomalous features and reduces detection performance [18]. Additionally, high-dimensional data exhibit complex structures. Nonlinear couplings may exist across dimensions, making it difficult to accurately capture anomaly patterns using a single modeling approach. Furthermore, uncertainties such as ambiguous boundary samples, class overlap, and noise interference also affect the stability and reliability of anomaly classification [19]. Existing methods largely rely on a single representation space or a fixed modeling framework, lacking the ability to jointly model structural information and uncertainty, and thus struggle to fully uncover potential discriminative features. Furthermore, in complex scenarios, anomaly patterns may also exhibit diversity and dynamic changes; some anomalies manifest as gradual or weak anomalies, placing higher demands on the model's expressive and adaptive capabilities.

To address the aforementioned challenges, this paper proposes a modeling framework that integrates high-dimensional feature selection with anomaly detection, MSFO-EIF. First, during the feature selection stage, label information is incorporated into a pre-screening strategy based on Maximum Relevance Minimum Redundancy (mRMR) to reduce the search space. Subsequently, based on symmetric uncertainty, a feature grouping mechanism is constructed to model feature correlations and reduce redundant interference. Building on this foundation, we designed an improved hybrid breeding optimization that integrates Latin Hypercube Sampling (LHS), elite backpropagation, roulette wheel selection, and Grey Wolf Optimization (GWO). This approach enhances global search capabilities and population diversity in high-dimensional spaces. Furthermore, during the anomaly detection phase, the optimized feature subset is input into an extended isolation forest, and uncertainty is modeled using rough set theory. This enhances the model's ability to capture complex data distributions, enabling high-dimensional anomaly detection.

The main contributions of this paper are as follows:

1. We propose an improved hybrid breeding optimization algorithm integrating roulette wheel selection, grey wolf-guided updating, and elite reverse learning, thereby enhancing global optimization capability and convergence stability in high-dimensional search spaces.
2. We construct a high-dimensional feature selection method based on feature grouping. By combining mRMR pre-screening with a symmetric uncertainty grouping strategy, we effectively reduce feature redundancy and improve the quality of the feature subsets.
3. We propose an extended Isolation Forest model for high-dimensional anomaly detection that integrates rough sets, incorporating a five-region uncertainty partitioning mechanism to enhance the accuracy and robustness of anomaly detection.
4. Experimental results demonstrate that the proposed model surpasses existing mainstream models in detection performance and stability.

2 Related Work

2.1 High-Dimensional Data Anomaly Detection

Anomalous patterns often exhibit strong sparsity, uneven distribution, and overlap with normal samples. This makes it challenging to accurately characterize anomaly boundaries and to improve the sensitivity and robustness of detection. In early research, anomaly detection relied on statistical modeling and distance-based methods. Breunig et al. [9] proposed the Local Outlier Factor (LOF), identifying local anomalous points by characterizing the density deviation of a sample's relative region via local reachable density. However, such methods suffer from the "distance concentration phenomenon" in high-dimensional spaces, resulting in reduced distance variations between samples and diminished detection performance.

With the advancement of machine learning, researchers have begun to utilize models to learn data distributions in order to enhance anomaly detection capabilities. One-Class SVM [20] identifies anomalous samples by learning the minimum enclosing boundary of normal samples in a high-dimensional feature space. Sakurada and Yairi [21] proposed an autoencoder-based anomaly detection method that identifies anomalies by minimizing the reconstruction error of normal samples. Additionally, DAGMM [22] combines an autoencoder with a Gaussian mixture model to perform both reconstruction and density estimation in the latent space, thereby improving anomaly detection performance. Although the above methods have advantages in nonlinear modeling, they typically rely on large amounts of training data and are prone to overfitting when outliers are scarce, resulting in poor interpretability.

In addition, some researchers have explored the integration of isolation mechanisms with representation learning. Deep Isolation Forest [23] incorporates nonlinear mappings from deep neural networks and

combines isolation mechanisms to achieve anomaly scoring. To improve model interpretability, Arcudi et al. [24] proposed a function-based isolated forest, which generalizes random threshold splitting into real-valued function splitting, enabling the tree structure to better capture complex segmentation boundaries. Furthermore, to address the issue of abnormal definition relying on domain knowledge, the Bayesian active learning isolated forest [25] combines a few annotations with an active learning strategy to select key samples for feedback-based updates, thereby improving detection performance with limited annotations.

Furthermore, given that anomalies are inherently uncertain and ambiguous, some researchers have incorporated rough set theory into anomaly detection. Song et al. [26] constructed tolerance relations based on multi-valued information systems and designed outlier factors using rough set theory to detect anomalies in uncertain data. Hu et al. [27] constructed fuzzy similarity relations and fuzzy rough approximation models to select informative subspaces and detect anomalies under uncertain nominal data.

Although existing methods have improved the performance of isolated forests, their binary partitions still struggle to capture the anomaly hierarchy and the boundary uncertainty. Thus, it is necessary to introduce finer-grained partitions to enhance the ability to represent complex anomalies.

2.2 Anomaly Detection in High-Dimensional Data Based on Feature Selection

High-dimensional data often contain redundant features and noise, and performing anomaly detection directly in the original feature space can reduce discriminative power. Therefore, the key challenge in anomaly detection is how to enhance the separability of anomalous patterns through feature selection and subspace modeling. Early research primarily relied on statistical correlations or information-theoretic criteria for feature selection. Peng et al. [28] proposed the mRMR method, which maximizes the correlation between features and the target and minimizes redundancy among features, thereby obtaining a feature subset with strong discriminative power. However, such methods typically rely on greedy strategies and struggle to capture the complex interdependencies among high-dimensional features.

To enhance feature representation capabilities in unsupervised settings, researchers have further explored feature subspace learning methods for anomaly detection. HiCS [29] identifies feature combinations that deviate significantly from the normal distribution by searching for high-contrast subspaces, thereby improving the performance of high-dimensional anomaly detection. Subsequently, some studies have begun to integrate feature selection with specific anomaly detection tasks. For example, Najarbashi [30] proposed a feature selection method based on the Squirrel optimization algorithm for global feature screening and combined with multi-classification models to enhance financial anomaly detection capabilities. While Time-EAPCR-T [31] optimizes temporal modeling and feature fusion for industrial multi-source heterogeneous data, thereby improving the ability to represent anomalous features. However, such methods focus on subspace modeling or feature fusion but lack a unified optimization framework, resulting in feature selection that relies on heuristic learning. Consequently, it is difficult to obtain globally optimal and stable discriminative feature combinations in high-dimensional, complex spaces.

In recent years, evolutionary optimization algorithms have been widely adopted in feature selection processes to enhance the ability to search for feature subsets. MGA-IDS [32] improves genetic algorithms to optimize feature subsets and combined multi-classification models to enhance intrusion detection performance and generalization capabilities. Furthermore, to address the trade-off between search efficiency and solution quality in high-dimensional spaces, researchers have introduced more refined search strategies and multi-objective optimization mechanisms. For example, Li et al. [33] proposed a dual-objective evolutionary algorithm for binary individual search, which enhances feature discrimination and reduces dimensionality through pre-screening and adaptive operations, achieving a more optimal representation for anomaly detection. Building on this, recent works have explored multi-task and adaptive optimization frameworks.

AMFEA [34] proposes adaptive multi-factor evolutionary feature selection, which shares information across related tasks through local search and dynamic knowledge transfer.

Although the above methods have made progress in feature selection and search strategies, they primarily focus on subset optimization and efficiency, and lack a joint modeling of feature structure and anomaly uncertainty. This limits their applicability in complex industrial scenarios.

3 Proposed Method

3.1 Overview of the Proposed Framework

As shown in Fig. 1, the proposed model consists of three modules: high-dimensional feature space optimization, multi-strategy feature selection, and rough-set-integrated anomaly detection, following a progressive "feature compression–subset optimization–anomaly discrimination" framework. First, mRMR with label information is used to preselect features, reducing irrelevant and redundant interference, while symmetric uncertainty is adopted for structured feature grouping. Then, in the feature selection module, a multi-strategy cooperative evolutionary algorithm optimizes feature subsets in a label-informed search space, where LHS, elite opposition-based learning, roulette wheel selection, GWO, and entropy-based simulated annealing are combined to balance global exploration and local exploitation. Finally, the optimized feature subset is fed into a rough-set-integrated Extended Isolation Forest anomaly detection module, where rough set approximation regions are introduced to characterize sample distributions through multi-region partitioning, and weighted path length is used to calculate anomaly scores. Since five-region partitioning depends on the distribution and boundary structure of the selected feature space, prior feature optimization reduces the influence of noisy dimensions and improves modeling of complex distributions and uncertain data.

