

ORIGINAL RESEARCH

Open access

A Hybrid Marine Predators Algorithm for Optimized Gene Selection in Cancer Classification Using Microarray Data

Camila Ortega^{1*}, Diego Molina¹, Andres Perez², Valeria Soto¹

Abstract

DNA microarray technology relies on a chip containing thousands of DNA probes to simultaneously quantify the activity levels of many genes. The resulting gene expression data, typically arranged in matrix form, serves as the foundation for pattern-recognition techniques that separate samples taken from healthy subjects from those obtained from cancer patients. Nonetheless, extracting genuinely informative biomarkers from such datasets remains extremely difficult due to the curse of dimensionality and the substantial presence of uninformative or noisy genes. This study presents an effective gene selection framework that combines the powerful Minimum Redundancy Maximum Relevance (MRMR) filter with a newly designed hybrid metaheuristic. The proposed optimizer merges an enhanced version of the Marine Predator Algorithm (MPA) with a crossover mechanism, yielding the so-called MPAC algorithm. MPAC is tailored to identify a small yet highly discriminative set of genes that dramatically improves cancer classification accuracy. Classification is performed using the k-nearest neighbors (k-NN) classifier. The main strength of MPAC resides in its marked improvement of the exploration and exploitation capabilities of the original MPA population. It systematically identifies optimal gene subsets suitable as cancer biomarkers while carefully balancing global search (exploration) and local refinement (exploitation). The proposed method was comprehensively benchmarked on nine widely recognized microarray benchmark datasets. Its classification performance was rigorously compared with several standard and recently published optimization-based approaches. Results indicate that MPAC achieves superior accuracy across most datasets and delivers very competitive results on the remaining datasets.

Keywords Marine predator algorithm, Crossover, Gene selection optimization, rMRMR, Classification

*Correspondence:

Camila Ortega
camila.ortega@gmail.com

¹ Department of Smart Grid Control Systems, Faculty of Engineering, University of Lima, Lima, Peru

² Department of Renewable Energy Analytics, Faculty of Engineering, Pontifical Catholic University of Peru, Lima, Peru

Introduction

Thanks to DNA microarray technology, researchers can now simultaneously monitor the activity of thousands of genes and generate gene expression profiles [1-3]. These profiles, however, belong to the category of high-dimensional data: they contain far more gene variables (features) than patient samples. In the standard data arrangement, each column corresponds to a gene, and each row corresponds to a biological sample collected from patients [4, 5]. Variations in gene expression values provide critical clues for identifying and categorizing different types of cancer. Multiple studies have shown that relying on single genes or poorly chosen gene groups frequently leads to degraded classification performance because irrelevant or redundant genes introduce noise [6, 7]. To improve prediction reliability and reduce the impact of unhelpful features, a key data-preprocessing technique called gene (or feature) selection is commonly applied before classification [8, 9].

Broadly speaking, gene selection algorithms fall into three major families: filter-based, wrapper-based, and hybrid methods [8]. Wrapper methods involve two phases—candidate subset generation (usually handled by a search algorithm) and subset evaluation (typically performed by training and testing a classifier). Filter methods, by contrast, evaluate genes or gene subsets independently of any classifier, relying solely on statistical characteristics computed from the training data. This independence makes filters considerably faster than wrappers, although they usually sacrifice some predictive accuracy [10]. Hybrid techniques attempt to merge the computational efficiency of filters with the higher accuracy typically achieved by wrappers. Empirical evidence consistently shows that hybrid approaches tend to deliver better overall performance than either pure filter or pure wrapper strategies used in isolation [4, 11, 12].

Salem et al. [13] proposed a novel approach for classifying human cancer samples using gene expression profiles. This technique leverages initial knowledge to remove redundant and irrelevant genes. Subsequently, a genetic algorithm refines the pre-filtered genes, yielding more useful and informative features. By performing an initial filtering step, the genetic algorithm (GA) can explore a reduced yet relevant portion of the gene space, thereby identifying more valuable and promising gene candidates. Final classification is achieved through genetic programming. Published experimental findings confirmed that the IG/GA method outperforms competing techniques.

Yan et al. [14] applied the Coral Reefs Optimization method, incorporating Tournament Selection as the selection mechanism and Simulated Annealing for local search, to solve the gene selection challenge. On two benchmark datasets (Colon and Nervous System), their proposed approach achieved average classification accuracies of 82.00% and 92.31% while utilizing 21.4 and 20.5 features, respectively.

Zhang et al. [15] employed Information Gain (IG) in combination with an enhanced krill herd optimization algorithm within a two-phase gene selection framework. This strategy enabled a more thorough exploration of the gene space and promoted greater diversity in the search. Compared with other state-of-the-art methods in the literature, their approach delivered superior results in terms of the number of selected genes, convergence speed, and overall classification performance.

Pashaei [16] introduced a hybrid gene selection framework that integrates an improved Aquila optimizer into the wrapper phase, with Minimum Redundancy Maximum Relevance (MRMR) as the filter. The method features a binary version of the Aquila optimizer, equipped with a time-varying mirrored S-shaped transfer function, further augmented with a mutation operator to enhance its exploration. Extensive evaluations across 11 biomedical datasets showed that the proposed technique consistently outperformed prior gene selection strategies.

El Akadi et al. [4] presented a two-phase gene selection procedure for microarray data classification. The initial phase applies MRMR filtering to rank genes, advancing only the highest-ranked ones to the subsequent stage. There, a genetic algorithm (GA) evaluates interactions among these top genes to pinpoint the most effective subset. Support Vector Machines and Naive Bayes serve as classifiers, with classification accuracy used as one of the fitness criteria. Experimental results indicated that applying MRMR and GA independently or in a single combined phase does not perform as well on microarray data as the sequential two-stage integration.

Chuang et al. [17] proposed a hybrid strategy to address gene selection issues by combining Binary Particle Swarm Optimization (PSO) with a competitive genetic algorithm. Candidate gene subsets are evaluated via the K-Nearest Neighbor (KNN) classifier under leave-one-out cross-validation. The experimental outcomes highlighted the method's effectiveness in reducing both the size of gene subsets and classification error rates.

Alomari et al. [18] combined MRMR with the Bat algorithm to identify optimal gene subsets. MRMR first filters genes to select the most relevant ones, which are then passed to a GA-based wrapper. Support Vector Machine is employed for classification to compute the fitness of each candidate subset. The results achieved by this method were found to be on par with leading approaches reported in the literature.

Due to the complex interdependencies among genes and the high dimensionality of the search space, most of the methods described above are prone to local optima [19, 20]. Consequently, a more robust and efficient optimization technique is required to mitigate these limitations and effectively capture the majority of genes relevant to cancer biomarkers for accurate classification.

Faramarzi *et al.* [21] presented the Marine Predators Algorithm (MPA), a nature-inspired metaheuristic that mimics the foraging and social dynamics of marine predators. The algorithm's update mechanism relies on randomized population adjustments and incorporates two distinct random-walk strategies: Brownian motion and Lévy flights. Among its key strengths are a simple parameter-tuning process, broad applicability, ease of implementation, and strong global search capability. MPA has proven highly effective across diverse domains, including ECG signal classification [22], dynamic clustering [23], fog computing [24], COVID-19 medical image segmentation [25, 26], photovoltaic array reconfiguration [27], and gene selection tasks [28, 29].

This study introduces and develops an innovative hybrid optimization framework, grounded in the Marine Predators Algorithm (MPA), to effectively address the gene selection challenge. The newly presented hybrid technique, termed MPAC, integrates key elements and behavioral patterns drawn from an enhanced optimization method referred to as crossover (CO). By carefully calibrating the search dynamics, MPAC strives to enhance both exploitative and exploratory strengths while markedly improving the core effectiveness of the standard MPA. The central contributions of the present work include:

1. A fresh hybrid filter-wrapper approach to gene selection is proposed and implemented, using rMRMR for filtering and MPAC for wrapper selection.
2. Adaptation and application of the MPA are used to address the gene selection issue, to lower the fitness score by pinpointing a compact set of informative biomarker genes that improve classification performance.
3. MPAC is formulated as an original fusion of crossover (CO) mechanisms with MPA agents and structural features, specifically engineered to heighten the proficiency of MPA search agents in pinpointing superior candidate solutions. In addition, the designed hybridization strategy amplifies exploratory reach and exploitative precision, secures an optimal trade-off between them, and avoids premature convergence to suboptimal regions.
4. The performance of the introduced MPAC is rigorously assessed across 10 well-established microarray benchmark datasets, confirming its ability to substantially reduce gene subset size without compromising classification quality.

Although metaheuristic-driven feature selection has received considerable attention in the literature, the MPAC framework brings forward two significant novelties. Firstly, it merges the Marine Predators Algorithm with an ensemble of crossover operators—an unprecedented combination crafted to accelerate convergence rates and enrich population diversity. Secondly, it pairs this evolutionary core with a robust MRMR-based preprocessing filter that systematically applies diverse relevance metrics to discard irrelevant and redundant features. The resulting dual-phase structure narrows the dimensionality of the problem space and promotes greater consistency and biological meaningfulness in the chosen gene signatures. Distinct from previous efforts, this method purposefully links solid dimensionality reduction upfront with responsive evolutionary fine-tuning, positioning it as particularly well-suited to the intricate demands of high-dimensional oncology microarray datasets.

Research Background

This section traces the evolution of the Marine Predators Algorithm, which has matured into a concise yet highly competitive metaheuristic for optimization.

Marine Predators Algorithm (MPA)

Like many other metaheuristic techniques, the MPA operates as a population-based method. At the outset, candidate solutions are uniformly scattered throughout the search space in the initial iteration.

Drawing from the concept of survival of the fittest observed in natural ecosystems, the most successful foragers emerge as apex predators. Correspondingly, MPA identifies the best-performing solution as the top predator. This elite solution is then employed to build the Elite matrix, consisting of vectors that direct the search by incorporating positional information from the prey, as expressed in [Eq. 1](#).

$$\text{Elite} = \begin{bmatrix} X_{1,1}^I & \dots & X_{1,d}^I \\ \vdots & \ddots & \vdots \\ X_{n,1}^I & \dots & X_{n,d}^I \end{bmatrix} \quad (1)$$

The top predator vector, denoted as X^I , is duplicated n times to populate the Elite matrix, where n equals the total number of search agents and d denotes the dimensionality of the problem. Within this framework, both predators and prey function as search agents: predators pursue prey, while prey seek their own resources. Following each iteration, the Elite matrix is refreshed whenever a superior solution displaces the current top predator.

A companion matrix called prey is initialized to have the same dimensions as the Elite matrix. Position updates for predators are guided by information extracted from the Prey matrix. In essence, the algorithm begins by randomly generating an initial Prey population, from which the strongest individual (the predator) forms the Elite matrix. The Prey matrix is structured as indicated in [Eq. 2](#).

$$\text{Prey} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,d} \\ \vdots & \vdots & X_{i,j} & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,d} \end{bmatrix} \quad (2)$$

Here, $X_{i,j}$ denotes the j^{th} dimension of the i^{th} prey individual within the Prey matrix. It should be emphasized that the core optimization mechanism revolves directly around these two matrices—Elite and Prey—which drive the search and refinement process.

MPA optimization scenarios

The MPA search mechanism unfolds across three distinct phases, each governed by varying velocity ratios that mirror real-world predator-prey dynamics throughout their lifecycle. These phases are outlined below:

1. High Velocity Ratio: This stage occurs when the prey outpaces the predator. A designated portion of the total iterations is allocated to this phase, adhering to biologically inspired movement patterns. The mathematical representation is given by:

$$\vec{stepsize}_i = \vec{R}_B \otimes (\vec{Elite}_i - \vec{R}_B \otimes \vec{Prey}_i) \quad i = 1, \dots, n \quad (3)$$

$$\begin{aligned}\vec{Prey}_i &= \vec{Prey}_i + P \cdot \vec{R} \otimes \vec{stepsize}_i \\ \text{Subject to} \\ Iter &< \frac{1}{3} Max_{iter}\end{aligned}\quad (4)$$

Where RB is a vector of random values drawn from a Normal (Gaussian) distribution to model Brownian motion, the symbol $\otimes$ indicates element-wise multiplication. Multiplying RB by the Prey matrix mimics the prey's motion. A fixed parameter $P = 0.5$ is applied, while R consists of uniformly distributed random numbers between 0 and 1. This high-velocity regime is active during roughly the first third of iterations, featuring large step sizes that promote strong global exploration. Iter stands for the current iteration count, and Maxiter denotes the total number of iterations allowed.

Unit Velocity Ratio: Here, predator and prey move at comparable speeds, establishing a state of equilibrium. A particular segment of iterations is reserved for this balanced phase, consistent with natural behavioral patterns. The update rule for the first half of the population is expressed as:

$$\vec{stepsize}_i = \vec{R}_L \otimes \left(\vec{Elite}_i - \vec{R}_L \otimes \vec{Prey}_i \right) \quad i = 1, \dots, \frac{n}{2} \quad (5)$$

$$\begin{aligned}\vec{Prey}_i &= \vec{Prey}_i + P \cdot \vec{R} \otimes \vec{stepsize}_i \\ \text{Subject to} \\ \frac{1}{3} Max_{iter} &< Iter < \frac{2}{3} Max_{iter}\end{aligned}\quad (6)$$

Where RL is a vector of random numbers sampled from a Lévy distribution, emulating Lévy-flight behavior. Element-wise multiplication of RL with the Prey matrix replicates the prey's Lévy-style movements, and an additional step-size term is incorporated to further model displacement. Notably, Lévy distributions yield smaller average steps than Brownian motion, shifting the emphasis toward local exploitation in this phase.

The update for the second half of the population is formulated as follows:

$$\vec{stepsize}_i = \vec{R}_B \otimes \left(\vec{R}_B \otimes \vec{Elite}_i - \vec{Prey}_i \right) \quad i = \frac{n}{2}, \dots, n \quad (7)$$

$$CF = \left(1 - \frac{itr}{max_itr} \right)^{\left(\frac{2itr}{max_itr} \right)} \quad (8)$$

$$\vec{Prey}_i = \vec{Elite}_i + P \cdot CF \otimes \vec{stepsize}_i \quad (9)$$

In this case, the RB vector multiplied by the Elite matrix simulates the predator's Brownian-motion trajectory. Simultaneously, prey positions adapt in reaction to these predator movements. This reciprocal interaction between

predator and prey trajectories plays a pivotal role in shaping position updates and overall convergence behavior during the optimization.

3. Low Velocity Ratio: This stage occurs when the predator moves faster than the prey. As with the earlier phases, a designated portion of the iterations is allocated to this phase, in accordance with biologically inspired guidelines for predator–prey locomotion. The mathematical expression for this phase is given by:

$$\vec{stepsize}_i = \vec{R}_L \otimes (\vec{R}_L \otimes \vec{Elite}_i - \vec{Prey}_i) \quad i = 1, \dots, n \quad (10)$$

$$\vec{Prey}_i = \vec{Elite}_i + P.CF \otimes \vec{stepsize}_i \quad (11)$$

Here, element-wise multiplication of RL with the Elite matrix replicates the predator's Lévy-flight movement pattern. An additional step-size term is incorporated into the Elite position to model better the predator's displacement, which, in turn, influences the subsequent update of the prey's position. This mutual influence between predator trajectory and Elite-guided adjustments plays a key role in driving the overall optimization dynamics for both predator and prey populations.