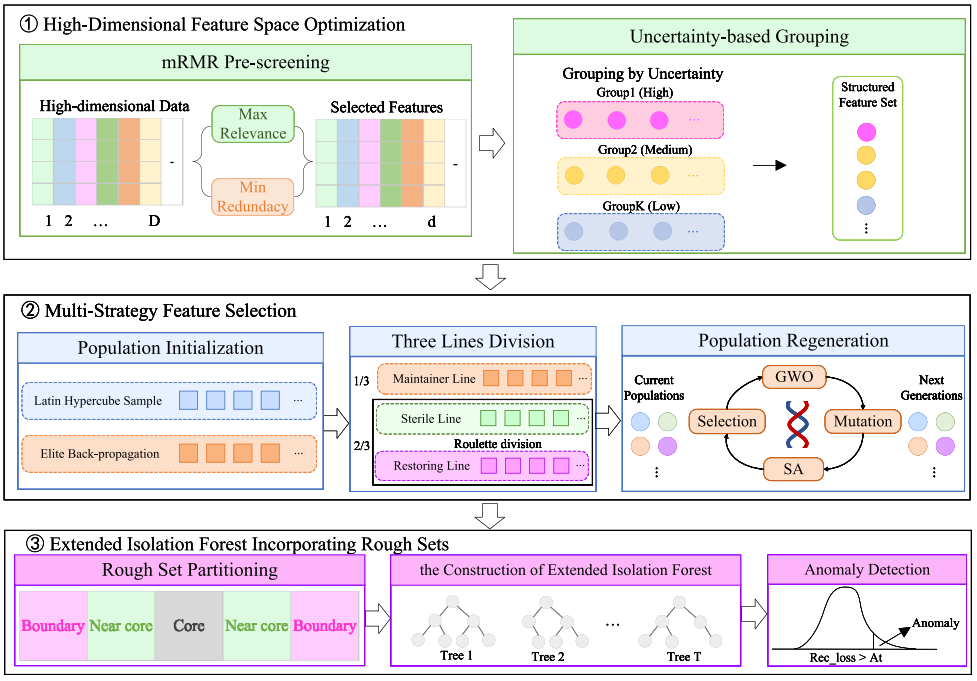


Figure 1: The overall structure of MSFO-EIF model.

3.2 High-Dimensional Feature Space Optimization Based on mRMR and Feature Clustering

High-dimensional biomedical data are characterized by high dimensionality, small sample sizes, and severe feature redundancy. Direct feature selection in the original space enlarges the search space and may repeatedly select redundant features, reducing subset effectiveness. Therefore, we optimize the feature space before formal feature selection. First, mRMR is used to screen candidate features that are relevant to the class and have low redundancy, reducing the influence of irrelevant features on subsequent relevance evaluation. Second, symmetric uncertainty is introduced to evaluate the importance and dependency relationships of the pre-screened features, thereby constructing multiple differentiated feature subspaces and providing more complementary candidate spaces for subsequent feature subset search.

3.2.1 High-Dimensional Feature Pre-Screening Based on mRMR

To improve the efficiency of subsequent feature selection algorithms and reduce the search space, we introduce the mRMR criterion. It ensures strong relevance between features and class labels while minimizing redundancy among the selected features. Specifically, let the original feature set be $F = \{f_1, f_2, \cdots, f_D\}$, the class labels be $c \in C$, and the set of candidate features be $S \subseteq$ F (of size k). For each candidate feature f_i, we compute its mRMR score as follows:

$$J(f_i) = I(f_i; c) - \frac{1}{|S|} \sum_{f_j \in S} I(f_i; f_j) \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information. The first term measures the correlation between feature f_i and label c, while the second term measures the average redundancy between f_i and the previously selected features. Features are selected based on the magnitude of $J(f_i)$, resulting in a set of candidate features F'.

$$F' = \{f'_1, f'_2, \ldots, f'_d\}, d \ll D \tag{2}$$

where D denotes the number of original features, and d denotes the dimensionality of the candidate feature set after mRMR pre-screening. This process can preliminarily remove weakly correlated features and some redundant features, thereby reducing the search dimension for subsequent optimization algorithms.

3.2.2 Feature Clustering Based on Symmetric Uncertainty

After mRMR pre-screening, the candidate feature set F' may still contain local redundancy. To further characterize feature importance, we employ Symmetric Uncertainty (SU) to measure the degree of association between feature f'_i and label c—that is, the feature's importance—as shown in Eqs. (3) and (4):

$$I(f'_i; c) = \sum_{f'_i \in F'} \sum_{c \in C} p(f'_i, c) \cdot \log \frac{p(f'_i, c)}{p(f'_i)p(c)} \tag{3}$$

$$SU(f'_i, c) = 2\frac{I(f'_i; c)}{H(f'_i) + H(c)} \tag{4}$$

where $p(f'_i)$ and $p(c)$ denote the marginal probabilities of the feature value and the class label, respectively, and $p(f'_i, c)$ denotes their joint probability. $I(f'_i; c)$ denotes the mutual information between feature f'_i and class label c, $H(f'_i)$ and $H(c)$ represent their information entropies, respectively. $SU(f'_i, c) \in [0, 1]$ is a metric such that a higher value indicates a stronger association between the feature and the class label. By sorting

the candidate features in descending order based on $SU(f_i', c)$, we obtain a new feature space F_{sort}, which is then divided into three buckets with high, medium, and low correlations.

$$F_{sort} = B_1 \cup B_2 \cup B_3 \tag{5}$$

Next, features are selected from different feature buckets to form a feature subspace Ω_g, ensuring that each subspace contains features with strong, moderate, and weak correlations. This approach prevents highly correlated features from clustering together and enhances the distinctiveness of the subspaces as well as the diversity of the search. For the g-th feature subspace Ω_g, its importance is determined jointly by the average correlation between features and class labels and the average redundancy within the subspace:

$$FI_g = \frac{1}{|\Omega_g|} \sum_{f_i \in \Omega_g} SU(f_i, c) - \eta \frac{2}{|\Omega_g|(|\Omega_g| - 1)} \sum_{f_i, f_j \in \Omega_g, i<j} SU(f_i, f_j) \tag{6}$$

where the first term denotes the average importance of features within the subspace, and the second term denotes the average redundancy among features. And η is the redundancy penalty coefficient. The larger the FI_g, the stronger the discriminative power of the subspace and the lower its internal redundancy. Subsequently, the total population size P is allocated across G feature subspaces using the Softmax function:

$$P_g = P \times \frac{\exp(FI_g)}{\sum_{k=1}^{G} \exp(FI_k)} \tag{7}$$

where P_g denotes the initial number of individuals allocated to the g-th subspace. Through this strategy, subspaces with higher importance receive more search resources. Meanwhile, other subspaces still retain a certain degree of exploration opportunity, thereby balancing search efficiency with feature diversity.

3.3 Multi-Strategy Integrated Improved Hybrid Breeding Feature Selection

To address the issues of uneven population distribution, limited adjustment mechanisms, and inefficient subpopulation updates in Hybrid Breeding Optimization algorithms [35,36] during complex optimization, we propose a multi-strategy improved hybrid breeding optimization algorithm (MIHBO) for efficient feature subset search within the differentiated feature subspaces. The algorithm enhances global search capability and population diversity in high-dimensional spaces, dynamically balances exploration and convergence, and obtains a more stable optimized feature subset, thereby providing a clearer feature representation for the subsequent region partitioning of the rough isolation forest. To ensure both the discriminative power and compactness of feature subset, we construct the following fitness function:

$$y(x_i) = \mu Q(A_i) + (1 - \mu)(1 - \frac{|A_i|}{d}) \tag{8}$$

where the first term is used to evaluate the anomaly detection performance of the selected feature subset, and the second term is used to penalize the proportion of selected features. A_i denotes the feature subset corresponding to an individual, $Q(A_i)$ denotes the anomaly detection performance metric derived from the current subset of features. $|A_i|$ represents the number of selected features, and $\mu \in [0, 1]$ is the balancing coefficient, and the influence of different μ values is discussed in the hyperparameter analysis section.

3.3.1 Initialization Strategies Based on Latin Hypercube Sampling and Elite Backpropagation

First, LHS is used to generate the initial individuals, ensuring that the population remains relatively uniformly distributed across the search space. Specifically, let the population size be P, the number of feature dimensions be d, and the i-th individual be represented as:

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{id}), \quad i = 1, 2, \ldots, P \tag{9}$$

where $x_{ij} \in [0,1]$ denotes a continuous encoded value. LHS divides the search interval for each dimension into N equal-probability subintervals and performs random sampling within each subinterval, thereby obtaining an initial population with better coverage. Since feature selection is essentially a binary selection problem, it is necessary to map continuous individuals to a subset of binary features:

$$b_{ij} = \begin{cases} 1, & x_{ij} \geq \theta \\ 0, & x_{ij} < \theta \end{cases} \tag{10}$$

where θ denotes the binary threshold, $b_{ij} = 1$ denotes the selection of the j-th feature, and $b_{ij} = 0$ denotes its non-selection. To further improve the quality of the initial population, we introduce an elite back-propagation strategy. For the current individual X_i, its back-propagated counterpart is defined as:

$$\tilde{x}_{ij} = r_{\min} + r_{\max} - x_{ij} \tag{11}$$

where $r_{\min}$ and $r_{\max}$ denote the lower and upper bounds of the j-th dimensional variable, respectively. The initial population generated by LHS is merged with the reverse population, and the top N individuals are selected as the final initial population based on the fitness function. This strategy retains the uniform coverage advantage of LHS while expanding the potential high-quality region through reverse search.