Collectively, these three phases are engineered to faithfully reproduce the characteristic movement behaviors seen in natural predator–prey systems, thereby providing a well-organized framework for searching within the MPA algorithm.

Eddy formation and FADs' effect

Environmental factors, such as eddy currents and the presence of Fish Aggregating Devices (FADs), can significantly influence the behavior of marine predators. For instance, research indicates that sharks devote over 80% of their time to areas near FADs, treating them as dependable feeding zones. The remaining 20% is typically spent performing extended leaps across different dimensions, likely in pursuit of new prey distributions in alternative habitats. In the context of optimization, FADs are interpreted as local optima within the search landscape, where agents risk becoming trapped.

To mitigate the risk of premature convergence to these suboptimal points, the algorithm introduces occasional long-range jumps. Such mechanisms help escape local entrapment and facilitate broader exploration of the search space. The mathematical modeling of the FADs effect is described in [Eq. 12](#).

$$\vec{Prey}_i = \begin{cases} \vec{Prey}_i + CF \left[\vec{X}_{min} + \vec{R} \otimes (\vec{X}_{max} - \vec{X}_{min}) \right] \otimes \vec{U} & \text{if } r \leq FADs \\ \vec{Prey}_i + [FADs(1-r) + r] (\vec{Prey}_{r1} - \vec{Prey}_{r2}) & \text{if } r > FADs \end{cases} \quad (12)$$

Here, FADs = 0.2 indicates the probability threshold at which the FADs mechanism influences the optimization procedure. The binary vector U is formed by drawing a random vector from the interval [0,1] and setting each element to 0 if below 0.2, or to 1 if above 0.2. The term r denotes a uniformly distributed random value in [0,1].

The boundary vectors Xmin and Xmax specify the lower and upper limits for each dimension of the problem. The subscripts r1 and r2 represent randomly chosen indices from the prey population. These indices identify specific positions within the Prey matrix that are subject to modification by the FADs mechanism during the search process.

Marine memory

The MPA incorporates a memory mechanism to emulate the natural tendency of marine predators to recall and revisit locations associated with successful foraging. Following the position update of the Prey matrix and the application of the FADs effect, the fitness value of every solution in the current generation is computed and compared against the fitness of its counterpart from the immediately preceding iteration. Should the newly obtained solution exhibit superior fitness, it replaces the older one in the population.

This replacement strategy progressively refines solution quality across iterations and mirrors the real-world behavior of predators returning to previously rewarding hunting grounds.

Crossover

To bolster the exploitative strength of the MPA, crossover operators are integrated into the algorithm's framework. These operators combine an individual's position with the population's best-performing individual's position, as shown in [Eq.13](#). The resulting positions become new candidate solutions, enabling more effective probing of high-potential regions and improving overall solution quality.

$$X_m^t = \bowtie \left(X_m^{(t)}, X_f^{(t)} \right) \quad (13)$$

The operator $\bowtie$ represents the crossover operation used to blend two binary-encoded solutions. Specifically, it combines the current solution X_m (at index m) with a superior solution X_f . Selection of X_f is performed via the roulette wheel method—a probabilistic technique commonly employed in evolutionary computation that assigns higher selection probability to individuals with better fitness scores. This mechanism enhances the transmission of advantageous traits from elite solutions to newly created offspring.

Three distinct crossover operators are employed in this work. These operators are activated during the algorithm's reproduction stage to produce novel candidate solutions (offspring). The specific crossover types integrated into the MPA framework are as follows:

1. **Single-Point Crossover:** This technique selects a random cut point along the binary string of each parent and exchanges all bits after that point between the two parents, thereby forming new offspring.
2. **Two-Point Crossover:** This method operates in a manner analogous to single-point crossover, randomly selecting two distinct positions within the binary representations of the parents. The segment lying between these two cut points is then swapped between the parents to yield the offspring, as shown in [Figure 1](#).
3. **Uniform Crossover:** Under this operator, every bit position in the offspring is independently determined by selecting the corresponding bit from one of the two parents with equal probability (0.5). Consequently, the resulting offspring inherit a randomly distributed combination of genetic information from both parents, as illustrated in [Figure 2](#).

At the conclusion of the crossover operation, after one of the above operators has been used to exchange elements between the current solution X_m^t and the selected high-quality solution X_f^t —the principle of survival of the fittest is applied. More precisely, the fitness value of the newly created solution X_m^t is evaluated and compared against the fitness of its predecessor X^{t-1} from the previous generation. The superior solution (in terms of fitness) is retained in the population.

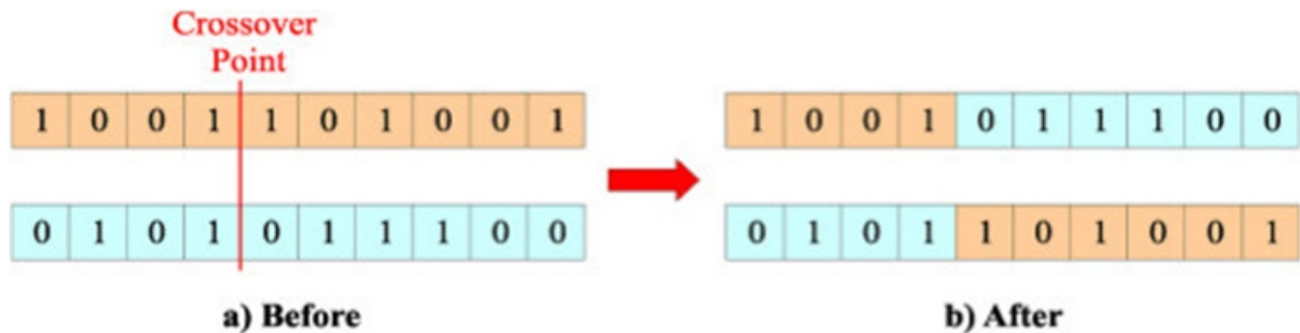


Figure 1. Two-point crossover.

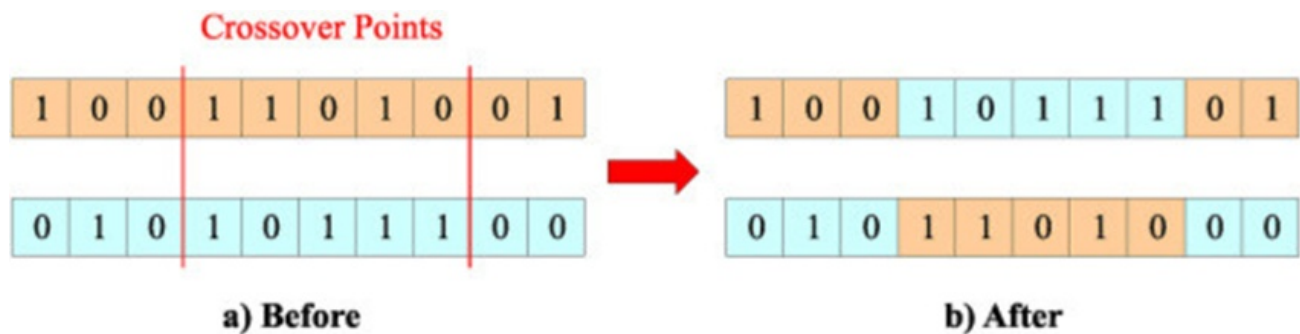


Figure 2. Uniform crossover.

Materials and Methods

Stage 1: Filter approach

The filtering stage constitutes a critical component of gene selection, as it substantially reduces the dataset's dimensionality by retaining only the most informative genes for subsequent processing by the wrapper method. The filtering technique adopted in the present study is an enhanced version of MRMR, termed resilient MRMR (rMRMR), originally proposed by Alomari *et al.* [30].

In gene selection tasks, rMRMR assigns a ranking score to each gene. Consistent with the original MRMR, it prioritizes genes that are most relevant to the target class (output) while minimizing redundancy among the selected genes. However, the conventional single-filter MRMR approach suffers from noticeable instability in classification performance, largely because it relies solely on mutual information to quantify relevance. To overcome this drawback, rMRMR was developed to compute gene relevance by integrating multiple complementary filter criteria, including distance-based measures, probability distribution characteristics, and information-theoretic metrics. The general framework of the rMRMR method is outlined in the subsections that follow.

The rMRMR filter plays a crucial role in mitigating the adverse effects of noisy and non-informative genes in microarray datasets. By combining multiple filter criteria—namely ReliefF, Chi-Square, and Kullback–Liebler—this approach captures a broad spectrum of gene relevance and discriminatory capability, thereby minimizing dependence on any single evaluation metric. The resulting multifaceted ranking procedure effectively removes irrelevant and redundant genes that might otherwise interfere with the subsequent optimization stage. As a result, the feature space forwarded to the wrapper phase is significantly reduced in size and enriched with more relevant genes.

Step 1: Initialization.

In the initialization phase, rMRMR employs three well-established filter methods: ReliefF [31], Chi-Square [32], and Kullback–Liebler [33]. These filters are executed separately. Genes are then evaluated according to their discriminative ability. Individual gene scores from each filter are summed to produce a unified gene list, which is then sorted by the average score across the three filters.

Step 2: Hybridization.

In the hybridization phase, an attempt is made to integrate the ranked gene list into the conventional MRMR filtering framework. First, during an intermediate assessment, preliminary gene relevance values are assigned based on the average scores from the ranked gene list. Next, gene relevance is computed using a dual-metric approach that incorporates $I(Gx, c)$ (mutual information) and $R(G_i)$ (the ranked gene list score). This combined strategy mitigates the instability inherent in relying on a solitary measure and ensures that gene relevance is assessed from multiple data perspectives. Consequently, the MRMR process achieves greater robustness and consistency.

Step 3: Filtering process outcomes.

The subsequent wrapper stage selects the most discriminative gene subset by retaining only the highest-ranked genes that satisfy predefined user-specified thresholds.

Stage 2: Wrapper approach

This section describes, explains, and outlines the proposed wrapper method. The wrapper phase further refines and identifies the most informative gene subset initially identified by rMRMR. MPAC, a novel hybrid optimization algorithm, is introduced and applied; it merges the strengths of MPA with crossover (CO) mechanisms. The CO component specifically targets and improves the weakest-performing MPA individuals within the population, thereby elevating the overall quality of MPA solutions. This integrated approach aims to achieve maximum classification accuracy using the smallest possible number of selected genes by guiding distant (poorly performing) solutions toward optimal regions and refining them through CO-based enhancements.

Solution representation

As noted earlier, gene selection constitutes an NP-hard combinatorial optimization problem. The search space comprises all possible gene subsets, and its complexity grows exponentially with the number of genes [34, 35]. To represent solutions within this space, let N denote the total number of genes; the total number of possible subsets is therefore 2^N .

As shown in Eq. 14, each candidate solution in the gene selection search space is encoded as a binary vector x of length N .

$$x = (x_1, x_2, \dots, x_N), \quad (14)$$

A gene i is included in the subset if $x_i = 1$; otherwise, it is excluded ($x_i = 0$).

Fitness function

As previously described, gene subsets are encoded in binary format (0 or 1), with only genes assigned a value of 1 being evaluated. The fitness of each solution is computed using the function defined in Eq. 15 [36].

$$\alpha \times R(D) + \beta \times \frac{|C| - |R|}{|C|} \quad (15)$$

In this formulation, a k-Nearest Neighbors (KNN) classifier is applied to the training dataset using the selected gene subset R to predict the class label D . The classifier performs 10-fold cross-validation repeated multiple times, yielding a classification accuracy denoted $\alpha_R(D)$. The cardinality of the gene subset is represented by $|R|$, while the total number of genes is $|C|$. The subset length is β , and classification accuracy α lies in the interval $[0, 1]$, with $\beta = 1 - \alpha$. Given the high priority placed on classification performance in gene selection applications, the weighting parameter α is fixed at 0.8 in this study [37-39].

Although k-NN is a widely adopted and straightforward classifier in gene selection research owing to its simplicity and ease of implementation, it is not without limitations. Its performance can be sensitive to the choice of k and the distance metric, and it may suffer degradation in the presence of high-dimensional noise or irrelevant features, potentially affecting both accuracy and generalization. Furthermore, as a lazy learning algorithm, k-NN stores the entire training set and performs comparisons at prediction time, rendering it computationally less efficient for very large datasets. Nevertheless, k-NN remains a dependable and interpretable option for assessing the discriminative quality of selected gene subsets, particularly when repeated cross-validation is used to ensure consistent evaluation across multiple runs.

Marine predator optimizer-based crossover

This section describes the rMRMR-MPAC framework, an advanced variant of the MPAC algorithm tailored explicitly for gene selection applications. The core objective of rMRMR-MPAC is to pinpoint biologically meaningful genes via the rMRMR preprocessing step. These prioritized genes then serve as the input features for the MPAC optimization process, thereby further refining and strengthening the gene selection outcomes.

The MPAC procedure comprises five primary sequential phases, which are elaborated as follows.

Step 1: MPAC Initialization.

In this initial phase, the essential parameters for both the MPA algorithm and the crossover mechanism are configured. For the MPA component, this includes setting β , P , FADs, the maximum number of iterations (MaxIter), the population size (N), and the search domain boundaries defined by the lower bound (lb) and upper bound (ub). The crossover operator itself requires no dedicated parameter adjustment.

Step 2: Population Initialization.

Here, random candidate solutions are generated to constitute the initial population for the gene selection problem. The population is represented as an $N \times D$ matrix, where N denotes the population size and D the number of decision variables (genes). The starting population for MPAC is mathematically expressed in Eq. 16.

$$MPACP = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_D^1 \\ x_1^2 & x_2^2 & \cdots & x_D^2 \\ \vdots & \vdots & \cdots & \vdots \\ x_1^N & x_2^N & \cdots & x_D^N \end{bmatrix}, \quad (16)$$

where $x \in 0, 1$.

Step 3: Population Evaluation.

During this stage, the fitness score of each individual in the population is computed using the objective function in Eq. 15. This evaluation assesses each candidate solution against the specific objectives and limitations of the gene selection

problem.

Upon completion of fitness calculations, the solution exhibiting the highest fitness value is selected and designated as the top predator, represented by XI within the MPA structure. This individual is considered the best-performing candidate within the current population.

Furthermore, the Elite matrix is refreshed to capture and retain details of this superior solution. As a form of memory, the Elite matrix stores the strongest solutions encountered during the optimization run. Updating the Elite matrix ensures that valuable information about elite solutions is maintained.

Step 4: Update population by MPA.

In this phase, the MPA search agents begin exploring the solution landscape by adjusting their positions to find superior candidates. The Elite matrix directs position modifications and follows the mathematical rules, specifically employing [Eqs. 4, 6, 9, and 11](#).

Subsequently, the influence of FADs is incorporated and applied via [Eq. 12](#), facilitating a balanced interplay between exploration and exploitation throughout the search.

The resulting candidate solutions from the MPA agents are then evaluated using [Eq. 15](#), which defines the objective function for gene selection. Depending on these fitness outcomes, both the Elite matrix and the associated memory are updated to safeguard information about the newly identified best solutions.

Step 5: Update population by the crossover operator.

This phase enhances the MPA's exploitation capacity and enables the discovery of improved solutions through crossover operations. After Step 4 has completed position updates, fitness evaluations, and assignments to the Elite matrix and memory, the crossover mechanism is triggered. It begins by randomly generating an integer value of 1, 2, or 3 to determine the crossover type: 1 corresponds to Single-Point Crossover, 2 to Two-Point Crossover, and 3 to Uniform Crossover.