3.3.2 Three-Classification Strategy Based on the Roulette Wheel Concept

To better capture the continuity of individual variation and the dynamic changes in the population, and to enhance population diversity while maintaining evolutionary efficiency, we introduce a roulette wheel mechanism to replace the static ranking strategy. Under this strategy, individuals in the population are first ranked from highest to lowest fitness. The top one-third of individuals are assigned to the maintainer line M, while the remaining two-thirds undergo the roulette wheel operation. Specifically, first, normalize the remaining two-thirds of the individuals, as shown in Eq. (12):

$$p_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min} + \varepsilon}, \quad \varepsilon = 10^{-8} \tag{12}$$

where y_i denotes the fitness value of the current individual, and $y_{\max}$ and $y_{\min}$ denote the best and worst fitness values among the remaining individuals, respectively. Next, we generate a random number $r \sim U(0,1)$. If $r \leq p_i$, the individual is assigned to the restoring line R; otherwise, it is assigned to the sterile line O. This process is repeated until the entire population has been partitioned, thereby shifting the grouping strategy from a global static approach to a local dynamic one. This prevents the partitioning results from being entirely determined by the sorting order, thereby enhancing the flexibility and diversity of the population structure.

3.3.3 Individual Update Strategy for Retention Systems Based on the Grey Wolf Optimization Algorithm

To balance guided search with global exploration, we introduce GWO to maintain the evolution of the population. Through the gray wolf hierarchy and the capture mechanism, individuals retain high-quality

genes and continuously converge toward the optimal region, thereby enhancing internal search capabilities and adaptability. Specifically, the top three individuals in terms of fitness are selected as α, β, and δ. For an individual ω in the preserved population, its distance to these three guiding individuals is defined as:

$$D_\alpha = |C_1 \cdot x_\alpha - x_\omega|;\ D_\beta = |C_2 \cdot x_\beta - x_\omega|;\ D_\delta = |C_3 \cdot x_\delta - x_\omega| \tag{13}$$

where $C_k = 2r_{1,k}$, $k = 1, 2, 3$, $r_{1,k} \in [0, 1]$ is a random number. Then, individual ω is updated:

$$x_1 = x_\alpha - A_1 \cdot D_\alpha;\ x_2 = x_\beta - A_2 \cdot D_\beta;\ x_3 = x_\delta - A_3 \cdot D_\delta \tag{14}$$

$$A_i = 2a \cdot r_2 - a \quad i = \{1, 2, 3\} \tag{15}$$

where A_i is used to control the search step and direction during position updating. r_2 denotes a random number between [0, 1], a is the linear attenuation factor. To avoid insufficient early search and slow late convergence in the traditional GWO, we introduce a cosine-nonlinear decay strategy to optimize the parameter a curve. This strategy enhances global search by using larger step sizes in the early stages and gradually reduces the step size later on to improve local convergence accuracy. The dynamic update formula for the decay parameter a is as follows:

$$a = (a_{\max} - a_{\min}) \cdot \cos\left[\frac{\pi}{2} \cdot \left(\frac{iter}{\max\_\ iter}\right)^{2.5}\right] \tag{16}$$

where $iter$ denotes the current iteration number and $\max\_\ iter$ denotes the maximum number of iterations. $a_{\max}$ and $a_{\min}$ denote the upper and lower bounds of a, respectively. The final update location is:

$$x_{new} = \frac{x_1 + x_2 + x_3}{3} \tag{17}$$

3.3.4 Inbreeding-Free Individual Update Strategy Based on Simulated Annealing and Shannon Entropy

To address the evolutionary characteristics of infertile lineages, we have introduced a dynamic update strategy based on simulated annealing. This strategy incorporates information entropy into the temperature decay mechanism and adapts the temperature decay rate based on the population distribution. First, sterile individuals generate new individuals according to Eq. (18):

$$x_i^{new} = x_i + rand\frac{T}{\max\_\ iter} \tag{18}$$

where x_i denotes the current individual from the retained or eliminated branch, $rand$ denotes a random number between $(0, 1)$. And T is the temperature parameter for the current iteration, $\max\_\ iter$ denotes the maximum number of iterations for the population. If the fitness of the new individual is better, then $x_i = x_i^{new}$. Otherwise, the population accepts the inferior solution with a certain probability:

$$p_{accept} = \exp\left(\frac{y(x_{new}) - y(x_i)}{T}\right) \tag{19}$$

where $y(\cdot)$ denotes the fitness function, T is the temperature parameter at the current iteration, which controls the probability of accepting suboptimal solutions. It gradually decreases as the iteration progresses.

To adaptively control the temperature decay rate based on population diversity, we employ the Shannon entropy value H_k to regulate the temperature reduction, characterizing the distributional complexity of feature subsets. Specifically, for the sterile population S, the probability that the j-th feature equals 1 is:

$$c_j^1 = \frac{1}{NP}\sum_{i=1}^{NP} x_{ij} \tag{20}$$

where NP denotes the current population size of the sterile line. Therefore, the probability of obtaining 0 is denoted by $c_j^0 = 1 - c_j^1$. Subsequently, the information entropy of the j-th characteristic bit is calculated based on the Shannon entropy formula.

$$H_j = -\left(c_j^0 \log_2 c_j^0 + c_j^1 \log_2 c_j^1\right) \tag{21}$$

When $c_j^0 = c_j^1 = 0.5$, the entropy value is at its maximum, indicating that the dimension is the most uncertain and diverse. When c_j^0 or c_j^1 equals 1, the entropy value H_j is 0, indicating that the dimension has stabilized within the population. Finally, the entropy values of all feature positions are averaged to serve as a diversity metric for the entire population.

$$H_k = \frac{1}{d}\sum_{j=1}^{d} H_j \tag{22}$$

where d denotes the number of candidate features. The larger the value of H_k, the more dispersed and structurally diverse the inbred lines are as a whole; conversely, the smaller the value, the more concentrated and convergent they are. Therefore, the temperature attenuation mechanism is:

$$\alpha = \frac{1}{1 + H_k};\ T = T \cdot (1 - \alpha) \tag{23}$$

where α is the temperature decay factor, T denotes the temperature parameter in the simulated annealing process.

3.4 Extended Isolation Forests Incorporating Rough Sets

Building upon the Extended Isolation Forest (EIF) [37] framework, we introduce the concept of rough set approximate regions to propose a five-way partition-based rough isolation forest model, whose partitioning mechanism is illustrated in Fig. 2. This module takes the optimized feature subset obtained from the previous stage as input, reducing the interference of irrelevant and redundant dimensions in random hyperplane partitioning. Unlike traditional EIF, which relies solely on random hyperplanes for binary partitioning, rough approximate regions are constructed around the hyperplanes to partition the node space into five subregions with clear semantic meanings. In the figure, black scatter points represent the current node samples, green lines represent the random hyperplane, blue lines represent the boundaries of the approximate regions derived from its extension. And the red dashed lines indicate the boundary zone near the random hyperplanes, where sample membership is uncertain. This five-way partition captures the differences in certainty, uncertainty, and anomaly levels of samples relative to the partition boundaries, thereby enhancing the precision and robustness of anomaly isolation.

In the proposed model, each isolation tree still follows EIF's random attribute selection and recursive construction strategy. However, it has been extended in terms of node partitioning, the number of subspaces, and the anomaly metric. The overall process includes node statistical parameter estimation, five-region partitioning based on rough set, recursive tree construction, and anomaly scoring based on weighted paths.

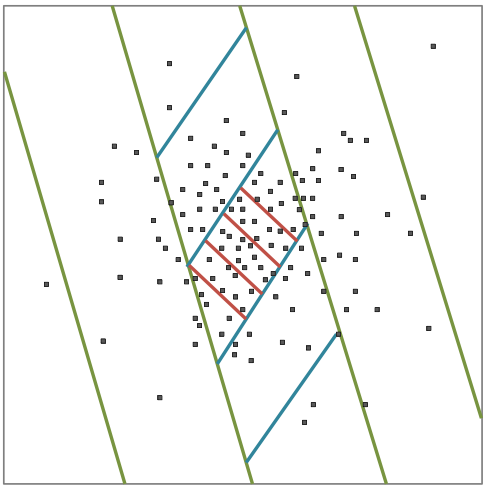


Figure 2: Schematic diagram of the extended isolated forest algorithm based on rough sets.