Once the crossover operator has modified the solutions, the updated candidates are evaluated for fitness using the objective function. If any newly generated solution demonstrates superior fitness, it is incorporated into the population via the MPA's update rules.

Step 6: Check the stop criterion

Phases 3, 4, and 5 are repeated in a loop until the predefined termination condition is met.

Owing to the inherently high dimensionality of microarray gene expression data—where the gene count vastly exceeds the number of samples—gene selection procedures are especially prone to overfitting and compromised generalization. Moreover, the deployment of potent hybrid metaheuristics, such as MPAC, further heightens this risk due to their pronounced exploitative tendencies. To counteract these challenges, the proposed framework adopts a dual-stage methodology. Initially, the rMRMR filter dramatically reduces dimensionality by excluding irrelevant and redundant genes using diverse statistical measures, yielding a more concise and pertinent feature set. Subsequently, the MPAC algorithm optimizes within this narrowed space to fine-tune the identification of highly discriminative genes. In combination, these two stages work to alleviate the difficulties posed by high dimensionality and to enhance the overall dependability and resilience of the gene selection procedure.

Experimental Setup and Results

This section assesses the performance of the proposed MPAC method by conducting experiments on 11 gene expression datasets. All datasets are publicly accessible at <http://csse.szu.edu.cn/staff/zhuzx/Datasets.html>. The experimental evaluation is structured into three separate scenarios, each intended to measure MPAC's effectiveness relative to competing approaches.

To illustrate the advantage of incorporating the crossover operator into MPAC, MPAC results are directly contrasted with those from the standard MPA algorithm. Furthermore, the same crossover mechanism is applied to six additional optimization algorithms, demonstrating its consistent and reliable performance enhancement across different metaheuristic frameworks. This analysis seeks to underline the value of embedding crossover within these algorithms and its overall contribution to superior search efficiency.

In the second experimental scenario, MPAC's outcomes are benchmarked against several other hybrid models that integrate the crossover operator with various base optimization techniques. This comparison provides a clear perspective on how MPAC compares with alternative crossover-enhanced optimization strategies.

In the third and final evaluation, the proposed MPAC approach is pitted against 12 well-established state-of-the-art gene selection techniques that have been tested on the same dataset collection. Through this head-to-head comparison, the relative strengths and competitive advantages of MPAC become evident when measured against leading methods in the field. Collectively, these assessments yield a thorough and multifaceted picture of MPAC's capabilities in gene selection tasks.

Dataset

The datasets examined in this research comprise Leukemia, Leukemia_3c, Leukemia_4c, Colon, Lymphoma, CNS, SRBCT, MLL, Ovarian, and Lung_Cancer. These datasets are among the most frequently used benchmarks in the gene selection literature and vary considerably in scale, ranging from small to large feature dimensions. A concise overview of their key properties is presented in [Table 1](#).

These particular datasets were intentionally chosen to represent a broad spectrum of feature counts, sample sizes, and class structures, encompassing both binary and multiclass classification problems. Such heterogeneity ensures a comprehensive test of the proposed MPAC algorithm's ability to scale effectively and generalize across problems of differing complexity. All selected datasets are well-known, preprocessed microarray benchmarks commonly referenced in prior studies. To preserve alignment with published results and facilitate fair comparisons, no additional preprocessing steps (such as normalization or scaling) were performed.

Although microarray datasets continue to serve as standard testbeds for gene selection research, they are subject to certain inherent limitations. These include limited dynamic range, potential cross-hybridization effects, and reliance on fixed probe sets that may overlook previously uncharacterized or lowly expressed transcripts. Nevertheless, their standardized format, widespread use, and extensive documentation in the literature make them highly appropriate for conducting reliable comparative evaluations of gene selection algorithms.

Table 1. Datasets characteristic.

Datasets	# Classes	# Samples	# Genes
Colon			
CNS	2	60	7129
Leukemia_3c	3	72	7129
Leukemia_4c	4	72	7129

Leukemia	2	72	7129
Lung Cancer	5	203	12 601
Lymphoma	3	66	4026
MLL	3	72	12 582
Ovarian Cancer	2	253	15 154
SRBCT	4	83	2308

Parameter settings

To ensure impartial and uniform benchmarking across competing optimization techniques, the same foundational hyperparameters are enforced for all algorithms. In concrete terms, the population is fixed at 100 individuals, and the iteration limit is consistently capped at 100 cycles per run across all evaluated methods.

While these baseline controls remain identical, each algorithm possesses its own distinctive internal hyperparameters, which are fully documented in Table 2. Because all tested algorithms are probabilistic, outcomes inevitably exhibit some degree of run-to-run variability. To reliably quantify their consistency in addressing gene selection challenges, each algorithm was run 30 times per dataset, yielding a statistically meaningful distribution of performance values.

Table 2. Parameter settings of optimization algorithms.

Approach	Parameters
AOA	$\alpha=5, \mu=0.5$
BAT	$f_{max} \ \& \ f_{min} = 0 \ \& \ 2, \alpha = 0.9, \gamma = 0.9, A_{max} = 2, r0_{max} = 1$
EO	$a1 = 2, a2 = 1, GP = 0.5, V = 1$
GWO	
MPA	$\beta = 1.5, P = 0.5, FADs = 0.2$
PSO	$c1 = 2, c2 = 2$
WSO	$f_{min} = 0.07, f_{max} = 0.75, a0 = 6.25, a1 = 100, a2 = 0.0005, \tau = 4.125$

Effect of crossover operator on the optimization algorithms

The present subsection conducts a targeted comparative investigation to quantify the added value brought by integrating the proposed crossover strategy into leading optimization algorithms. Each baseline method is evaluated side-by-side with its enhanced variant that embeds the MPAC crossover mechanism. The base algorithms under scrutiny are the Arithmetic Optimization Algorithm (AOA) [40], Bat Algorithm (BAT) [41], Particle Swarm Optimization (PSO) [42], and White Shark Optimizer (WSO) [43]. Their respective crossover-augmented versions are labeled AOAC, BATC, PSOC, and WSOC.

Classification accuracy serves as the principal performance indicator, with both the arithmetic mean (AVG) and the standard deviation (STD) computed to capture central tendency and variability. Table 3 provides a complete, side-by-side summary that vividly illustrates how hybridization influences algorithmic effectiveness.

[illegible]

Table 4 further corroborates that fusing the original algorithms with the MPAC crossover framework yields consistent F1-score gains across the evaluated datasets. These uplifts are reinforced by concurrent reductions in standard deviation, reflecting markedly more dependable predictive behavior.

Table 4. Comparison between the proposed and original methods based on F1.

Datasets	Measure	MPAC	PSO	PSOC	WSO	WSOC	AOA	AOAC	BAT	BATC	MPA
CNS	AVG	0.96107	0.89624	0.89433	0.95971	0.96370	0.95252	0.95683	0.94887	0.96284	0.959
	STD	0.00018	0.01460	0.01325	0.00537	0.00542	0.01082	0.00700	0.00788	0.00466	0.004
Colon	AVG	0.95206	0.83540	0.83474	0.89056	0.90596	0.92913	0.91248	0.92099	0.92096	0.958
	STD	0.01447	0.01333	0.01744	0.01553	0.00898	0.02955	0.02330	0.01781	0.02220	0.012
Leukemia_3c	AVG	1	0.98904	0.98977	1	1	1	1	1	1	1
	STD	0	0.00730	0.00834	0	0	0	0	0	0	0
Leukemia_4c	AVG	0.98510			0.97065	0.99167			0.96226	0.98477	0.983
	STD	0.00446			0.01652	0.01172			0.01913	0.01745	0.007
Leukemia	AVG	1	0.98949	0.99193	1	1	1	1	1	1	1
	STD	0	0.00387	0.00453	0	0	0	0	0	0	0
Lung_Cancer	AVG	0.99362	0.98667	0.98654	0.99244	0.99300	0.99109	0.99215	0.99146	0.99383	0.993
	STD	0.00002	0.00128	0.00160	0.00155	0.00128	0.00192	0.00214	0.00153	0.00165	0.000
Lymphoma	AVG	1	1	1	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0	0	0	0
MLL	AVG	0.99344	0.96588	0.96574	0.98832	0.99599	0.98139	0.98637	0.98743	0.99589	0.994
	STD	0.00715	0.00721	0.00711	0.00680	0.00718	0.00890	0.00957	0.00502	0.00641	0.006
Ovarian	AVG	1	1	1	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0	0	0	0
SRBCT	AVG	1	1	1	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0	0	0	0

Focusing specifically on the CNS dataset, the hybridized versions AOAC, BATC, and WSOC exhibit clear superiority relative to their base forms AOA, BAT, and WSO. Of particular note, BATC improves its F1-score from 0.94887 to 0.96284 while simultaneously reducing the standard deviation from 0.00788 to 0.00466. Such simultaneous improvement in both average performance and consistency highlights the crossover mechanism's capacity to refine the precision–recall equilibrium—an especially valuable trait when dealing with intricate, multiclass cancer gene-expression profiles.

The Colon dataset reveals that WSOC elevates the F1-score from 0.89056 to 0.90596. In more demanding classification scenarios requiring both high sensitivity and specificity—such as those involving Leukemia_3c, Leukemia_4c, and MLL—the hybridized versions either preserve or exceed the F1 Scores of their baseline counterparts. A particularly striking gain occurs in Leukemia_4c, where BATC boosts its F1-score from 0.96226 to 0.98477, clearly illustrating the substantial contribution of MPAC integration toward superior classification outcomes.

Table 5 visualizes the fitness value improvements achieved by hybridizing conventional optimization algorithms with MPAC. It contrasts the fitness performance of the standalone methods (AOA, BAT, MPA, PSO, WSO) against their hybridized forms (AOAC, BATC, MPAC, PSOC, WSOC) across the evaluated datasets.

In the majority of cases, the hybrid algorithms deliver noticeably better results, as evidenced by lower mean fitness values—an indicator of stronger optimization performance. On the CNS dataset, for example, AOAC lowers the average fitness from 0.06220 (AOA) to 0.05652. Likewise, BATC markedly outperforms BAT by reducing fitness from 0.06911 to 0.04941, reflecting more thorough and efficient traversal of the solution landscape thanks to the MPAC component.

Hybrid configurations also exhibit considerably lower standard deviations across most datasets, indicating greater process stability and reproducibility. In Leukemia_4c, BATC not only achieves a higher average fitness but also narrows the standard deviation from 0.01913 to 0.01745, underscoring the increased reliability of the hybridized search.

On particularly critical datasets such as Leukemia and Lymphoma—where near-optimal fitness is essential due to the high-stakes nature of the domain—hybrids including MPAC and WSOC consistently achieve extremely low fitness scores, approaching zero, with very tight standard deviations. These results demonstrate the hybrids' ability to converge to the global optimum reliably.

Furthermore, the hybridized approaches display particularly strong and consistent advantages on more complex datasets such as MLL and Lung Cancer, suggesting that MPAC integration is particularly advantageous in difficult optimization environments where baseline algorithms tend to underperform. In the MLL dataset, WSOC dramatically cuts the average fitness from 0.01482 to 0.00607, while BATC likewise records substantial progress in Leukemia_4c.

Table 6 summarizes the gains in Matthews Correlation Coefficient (MCC) delivered by the hybridized methods across the various datasets. Clear advances are evident in several hybrids—most notably AOAC, BATC, and WSOC—which post higher MCC scores and narrower standard deviations, reflecting improved reliability and uniformity in classification results. On the CNS dataset, for instance, BATC improves its MCC from 0.85584 to 0.89657 while reducing variability, indicating more dependable classification performance. These gains are particularly important for datasets that demand precise and trustworthy predictions, such as Lung Cancer, where both BATC and WSOC achieve substantial improvements in MCC alongside greater stability.

Table 5. Comparison between proposed and original methods based on Fitness Values.

Datasets	Measure	MPAC	PSO	PSOC	WSO	WSOC	AOA	AOAC	BAT	BATC	MPA
CNS	AVG	0.05103	0.06700	0.06875	0.05424	0.04791	0.06220	0.05652	0.06911	0.04941	0.052
	STD	0.00012	0.01687	0.01935	0.00664	0.00638	0.01373	0.00892	0.01022	0.00552	0.004
Colon	AVG	0.03605	0.08155	0.07484	0.08182	0.06790	0.05152	0.06334	0.05902	0.05756	0.031
	STD	0.01037	0.01614	0.01910	0.01036	0.00538	0.01941	0.01606	0.01186	0.01530	0.008
Leukemia_3c	AVG	0.00081	0.00215	0.00177	0.00314	0.00130	0.00139	0.00129	0.00319	0.00129	0.000
	STD	0.00005	0.00050	0.00059	0.00043	0.00044	0.00039	0.00046	0.00042	0.00035	0.000
Leukemia_4c	AVG	0.01431	0.02383	0.02206	0.01568	0.00820	0.02549	0.01952	0.02245	0.00956	0.014
	STD	0.00329	0.01330	0.01300	0.00450	0.00666	0.01534	0.01137	0.00605	0.00663	0.003
Leukemia	AVG	0.00060	0.00334	0.00223	0.00221	0.00065	0.00077	0.00071	0.00248	0.00081	0.000
	STD	0.00000	0.00438	0.00382	0.00053	0.00009	0.00016	0.00015	0.00043	0.00021	0.000

Table 6. Comparison between the proposed and original methods based on MCC.

[illegible]

[illegible]

Table 7 presents a side-by-side comparison of the precision improvements achieved by augmenting standard optimization algorithms with MPAC. The analysis indicates that hybrids—including AOAC, BATC, and WSOC—typically achieve higher average precision values and lower variability across several demanding datasets. In the CNS dataset, BATC markedly raises precision from 0.95444 to 0.97300, reflecting a clear gain in classification exactness. On Leukemia_4c, BATC further advances precision from 0.97637 to 0.99076, paired with a reduced standard deviation that signals not only elevated precision but also more consistent outcomes. Additionally, incorporating MPAC contributes to result stability, as demonstrated by lower standard deviations in datasets such as MLL, where WSOC increases precision from 0.98973 to 0.99627.

Table 7. Comparison between the proposed and original methods in terms of precision.

[illegible]

Table 8 demonstrates that hybrid approaches incorporating MPAC reliably deliver equal or better sensitivity across more than 90% of the datasets tested. To illustrate, AOAC exceeds AOA's sensitivity in five out of the ten datasets and matches it in four others, reflecting a clear and consistent advantage. BATC outperforms BAT across all datasets evaluated, with particularly substantial gains in CNS (where sensitivity rises from 0.94359 to 0.95299) and Leukemia_4c (where it climbs from 0.95484 to 0.98158).

Table 8. Comparison between the proposed and original methods in terms of sensitivity.