Node statistical parameter estimation. At each level of the tree construction, let the subset of samples in the current node be $S_t = \{x_i\}_{i=1}^{n_t}$, where n_t denotes the number of samples contained in the current node. Following the EIF construction method, a direction vector is randomly generated, and the samples x_i are projected onto this direction to obtain one-dimensional projection values $z_i = \mathbf{w}^T x_i$. Subsequently, the mean μ_t and standard deviation σ_t are computed based on the set of projection values $Z_t = \{z_i \mid l_i \in S_t\}$ for all samples in the current node: $\mu_t = \frac{1}{n_t}\sum_{i=1}^{n_t} z_i$, $\sigma_t = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}(z_i - \mu_t)^2}$.

where μ_t and σ_t represent the central tendency and dispersion of the samples at the current node along the random projection direction, respectively. To assess the overall stability of the model, the mean and standard deviation of the results were calculated through multiple independent repetitions. Furthermore, a neighborhood similarity matrix is computed based on the sample subset S_t, which is used for subsequent determination of the coarse approximation region and boundary region.

Five-region partitioning based on rough set. Based on rough set theory, the lower approximation, upper approximation, and boundary region at node x_i can be obtained, defined as $\underline{R_t}(X) = \{x_i \in S_t \mid z_i \le \mu_t + N\sigma_t\}$, $\bar{R}_t(X) = \{x_i \in S_t \mid z_i \le \mu_t + 2N\sigma_t\}$, and $BN_t(X) = \bar{R}_t(X) - \underline{R_t}(X)$, respectively. Where N is a scaling coefficient for the rough approximation regions, which controls the width of the boundary region along the random projection direction. Therefore, the approximation accuracy at node x_i is $\alpha_t(X) = \frac{|\underline{R_t}(X)|}{|\bar{R}_t(X)|}$; the larger the value, the smaller the region of uncertainty. To preserve directional information relative to the random hyperplane, the rough set-inspired regions described above are further decomposed into five mutually exclusive subregions: $G_1 = \{x_i : z_i < \mu_t - 2N\sigma_t\}$, $G_2 = \{x_i : \mu_t - 2N\sigma_t \le z_i < \mu_t - N\sigma_t\}$, $G_3 = \{x_i : \mu_t - N\sigma_t \le z_i \le \mu_t + N\sigma_t\}$, $G_4 = \{x_i : \mu_t + N\sigma_t < z_i \le \mu_t + 2N\sigma_t\}$, and $G_5 = \{x_i : z_i > \mu_t + 2N\sigma_t\}$. Among these, G_3 corresponds to the lower approximation, containing the most representative and stable samples, representing highly typical normal observations. G_2 and G_4 correspond to the upper approximation, which are rough boundary regions representing samples with a certain degree of uncertainty. These samples exhibit a higher degree of abnormality than those in the lower approximation region, but have not yet shown extreme deviations. G_1 and G_5 lie outside the Upper Approximation region in the negative domain, containing samples far from the central distribution that exhibit the strongest abnormal characteristics.

Through this five-branch structure, tree nodes not only partition the data but also explicitly indicate the position of each sample on a continuum ranging from "Typical—Uncertain—Anomalous."

Recursive Tree Construction. Building upon the EIF recursive tree construction method, this approach integrates the results of the five-region partitioning with the concept of rough set uncertainty. It optimizes hyperplane construction and node splitting strategies to improve the efficiency of outlier isolation. The specific steps are as follows:

Step 1: From the selected optimal feature subset, a subsample of size s is randomly drawn without replacement and placed at the root node of the current recursive tree. Node statistical parameters and region weight coefficients are loaded. The region weights are denoted as ω_k for G_k, $k = 1, \ldots, 5$. According to the semantics of the five-region partition are set to (0.5, 0.75, 1, 0.75, 0.5), respectively, with a larger weight assigned to the central lower-approximation region.

Step 2: A direction vector $\vec{w}$ is randomly generated in a d-dimensional feature space and then normalized. Combine this with the mean of the core region to calculate the hyperplane intercept θ (with a range of $\mu_j \pm 0.2\sigma_j$), ensuring that samples in the boundary region are prioritized for segmentation.

Step 3: Based on the hyperplane equation $\vec{w} \cdot \vec{x} = \theta$, divide the samples at the current node into left and right subtrees. Using the non-distinguishability property of rough sets, identify samples in the boundary region and process them first.

Step 4: Recursively perform parameter estimation, region partitioning, hyperplane construction, and sample partitioning on the left and right subtrees until a node contains only one sample or the tree depth reaches $\log_2 s$, where s denotes the subsample size. At that point, stop splitting and designate the node as a leaf node.

Step 5: Repeat the above steps to construct a recursive tree, which together forms an extended isolation forest based on rough sets. Each tree retains region weights and node splitting information to support anomaly scoring.

Recursive Tree Construction. Calculation of anomaly scores based on weighted paths: First, the test samples are input into each recursive tree, and the path length l_i $(i = 1, 2, \ldots, T)$ from the root node to the leaf node is recorded, with the number of nodes traversed reduced by 1. Second, based on the region type of the leaf node where the sample ultimately resides, a corresponding weight coefficient w_i is assigned. This weight is predefined based on the semantic regions of the five-region rough set partition; the closer to the central region, the greater the weight, and the farther from the central region, the smaller the weight.

Compute the weighted path length for each tree as $l_{w_i} = l_i \times w_i$ and calculate the average $\overline{l}_w$.

$$\overline{l}_w = \frac{1}{T} \sum_{i=1}^{T} l_{w_i} \tag{24}$$

Finally, referring to the EIF scoring formula and incorporating the uncertainty of the rough set boundary, we normalize $\overline{l}_w$ to obtain the anomaly score AS.

$$AS = 2^{-\frac{\overline{l}_w}{c(s)}} \tag{25}$$

where $c(s)$ is the normalization factor, defined as $c(s) = 2\log_2(s-1) + \gamma - \frac{2(s-1)}{s}$, where γ is Euler's constant and s is the sample size. The anomaly score $AS \in [0, 1]$, where the closer AS is to 1, the more likely the sample is an outlier.

3.5 Complexity Analysis

Let m denotes the number of training samples, K denotes the maximum number of iterations corresponding to max_ $iter$, and $|S|$ denotes the number of finally selected features. The computational cost of MSFO-EIF mainly comes from three stages. In the feature space optimization stage, mRMR pre-screening calculates feature-label relevance and feature-feature redundancy, leading to $O(mDd)$. The SU-based feature grouping further evaluates dependency relationships among the retained candidate features, with $O(md^2)$. In the multi-strategy feature selection stage, each iteration includes population update and fitness evaluation. Since the fitness evaluation requires building and testing the rough extended isolation forest for candidate feature subsets, this stage costs $O(KP(d + Td(s + m))\log s)$. After the optimal feature subset is obtained, the final anomaly detection stage constructs T rough extended isolation trees and computes anomaly scores, resulting in $O(T|S|(s + m)\log s)$. Therefore, the overall time complexity is $O(mDd + md^2 + KP(d + Td(s + m)\log s) + T|S|(s + m)\log s)$. For space complexity, $O(md)$ is required to store the candidate feature matrix after pre-screening, $O(d^2)$ is used for the feature dependency or redundancy matrix, and $O(Pd)$ is needed to maintain the population encoding during feature selection. In addition, the forest structure requires $O(T|S|s)$ space to store the split parameters and node information of all trees. Thus, the overall space complexity is $O(md + d^2 + Pd + T|S|s)$.

4 Simulation and Results Analysis

4.1 Dataset and Basline

To validate the effectiveness of the proposed anomaly detection model, MSFO-EIF, we selected three high-dimensional biomedical datasets-Prostate_GE [38], Prostate_Tumors [39], SMK_CAN_187 [40]-for comprehensive experiments. Details of these three experimental datasets are provided below.

Table 1 provides a statistical overview of these three datasets, detailing their sample sizes, feature dimensions, and the number of classes. For the baseline algorithms, we selected several classic machine learning models, including KNN [41], LOF [9], IF [42], EIF [37], IBBA-EIF [43], and RRSM [44], to evaluate the differences between them and our proposed model.

Table 1: Statistical overview of datasets.

ID	Dataset	Sample Size	Feature Dimension	Class
D1	Prostate_GE	102	5966	2
D2	Prostate_Tumors	102	10,509	2
D3	SMK_CAN_187	187	19,993	2

4.2 Experimental Details

To systematically evaluate the performance of the proposed anomaly detection model, we select accuracy, precision, F1-score, AUC, and AUC-PR as evaluation metrics. To ensure the fairness and reproducibility of the experimental results, all baseline models were evaluated using a standardized experimental environment and evaluation process, with the anomaly threshold uniformly set to 0.7. Specifically, the KNN and LOF models used default parameter settings to reflect their general performance in practical applications. The IBBA-EIF and RRSM models were configured according to the parameters recommended in the original papers to fully leverage their methodological capabilities. All experiments were conducted on a Windows 10 operating system with an Intel(R) Xeon(R) Gold 6230 CPU @ 2.10 GHz and 128 GB of RAM. All algorithms were implemented in Python. To minimize the impact of randomness on the experimental

results and enhance the reliability of model evaluation, five-fold cross-validation was employed to evaluate each model.

4.3 Analysis of Experimental Results

As shown in Table 2, MSFO-EIF generally achieved the best or most stable detection performance across the three datasets. On dataset D1, the accuracy, F1 score, AUC, and AUC-PR of MSFO-EIF reached 96.78%, 96.46%, 95.84%, and 15.68%, respectively, all of which were higher than those of the other comparison methods. Although basic baseline methods such as KNN, LOF, IF, and EIF can identify anomalous samples to a certain extent, their overall metrics are lower than those of IBBA-EIF, RRSM, and MSFO-EIF. This indicates that traditional distance, density, and basic isolation mechanisms remain susceptible to the effects of distance clustering, unstable local distributions, and interference from redundant features in high-dimensional data. In contrast, IBBA-EIF and RRSM achieved better results on the D1 dataset, indicating that the improved methods can enhance high-dimensional anomaly detection performance to a certain extent. Although RRSM slightly outperforms MSFO-EIF in terms of Precision, its Accuracy, F1-score, AUC, and AUC-PR are 1.09, 1.01, 1.28, and 1.37 percentage points lower than those of MSFO-EIF, respectively, which suggests that MSFO-EIF offers a more balanced overall classification performance. It also indicates that mRMR pre-screening and feature grouping help reduce high-dimensional redundancy interference and yield more stable feature representations. On dataset D2, the performance of the various models fluctuated somewhat, but the differences between the various types of methods remained quite pronounced. The overall performance of KNN, LOF, IF, and EIF remains at a relatively low level, indicating that basic anomaly detection methods have limited adaptability in complex high-dimensional spaces. RRSM ranks second-best in terms of accuracy, precision, and F1 score, while IBBA-EIF outperforms RRSM in AUC and AUC-PR, suggesting that these two categories of improved methods have different focuses in terms of classification stability and global discriminative capability. In contrast, MSFO-EIF achieved the best results across the board, with accuracy, precision, F1 score, AUC, and AUC-PR reaching 94.03%, 93.67%, 93.85%, 94.79%, and 15.36%, respectively. Specifically, MSFO-EIF's AUC improved by 2.08 and 4.43 percentage points compared to IBBA-EIF and RRSM, respectively, while its AUC-PR improved by 1.52 and 2.67 percentage points, respectively, indicating that its anomaly discrimination capability is more stable. This also demonstrates that MSFO-EIF is not a simple superposition of existing techniques, but rather improves the quality of feature subset search through multi-strategy feature optimization, enabling subsequent isolation detection to take place in a more effective feature space, thereby alleviating the problems of feature redundancy and classification instability in high-dimensional data. On dataset D3, although KNN, LOF, IF, and EIF still possess some detection capability, their overall performance is inferior to that of the improved methods, indicating that their ability to characterize complex feature relationships, boundary samples, and uncertain samples remains insufficient. IBBA-EIF and RRSM achieved better results compared to the baseline methods, suggesting that feature optimization or improved isolation mechanisms help enhance the effectiveness of high-dimensional anomaly detection. Building on this, MSFO-EIF achieved the best results across all metrics, with Accuracy, Precision, F1-score, AUC, and AUC-PR of 96.59%, 96.21%, 96.40%, 93.65%, and 15.14%, respectively. Compared with IBBA-EIF, MSFO-EIF achieved improvements of 7.21, 7.26, 7.23, 3.53, and 3.72 percentage points, respectively, across the five metrics; compared to RRSM, improvements of 2.25, 2.34, 2.30, 2.29, and 1.56 percentage points, respectively, were observed, further demonstrating that the five-region coarse isolation mechanism enhances the model's ability to characterize boundary and uncertain samples.

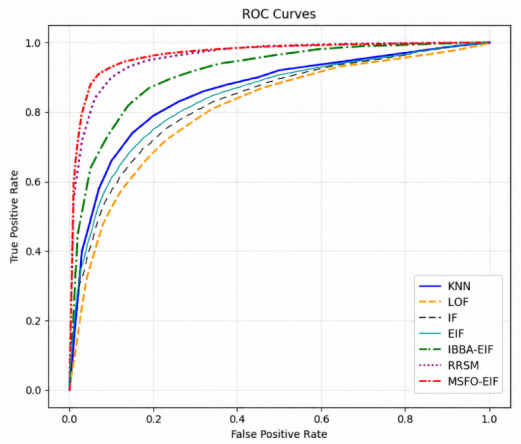
Table 2: Overview of experimental results.

Dataset	Model	Accuracy	Precision	F1-Score	AUC	AUC-PR
D1	KNN	82.30 ± 0.69	87.92 ± 0.34	88.11 ± 1.92	88.24 ± 0.58	9.24 ± 0.7
	LOF	83.97 ± 1.05	88.64 ± 0.73	89.30 ± 0.49	85.13 ± 3.06	8.17 ± 0.62
	IF	80.72 ± 0.83	86.05 ± 1.47	86.37 ± 0.56	86.44 ± 2.74	8.42 ± 0.91
	EIF	81.83 ± 0.42	87.24 ± 0.88	87.42 ± 1.16	87.68 ± 0.67	8.96 ± 2.31
	IBBA-EIF	94.45 ± 0.38	93.87 ± 1.21	94.16 ± 0.84	91.32 ± 0.57	12.86 ± 1.66
	RRSM	95.69 ± 0.71	95.71 ± 0.25	95.45 ± 0.93	94.56 ± 0.44	14.31 ± 0.36
	MSFO-EIF	96.78 ± 0.18	95.55 ± 0.64	96.46 ± 0.29	95.84 ± 0.41	15.68 ± 0.52
D2	KNN	85.17 ± 1.37	87.64 ± 0.66	87.90 ± 0.79	85.27 ± 2.56	8.06 ± 0.88
	LOF	83.23 ± 3.12	85.71 ± 0.94	85.97 ± 1.73	83.68 ± 0.69	7.12 ± 2.83
	IF	84.78 ± 0.96	87.06 ± 0.58	87.18 ± 2.18	84.73 ± 1.32	7.54 ± 0.74
	EIF	86.08 ± 0.47	88.37 ± 1.09	88.49 ± 0.63	86.18 ± 0.82	8.63 ± 1.51
	IBBA-EIF	91.56 ± 0.76	91.02 ± 0.33	91.29 ± 1.44	92.71 ± 0.52	13.84 ± 0.97
	RRSM	92.33 ± 0.59	91.85 ± 1.08	92.09 ± 0.48	90.36 ± 1.91	12.69 ± 0.71
	MSFO-EIF	94.03 ± 0.27	93.67 ± 0.41	93.85 ± 0.19	94.79 ± 0.68	15.36 ± 0.36
D3	KNN	85.04 ± 0.53	89.42 ± 1.41	89.73 ± 0.64	86.72 ± 0.92	8.79 ± 2.65
	LOF	84.61 ± 1.76	89.08 ± 0.47	89.34 ± 2.91	89.77 ± 0.38	9.86 ± 0.81
	IF	83.94 ± 2.07	88.22 ± 0.71	88.41 ± 1.29	87.61 ± 0.85	8.48 ± 0.44
	EIF	85.36 ± 0.62	89.15 ± 2.34	89.52 ± 0.89	88.88 ± 1.12	9.31 ± 0.57
	IBBA-EIF	89.38 ± 0.68	88.95 ± 0.95	89.17 ± 0.35	90.12 ± 1.58	11.42 ± 0.72
	RRSM	94.34 ± 0.31	93.87 ± 1.17	94.10 ± 0.74	91.36 ± 0.55	13.58 ± 0.96
	MSFO-EIF	96.59 ± 0.24	96.21 ± 0.13	96.40 ± 0.39	93.65 ± 0.76	15.14 ± 0.28

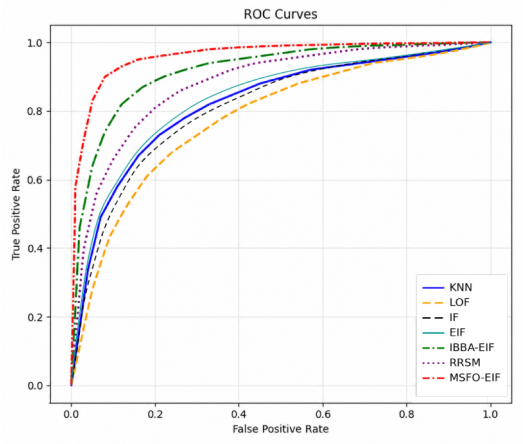
Overall, MSFO-EIF significantly outperforms baseline algorithms on all three datasets. Its performance improvements are primarily attributed to the following factors: First, the mRMR and feature grouping mechanisms effectively reduce feature redundancy and enhance discriminative information. Second, the multi-strategy feature selection algorithm improves global search capabilities and the quality of feature subsets. Finally, the extended isolation forest incorporating rough sets enhances the ability to characterize boundary and uncertain samples through multi-region partitioning, thereby significantly improving the performance and stability of anomaly detection.