Datasets	Measure	MPAC	PSO	PSOC	WSO	WSOC	AOA	AOAC	BAT	BATC	MPA
CNS	AVG	0.94957	0.89060	0.88718	0.94701	0.95385	0.94274	0.94701	0.94359	0.95299	0.948
	STD	0.00468	0.03006	0.03340	0.00936	0.01043	0.01457	0.01335	0.01242	0.01182	0.009
Colon	AVG	0.99242	0.79394	0.78636	0.88333	0.90606	0.95303	0.92576	0.94091	0.93182	0.993
	STD	0.02096	0.04089	0.03408	0.03086	0.02367	0.05402	0.04218	0.04329	0.04092	0.019
Leukemia_3c	AVG	1	0.98912	0.99124	1	1	1	1	1	1	1
	STD	0	0.00870	0.00770	0	0	0	0	0	0	0
Leukemia_4c	AVG	0.99386	0.80647	0.80283	0.96360	0.99342	0.92709	0.94326	0.95484	0.98158	0.990
	STD	0.00167	0.08066	0.07730	0.02936	0.01553	0.07387	0.06487	0.02967	0.02718	0.014
Leukemia	AVG	1	1	1	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0	0	0	0
Lung_Cancer	AVG	0.99872	0.99615	0.99444	0.99573	0.99936	0.99765	0.99829	0.99209	0.99850	0.997
	STD	0.00261	0.00548	0.00577	0.00350	0.00196	0.00394	0.00288	0.00435	0.00276	0.003
Lymphoma	AVG	1	1	1	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0	0	0	0
MLL	AVG	0.99270	0.96287	0.96310	0.98737	0.99589	0.98007	0.98561	0.98556	0.99563	0.994
	STD	0.00805	0.00719	0.00690	0.00746	0.00763	0.00930	0.01038	0.00576	0.00691	0.008
Ovarian	AVG	1	1	1	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0	0	0	0
SRBCT	AVG	1	1	1	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0	0	0	0

In a similar vein, MPAC registers modest gains over standard MPA on several datasets—for example, advancing sensitivity from 0.94872 to 0.94957 on CNS—while showing more noticeable progress on Leukemia_4c, where sensitivity improves from 0.99013 to 0.99386. Such increases in sensitivity are particularly important for medical diagnostics, as correctly identifying positive cases can directly and meaningfully affect clinical decision-making and patient prognosis.

PSOC and WSOC likewise exhibit enhanced performance relative to their baseline counterparts, PSO and WSO. PSOC achieves a small but positive lift over PSO on Leukemia_3c, moving sensitivity from 0.98912 to 0.99124. Meanwhile, WSOC delivers a striking improvement on Lung Cancer, boosting average sensitivity from 0.99573 to 0.99936.

Results on feature selection efficiency, summarized in **Table 9**, reveal that the hybridization strategy typically yields fewer selected features while preserving—or even improving—overall model effectiveness. On the CNS dataset, for instance, AOAC reduces the average feature count from 8.5 (AOA) to 7.6, indicating superior parsimony. Even more pronounced is BATC's performance on the same dataset, where the average number of selected features drops sharply from 18.3 to 10.53333, underscoring a markedly more streamlined selection process.

Table 9. Comparison between proposed and original methods based on Selected Features.

Datasets	Measure	MPAC	PSO	PSOC	WSO	WSOC	AOA	AOAC	BAT	BATC
CNS	AVG	7.63333	13.26667	11	15.43333	8.53333	8.50000	7.6	18.3	10.53333
	STD	0.61495	2.91173	3.33218	2.78770	1.83328	2.44597	1.30252	2.70568	2.16131
Colon	AVG	6.23333	11.73333	8.43333	17.83333	10.7	7.73333	7.5	15.4	11.23333
	STD	0.89763	2.61209	2.12835	2.91350	1.72507	2.89986	3.54041	3.65400	1.86960
Leukemia_3c	AVG	4.06667	10.73333	8.86667	15.7	6.5	6.93333	6.46667	15.93333	6.46667
	STD	0.25371	2.50425	2.93297	2.16795	2.19325	1.96404	2.31537	2.08332	1.75643
Leukemia_4c	AVG	5.53333	12.46667	8.93333	16.53333	10.63333	8.73333	7.73333	19.13333	10.4
	STD	1.63440	3.19194	2.62525	3.00268	3.45895	2.80312	2.46259	2.93297	3.93569
Leukemia	AVG	3	8.43333	4.96667	11.06667	3.26667	3.83333	3.56667	12.4	4.06667
	STD	0	2.66113	1.90251	2.65139	0.44978	0.79148	0.72793	2.14315	1.04826
Lung_Cancer	AVG	4.56667	11.06667	8.4	17.03333	9	7.1	6.93333	18.4	12.1
	STD	1.07265	1.96404	2.45792	4.35877	4.32315	2.32453	2.70291	3.13600	3.92472
Lymphoma	AVG	2	4.6	3	4.36667	2.03333	2.86667	2.73333	8.46667	2.93333
	STD	0	1.52225	0.78784	0.99943	0.18257	0.68145	0.52083	1.27937	0.73968
MLL	AVG	6.20000	12.06667	9.03333	17.23333	11.2	7.66667	8.03333	17.3	12.13333
	STD	3.11171	2.39156	2.82212	3.95390	1.93694	2.66954	2.73525	3.06425	2.90936
Ovarian	AVG	3	4.73333	3.3	5.16667	3	3.26667	3.03333	9.5	3.06667
	STD	0	1.25762	0.65126	1.26173	0	0.63968	0.18257	1.04221	0.25371
SRBCT	AVG	5.83333	10.33333	7.86667	11.23333	6.2	7.4	6.86667	13.3	7.43333
	STD	0.53067	2.00574	1.61316	1.54659	0.48423	1.61031	0.97320	1.36836	1.27802

Comparable reductions appear across additional datasets. In Leukemia_4c, MPAC trims the average feature count from MPA's 6.1 down to 5.53333, demonstrating an improved capacity to isolate the most discriminative genes. This ability to shrink the feature set without compromising predictive power is especially valuable in genomic studies, where compact and interpretable models facilitate faster computation and clearer biological insight.

Hybridization also tends to narrow the standard deviation of selected feature counts, pointing to greater run-to-run consistency in the feature selection outcome. This stabilization effect stands out particularly in challenging, high-dimensional datasets such as MLL and Lung Cancer, where the hybrids reliably identify fewer yet more informative

features. A clear example is WSOC on Lung Cancer, which lowers the average feature count from 17.03333 to 9, thereby substantially elevating the overall efficiency of the selection mechanism.

Comparison of the hybridization approaches

This section presents, visualizes, and evaluates the performance of the proposed MPAC method together with its modified variants (AOAC, BATC, EOC, GWOC, PSOC, and WSOC). The comparison relies on several key performance indicators: classification accuracy, sensitivity, specificity, number of selected features, and fitness scores.

According to the mean classification accuracy values reported in **Table 10**, WSOC achieved the best or joint-best performance on 8 of the 10 datasets. MPAC and BATC ranked next, each securing the top position on six datasets. AOAC and EOC achieved leadership in five datasets apiece, while GWOC led in four. When examining the standard deviation (STD) of accuracy, MPAC showed the smallest variation across eight of the ten datasets, indicating superior stability. BATC and WSOC showed similarly low variability across six datasets each.

The average F1-score results, summarized in [Table 11](#), show that WSOC attained the highest or tied for the highest value across eight of the ten datasets. MPAC and BATC were close behind, each claiming the lead in six datasets. AOAC and EOC performed best across five datasets each, and GWOC topped four. For the F1-score standard deviation, MPAC again proved most reliable, exhibiting the least fluctuation across eight datasets. BATC and WSOC also showed excellent stability, with each recording the lowest STD across six datasets.

Table 10. Performance of the MPAC and other modified versions in terms of classification accuracy.

[illegible]

	STD	0	0	0	0	0	0	0
SRBCT	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0

In terms of Matthew's Correlation Coefficient (MCC), the pattern remained consistent: WSOC recorded the highest or equivalent results in eight of the ten datasets, as indicated in **Table 12**. MPAC and BATC trailed closely, each leading in six datasets. AOAC and EOC excelled in five datasets each, whereas GWOC was the frontrunner in four. Looking at the STD of MCC, MPAC demonstrated the greatest robustness, showing the smallest variation in eight datasets. BATC and WSOC likewise achieved strong consistency, each placing first in stability for six datasets.