In terms of ROC curve analysis, Fig. 3 illustrates the classification performance of each model on datasets D1, D2, and D3, respectively. Overall, the ROC curves of MSFO-EIF are consistently closest to the top-left corner across the three datasets, indicating that it maintains a higher true positive rate at a lower false positive rate and therefore has stronger overall discriminative capability. On dataset D1, MSFO-EIF obtains the highest AUC, suggesting a more stable classification boundary. The curves of RRSM and IBBA-EIF rank second and generally outperform those of KNN, LOF, IF, and EIF. On dataset D2, MSFO-EIF still maintains the best ROC curve pattern, especially in the low-FPR region, where it achieves a higher TPR more rapidly. Its AUC improves by 2.08 and 4.43 percentage points compared with IBBA-EIF and RRSM, respectively, further indicating its more stable ability to identify anomalous samples. On dataset D3, MSFO-EIF achieves an AUC of 93.65%, which is higher than RRSM at 91.36% and IBBA-EIF at 90.12%, again confirming its incremental advantage over existing improved methods. In contrast, while KNN, LOF, IF, and EIF are capable of detecting anomalies to a certain extent, their ROC curves are generally lower than those of the improved

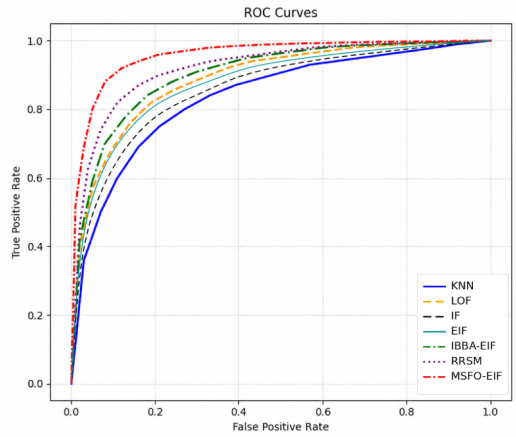
methods, indicating that the basic detection mechanisms still have limited ability to capture complex feature relationships and boundary anomaly samples.



(a) ROC curves on dataset D1



(b) ROC curves on dataset D2



(c) ROC curves on dataset D3

Figure 3: ROC curves of the comparison models.

These results are consistent with the quantitative metrics in Table 2, indicating that the performance improvement of MSFO-EIF is not limited to a single evaluation metric, but is consistently reflected in both the ROC curve patterns and the evaluation metrics. In particular, compared with IBBA-EIF and RRSM, the AUC improvements of MSFO-EIF on multiple datasets demonstrate that the coordinated design of feature space optimization and the five-region rough isolation mechanism further enhances the model's ability to identify boundary samples and weak anomalous samples. Overall, the ROC curves across the three datasets are generally closer to the top-left corner, indicating a better overall trade-off between the true positive rate and false positive rate over all possible classification thresholds. Therefore, it demonstrates stronger overall discriminative capability.

4.4 Generalization Experiment on High-Dimensional Text Data

To further evaluate the generalization ability of MSFO-EIF beyond biomedical datasets, an additional experiment was conducted on the R8 high-dimensional text dataset, using AUC-PR as the evaluation metric. R8 is a high-dimensional sparse text benchmark derived from Reuters-21578 and covering eight news

categories. We selected two representative deep anomaly detection methods, DIF [45] and RDP [46], as comparison baselines.

As shown in Table 3, MSFO-EIF achieves an AUC-PR value of 0.152 on the R8 dataset, outperforming DIF, and RDP, which obtain 0.145 and 0.146, respectively. Although the overall AUC-PR values are relatively low due to the sparsity and imbalance of high-dimensional text data, MSFO-EIF still achieves relative improvements of 4.83% and 4.11% over DIF, and RDP. This result indicates that the proposed feature optimization strategy and rough-set-based isolation mechanism can improve anomaly discrimination in sparse high-dimensional spaces, providing supplementary evidence for the adaptability of MSFO-EIF on non-biomedical data.

Table 3: AUC-PR comparison on the R8 dataset.

Method	AUC-PR
DIF	0.145 ± 0.031
RDP	0.146 ± 0.017
MSFO-EIF	0.152 ± 0.012

4.5 Ablation Experiments

To validate the effectiveness of each module in MSFO-EIF, we sequentially removed key modules and compared the detection performance of different variants on the Prostate_GE dataset. F1-score and AUC were selected as representative metrics to reflect the comprehensive detection performance and overall discriminative ability of the model. Specifically, we constructed the following ablation models: removal of the mRMR pre-screening module. Removal of the feature grouping mechanism based on SU. Removed the multi-strategy collaborative optimization mechanism and adopted HBO. Removed the rough set multi-region partitioning mechanism, reducing the model to standard extended isolation forests. Removed the front-end feature selection process, retaining only the five-region rough EIF.

The results are summarized in Fig. 4, and the following findings were observed: After removing mRMR, the F1-score and AUC decreased by 6.64% and 6.88%, respectively, indicating that prescreening effectively reduces high-dimensional redundant features. After removing SU grouping, both metrics decreased by 5.09 and 5.20 percentage points, respectively, indicating that feature grouping helps construct a more complementary candidate feature subspace. After replacing MIHBO with the original HBO, the F1-score and AUC decreased by 3.21% and 3.33%, respectively, demonstrating that the multi-strategy collaboration mechanism can improve the quality of feature search. Removing the five-region rough set partition also resulted in a decline in model performance, indicating that this mechanism helps enhance the ability to identify boundary and uncertain samples. Furthermore, the "w/o feature selection" configuration achieved the lowest performance, with the F1-score and AUC dropping to 86.91% and 85.97%, respectively. This result indicates that performing region partitioning and anomaly detection directly in the original high-dimensional feature space is subject to interference from redundant and noisy features, whereas the front-end feature selection process plays a crucial role in improving the partitioning quality and detection stability of the five-region rough EIF.

4.6 Parameter Sensitivity Analysis

This section discusses the impact of key parameters on the performance of MSFO-EIF, focusing primarily on the five-region partitioning coefficient N and the balance coefficient μ in the fitness function.

A hyperparameter sensitivity analysis was conducted on the Prostate_GE dataset, using F1 score and AUC as evaluation metrics. To ensure comparability, a one-factor-at-a-time sensitivity analysis was adopted. Only one hyperparameter was varied each time, while the others were fixed at the default settings with relatively better performance. When analyzing N, μ was fixed at 0.15; when analyzing μ, N was fixed at 1.5.

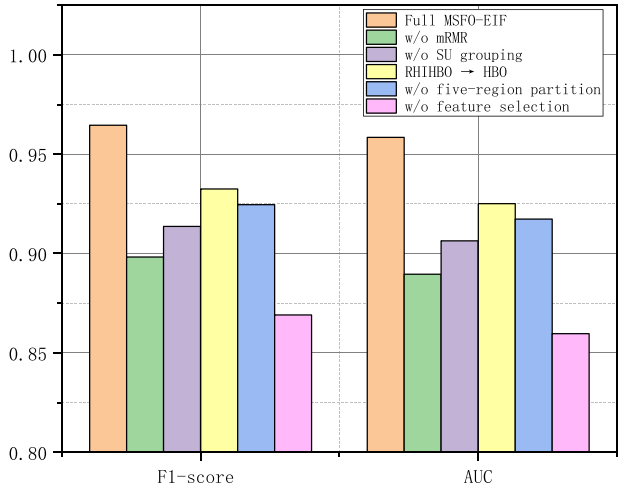
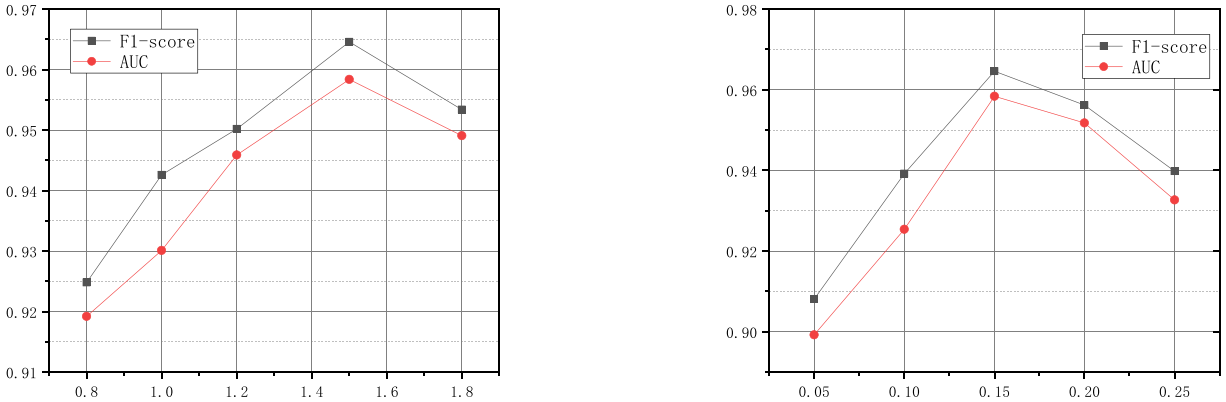


Figure 4: Results of ablation experiments on the Prostate_GE dataset.

The effect of parameter N is shown in Fig. 5a. As N increases from 0.8 to 1.5, the F1-score and AUC of MSFO-EIF generally show an upward trend, indicating that moderately expanding the near-core and boundary regions enhances the characterization of boundary and weakly anomalous samples. However, when N was further increased to 1.8, model performance declined slightly; this may be due to the fact that an excessively wide regional scope weakens the isolation effect of anomalous samples.



(a) F1-score and AUC on different N (b) F1-score and AUC on different μ

Figure 5: The result of parametric sensitivity analysis.

The effect of the balancing coefficient μ is shown in Fig. 5b. As μ increases from 0.05 to 0.15, model performance gradually improves, indicating that appropriate constraints on the number of features can reduce the interference of redundant features on anomaly detection. As μ continues to increase, the F1-score and AUC decline, suggesting that excessive feature compression may lead to the removal of some effective discriminative features, thereby affecting detection performance.

4.7 Statistical Significance Analysis

To further evaluate the statistical reliability and stability of the proposed MSFO-EIF model, Monte Carlo experiments were conducted on the Prostate_GE dataset. In each iteration, samples were randomly selected and evaluated under the same experimental settings, and the AUC value was recorded as the main statistical indicator. A total of 100 independent iterations were performed to examine whether the proposed method could maintain stable discrimination ability under different random sampling conditions.

As shown in Fig. 6, the AUC values fluctuate within a relatively narrow range from 0.9468 to 0.9679 across the 100 iterations. The mean AUC reaches 0.958, which indicates that MSFO-EIF maintains a stable anomaly discrimination performance on the Prostate_GE dataset. In addition, the standard deviation of the AUC values is 0.0052, suggesting that the performance fluctuation under repeated random sampling is limited. These results demonstrate that the proposed method has good stability and statistical reliability on the Prostate_GE dataset, and further support the effectiveness of MSFO-EIF in high-dimensional anomaly detection tasks.

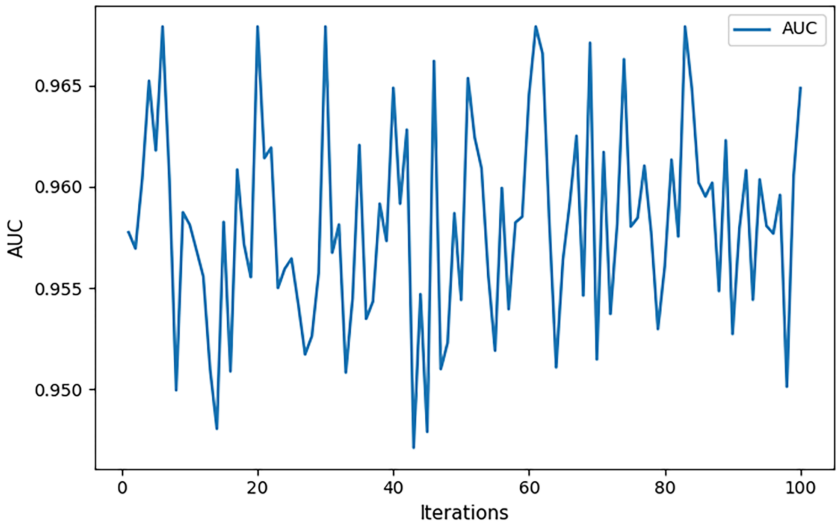


Figure 6: Results of statistical significance analysis on the Prostate_GE dataset.

5 Conclusion

To address the challenges of severe feature redundancy, concealed anomaly patterns, and the difficulty in characterizing uncertainty in high-dimensional anomaly detection, we propose a high-dimensional anomaly detection model, MSFO-EIF, which integrates multi-strategy feature optimization with an improved isolation mechanism. This model combines mRMR with symmetric uncertainty for feature pre-screening and grouping, and integrates Latin hypercube sampling, elite backpropagation, the roulette mechanism, and gray wolf optimization to achieve efficient feature subset search. During the detection phase, we introduce the concept of rough set approximate regions to construct a five-region partition in the extended isolation forest and enhance anomaly scoring capabilities through weighted path length. Experimental results demonstrate that this model outperforms KNN, LOF, IF, EIF, IBBA-EIF, and RRSM in metrics such as Accuracy, F1-score, AUC and AUC-PR, exhibiting excellent effectiveness and robustness. Future work will integrate deep representation learning and graph modeling, while incorporating uncertainty quantification and interpretability mechanisms to enhance practical application value.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Natural Science Foundation of China (62376089, U23A20318, 62302154), the Natural Science Foundation of Hubei Province (2024AFB882), the Program for Scientiffc and Technological Innovation Teams of Young and Middle-Aged Researchers in Higher Education Institutions of Hubei Province (T2023007) and the Hubei University of Technology under Grant XJ2024004202.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Jiaojiao Deng and Zhiwei Ye; methodology, Dong-Fang Wu, Jiaojiao Deng and Zhiwei Ye; software, Jiaojiao Deng; validation, Dong-Fang Wu, Jiaojiao Deng and Zhiwei Ye; formal analysis, Dingfeng Song; investigation,Jiaojiao Deng; resources, Rong Gao, Fan Ma and Dingfeng Song; data curation, Rong Gao; writing—original draft preparation, Jiaojiao Deng; writing—review and editing, Dong-Fang Wu and Zhiwei Ye; visualization, Dong-Fang Wu, Rong Gao, Fan Ma and Dingfeng Song; supervision, Dong-Fang Wu, Zhiwei Ye, Rong Gao and Fan Ma; project administration, Fan Ma; funding acquisition, Dong-Fang Wu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are openly available from public repositories. The Prostate_GE dataset is available from the scikit-feature repository, DOI: 10.1016/S1535-6108(02)00030-2. The Prostate_Tumors dataset is available from the GEMS/UniFeat public benchmark collection, DOI: 10.1016/j.ijmedinf.2005.05.002. The SMK_CAN_187 dataset is available from the scikit-feature repository, DOI: 10.1038/nm1556. The R8 high-dimensional text dataset is a single-label subset derived from the Reuters-21578 Text Categorization Collection. The original Reuters-21578 collection is available from the UCI Machine Learning Repository, DOI: 10.24432/C52G6M.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zamanzadeh Darban Z, Webb GI, Pan S, Aggarwal C, Salehi M. Deep learning for time series anomaly detection: a survey. ACM Comput Surv. 2024;57(1):1–42. doi:10.1145/3691338.
2. Salhi A, Alshamrani R, Althbiti A, Ismail A, Abd-ElRahman M, Hassan BM. Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema. Sci Rep. 2025;15(1):35038. doi:10.1038/s41598-025-08699-4.
3. Bao J, Sun H, Deng H, He Y, Zhang Z, Li X. BMAD: benchmarks for medical anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 17–21; Seattle, WA, USA. p. 4042–53. doi:10.1109/cvprsw63382.2024.00408.
4. Huang H, Wang P, Pei J, Wang J, Alexanian S, Niyato D. Deep learning advancements in anomaly detection: a comprehensive survey. arXiv:2503.13195. 2025.
5. Liu J, Gao R, Ye Z, Deng J, Zhang X, Zhang L, et al. AGDSF-Mamba: adaptive metric graph diffusion with sparse filtering for mamba guided anomaly detection via spatiotemporal learning in Industrial IoT. Internet Things. 2026;37(7):101934. doi:10.1016/j.iot.2026.101934.
6. Salem O, Liu Y, Mehaoua A. Anomaly detection in medical WSNs using enclosing ellipse and chi-square distance. In: Proceedings of the 2014 IEEE International Conference on Communications (ICC); 2014 Jun 10–14; Sydney, Australia. p. 3658–63. doi:10.1109/icc.2014.6883890.
7. Cauteruccio F, Fortino G, Guerrieri A, Liotta A, Mocanu DC, Perra C, et al. Short-long term anomaly detection in wireless sensor networks based on machine learning and multi-parameterized edit distance. Inf Fusion. 2019;52(4):13–30. doi:10.1016/j.inffus.2018.11.010.
8. Steinbuss G, Böhm K. Benchmarking unsupervised outlier detection with realistic synthetic data. ACM Trans Knowl Discov Data. 2021;15(4):1–20. doi:10.1145/3441453.

9. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data; 2000 May 16–18; Dallas, TX, USA. p. 93–104. doi:10.1145/335191.335388.
10. Yuan H, Cui N, Li C, Cui Z, Chang L. Early stage internal short circuit fault diagnosis for lithium-ion batteries based on local-outlier detection. J Energy Storage. 2023;57(3):106196. doi:10.1016/j.est.2022.106196.
11. Yuan Y, Wang S, Chen H, Luo C, Yuan Z. Anomaly detection based on fuzzy neighborhood rough sets. Inf Sci. 2025;709:122075. doi:10.1016/j.ins.2025.122075.
12. Duan J, Feng J. Dynamic anomaly detection in blast furnace operations using a transformer-enhanced isolation forest framework. Eng Appl Artif Intell. 2025;159(21):111724. doi:10.1016/j.engappai.2025.111724.
13. Leveni F, Magri L, Alippi C, Boracchi G. Preference isolation forest for structure-based anomaly detection. Pattern Recognit. 2025;172:112405. doi:10.2139/ssrn.5201368.
14. Xu J, Wang Q, Cao W, Qi X. Virtual node isolation forest: one-class isolation-based method for novelty detection. Eng Appl Artif Intell. 2026;164(14):113296. doi:10.1016/j.engappai.2025.113296.
15. Boschi T, Bonin F, Ordonez-Hurtado R, Pascale A, Epperlein J. A new computationally efficient algorithm to solve feature selection for functional data classification in high-dimensional spaces. Proc Mach Learn Res. 2024;235:4383–402.
16. Kim M, Choi HS, Kim J. Higher-order neural additive models: an interpretable machine learning model with feature interactions. In: Proceedings of the 2025 IEEE International Conference on Data Mining (ICDM); 2025 Nov 12–15; Washington, DC, USA. p. 1310–9.
17. Chang D, Rao C, Xiao X, Hu F, Goh M. Multiple strategies based grey wolf optimizer for feature selection in performance evaluation of open-ended funds. Swarm Evol Comput. 2024;86(1):101518. doi:10.1016/j.swevo.2024.101518.
18. Jiang X, Liu J, Wang J, Nie Q, Wu K, Liu Y, et al. Softpatch: unsupervised anomaly detection with noisy data. Adv Neural Inf Process Syst. 2022;35:15433–45. doi:10.52202/068431-1123.
19. Bindini L, Perini L, Nistri S, Davis J, Frasconi P. Dealing with uncertainty in contextual anomaly detection. arXiv:2507.04490. 2025. doi:10.24963/ijcai.2023/264.
20. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput. 2001;13(7):1443–71. doi:10.1162/089976601750264965.
21. Sakurada M, Yairi T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis; 2014 Dec 2; Gold Coast, QLD, Australia. p. 4–11. doi:10.1145/2689746.2689747.
22. Zong B, Song Q, Min MR, Cheng W, Lumezanu C, Cho D, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada.
23. Gałka Ł. Optimized deep isolation forest. Pattern Recognit Lett. 2025;197(6):88–94. doi:10.1016/j.patrec.2025.07.014.
24. Arcudi A, Ferreri A, Borsatti F, Susto GA. Function based isolation forest (FuBIF): a unifying framework for interpretable isolation-based anomaly detection. IFAC PapersOnLine. 2025;59(26):91–6. doi:10.1016/j.ifacol.2025.12.016.
25. Sartor D, Barbariol T, Susto GA. Bayesian active learning isolation forest (B-ALIF): a weakly supervised strategy for anomaly detection. Eng Appl Artif Intell. 2024;130(5):107671. doi:10.1016/j.engappai.2023.107671.
26. Song Y, Lin H, Li Z. Outlier detection in a multiset-valued information system based on rough set theory and granular computing. Inf Sci. 2024;657(1):119950. doi:10.1016/j.ins.2023.119950.
27. Hu Q, Yuan Z, Zhang J, Mi J. Fuzzy rough guided subspace anomaly detection in nominal data. Pattern Recognit. 2026;174(11):113024. doi:10.1016/j.patcog.2025.113024.
28. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38. doi:10.1109/tpami.2005.159.

29. Keller F, Muller E, Bohm K. HiCS: high contrast subspaces for density-based outlier ranking. In: Proceedings of the 2012 IEEE 28th International Conference on Data Engineering; 2012 Apr 1–5; Arlington, VA, USA. p. 1037–48. doi:10.1109/icde.2012.88.
30. Najarbashi M. Detection of fraudulent financial papers by picking a collection of characteristics using optimization algorithms and classification techniques based on squirrels. arXiv:2211.07747. 2022. doi:10.48550/arxiv.2211.07747.
31. Liang H, Wang D, Lu Y, Song M, Liu L, An L, et al. Time-EAPCR-T: a universal deep learning approach for anomaly detection in industrial equipment. arXiv:2503.12534. 2025. doi:10.48550/arxiv.2503.12534.
32. Aksu D, Aydin MA. MGA-IDS: optimal feature subset selection for anomaly detection framework on in-vehicle networks-CAN bus based on genetic algorithm and intrusion detection approach. Comput Secur. 2022;118(4):102717. doi:10.1016/j.cose.2022.102717.
33. Li T, Zhan ZH, Xu JC, Yang Q, Ma YY. A binary individual search strategy-based bi-objective evolutionary algorithm for high-dimensional feature selection. Inf Sci. 2022;610(2):651–73. doi:10.1016/j.ins.2022.07.183.
34. Li Z, Li H, Gao W, Xie J, Slowik A. Feature selection in high-dimensional classification via an adaptive multifactor evolutionary algorithm with local search. Appl Soft Comput. 2025;169(4):112574. doi:10.1016/j.asoc.2024.112574.
35. Ye Z, Ma L, Chen H. A hybrid rice optimization algorithm. In: Proceedings of the 2016 11th International Conference on Computer Science & Education (ICCSE); 2016 Aug 23–25; Nagoya, Japan. p. 169–74. doi:10.1109/iccse.2016.7581575.
36. Ŷe Z, Zhang S, Zhou W, Wu L, Cai T, Zhang M, et al. Hboffs: hybrid breeding optimization algorithm inspired federated feature selection for intrusion detection in IIoT. Knowl Based Syst. 2025;329(1):114419. doi:10.1016/j.knosys.2025.114419.
37. Hariri S, Kind MC, Brunner RJ. Extended isolation forest. IEEE Trans Knowl Data Eng. 2021;33(4):1479–89. doi:10.1109/TKDE.2019.2947676.
38. Chen C, Weiss ST, Liu YY. Graph convolutional network-based feature selection for high-dimensional and low-sample size data. Bioinformatics. 2023;39(4):btad135. doi:10.1093/bioinformatics/btad135.
39. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002;1(2):203–9. doi:10.1016/s1535-6108(02)00030-2.
40. Liu S, Mocanu DC, Matavalam ARR, Pei Y, Pechenizkiy M. Sparse evolutionary deep learning with over one million artificial neurons on commodity hardware. Neural Comput Appl. 2021;33(7):2589–604. doi:10.1007/s00521-020-05136-7.
41. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7. doi:10.1109/tit.1967.1053964.
42. Liu FT, Ting KM, Zhou ZH. Isolation forest. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining; 2008 Dec 15–19; Pisa, Italy. p. 413–22. doi:10.1109/ICDM.2008.17.
43. Guan YL. Research on extended isolation forest algorithm based on improved bat algorithm optimization [dissertation]. Shenyang, China: Shenyang University of Technology; 2023. (In Chinese). doi:10.27322/d.cnki.gsgyu.2023.000403.
44. Cheng YJ. Research on outlier detection algorithm and application of high-dimensional data based on RRSM [dissertation]. Zhenjiang, China: Jiangsu University; 2024. (In Chinese). doi:10.27170/d.cnki.gjsuu.2024.002177.
45. Xu H, Pang G, Wang Y, Wang Y. Deep isolation forest for anomaly detection. IEEE Trans Knowl Data Eng. 2023;35(12):12591–604. doi:10.1109/TKDE.2023.3270293.
46. Wang H, Pang G, Shen C, Ma C. Unsupervised representation learning by predicting random distances. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence; 2020 Jul 23–29; Vienna, Austria. p. 2950–6. doi:10.24963/ijcai.2021/408.