Table 11. Performance of the MPAC and other modified versions in terms of F1.

Datasets	Measure	MPAC	PSOC	WSOC	AOAC	BATC	EOC	GWOC
CNS	AVG	0.96107	0.89433	0.96370	0.95683	0.96284	0.95905	0.95473
	STD	0.00018	0.01325	0.00542	0.00700	0.00466	0.00693	0.00973
Colon	AVG	0.95206	0.83474	0.90596	0.91248	0.92096	0.92466	0.90290
	STD	0.01447	0.01744	0.00898	0.02330	0.02220	0.03105	0.02120
Leukemia_3c	AVG	1	0.98977	1	1	1	1	1
	STD	0	0.00834	0	0	0	0	0
Leukemia_4c	AVG	0.98510		0.99167		0.98477		
	STD	0.00446		0.01172		0.01745		
Leukemia	AVG	1	0.99193	1	1	1	1	0.99930
	STD	0	0.00453	0	0	0	0	0.00267
Lung_Cancer	AVG	0.99362	0.98654	0.99300	0.99215	0.99383	0.99235	0.99235
	STD	0.00002	0.00160	0.00128	0.00214	0.00165	0.00178	0.00196
Lymphoma	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
MLL	AVG	0.99344	0.96574	0.99599	0.98637	0.99589	0.98815	0.98283
	STD	0.00715	0.00711	0.00718	0.00957	0.00641	0.00937	0.00689
Ovarian	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
SRBCT	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0

Regarding specificity, detailed in **Table 13**, both WSOC and BATC exhibited the strongest or jointly strongest performance in seven of the ten datasets. MPAC ranked second, delivering top results across six datasets. AOAC and EOC distinguished themselves in five datasets each, while GWOC led in four. Regarding the standard deviation of specificity,

MPAC and BATC tied for the most consistent behavior, with the lowest variation across seven datasets. WSOC also maintained impressive reliability, achieving the smallest STD in six datasets.

Table 12. Performance of MPAC and other modified versions in terms of MCC.

Datasets	Measure	MPAC	PSOC	WSOC	AOAC	BATC	EOC	GWOC
CNS	AVG	0.89181	0.70548	0.89886	0.87980	0.89657	0.88580	0.87365
	STD	0.00044	0.03691	0.01418	0.01986	0.01213	0.01915	0.02878
Colon	AVG	0.92621	0.75635	0.85467	0.86473	0.87767	0.88389	0.85019
	STD	0.02249	0.02164	0.01305	0.03629	0.03443	0.04742	0.03175
Leukemia_3c	AVG	1	0.98446	1	1	1	1	1
	STD	0	0.01109	0	0	0	0	0
Leukemia_4c	AVG	0.98030		0.98964		0.98294		
	STD	0.00561		0.01334		0.01829		
Leukemia	AVG	1	0.97664	1	1	1	1	0.99797
	STD	0	0.01311	0	0	0	0	0.00773
Lung_Cancer	AVG	0.97229	0.94115	0.96950	0.96580	0.97321	0.96672	0.96673
	STD	0.00001	0.00711	0.00565	0.00948	0.00723	0.00784	0.00865
Lymphoma	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
MLL	AVG	0.99030	0.94864	0.99394	0.97955	0.99385	0.98223	0.97443
	STD	0.01056	0.01075	0.01082	0.01429	0.00957	0.01403	0.01041
Ovarian	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
SRBCT	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0

Shifting the focus to sensitivity, as outlined in [Table 14](#), WSOC either led or came very close to leading in 8 of the datasets considered. MPAC achieved top performance in six datasets. AOAC, EOC, BATC, and GEOC each delivered excellent results across five datasets. Regarding the STD of sensitivity scores, MPAC again showed the highest reliability, with the least fluctuation across eight datasets. WSOC similarly demonstrated strong stability, achieving the lowest STD across six datasets.

Table 13. Performance of MPAC and other modified versions in terms of specificity.

Datasets	Measure	MPAC	PSOC	WSOC	AOAC	BATC	EOC	GWOC
CNS	AVG	0.97289	0.90324	0.97382	0.96714	0.97300	0.96732	0.96467
	STD	0.00432	0.02701	0.00027	0.01396	0.00412	0.01364	0.01827

Colon	AVG	0.91532	0.89063	0.90642	0.90137	0.91108	0.91464	0.89571
	STD	0.02128	0.01249	0.01148	0.03292	0.01069	0.02599	0.02612
Leukemia_3c	AVG	1	0.98889	1	1	1	1	1
	STD	0	0.01123	0	0	0	0	0
Leukemia_4c	AVG	0.97758		0.99098		0.99076		
	STD	0.00700		0.01130		0.01013		
Leukemia	AVG	1	0.98403	1	1	1	1	0.99861
	STD	0	0.00896	0	0	0	0	0.00529
Lung_Cancer	AVG	0.98859	0.97882	0.98672	0.98611	0.98923	0.98693	0.98694
	STD	0.00254	0.00566	0.00341	0.00575	0.00407	0.00489	0.00515
Lymphoma	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
MLL	AVG	0.99445	0.96969	0.99627	0.98767	0.99631	0.98926	0.98496
	STD	0.00606	0.00784	0.00651	0.00860	0.00576	0.00851	0.00679
Ovarian	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
SRBCT	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0

Regarding the number of selected features, the proposed MPAC exhibited superior and highly competitive performance compared to alternative techniques. It achieved the best average (AVG) performance across 9 datasets and the best STD performance across 6 datasets, as detailed in **Table 15**. Such dominant outcomes clearly place MPAC ahead of the remaining approaches.

Table 14. Performance of MPAC and other modified versions in terms of sensitivity.

[illegible]

	STD	0	0	0	0	0	0	0
Lung_Cancer	AVG	0.99872	0.99444	0.99936	0.99829	0.99850	0.99786	0.99786
	STD	0.00261	0.00577	0.00196	0.00288	0.00276	0.00307	0.00350
Lymphoma	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
MLL	AVG	0.99270	0.96310	0.99589	0.98561	0.99563	0.98749	0.98131
	STD	0.00805	0.00690	0.00763	0.01038	0.00691	0.01013	0.00717
Ovarian	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0
SRBCT	AVG	1	1	1	1	1	1	1
	STD	0	0	0	0	0	0	0

Because performance varied across measures—with WSOC leading in some categories and MPAC in others—the fitness values of all methods were evaluated to determine which method best optimizes multiple objectives simultaneously. The fitness scores obtained by each approach appear in **Table 16**. According to these results, the proposed MPAC achieved the highest fitness across seven datasets, outperforming all other methods.

Table 15. Performance of MPAC and other modified versions across selected features.

Datasets	Measure	MPAC	PSOC	WSOC	AOAC	BATC	EOC	GWOC
CNS	AVG	7.63333	11	8.53333	7.6	10.53333	7.43333	6.56667
	STD	0.61495	3.33218	1.83328	1.30252	2.16131	1.27802	0.89763
Colon	AVG	6.23333	8.43333	10.7	7.5	11.23333	7.43333	6.56667
	STD	0.89763	2.12835	1.72507	3.54041	1.86960	1.27802	0.89763
Leukemia_3c	AVG	4.06667	8.86667	6.5	6.46667	6.46667	7.43333	6.56667
	STD	0.25371	2.93297	2.19325	2.31537	1.75643	1.27802	0.89763
Leukemia_4c	AVG	5.53333	8.93333	10.63333	7.73333	10.4	7.43333	6.56667
	STD	1.63440	2.62525	3.45895	2.46259	3.93569	1.27802	0.89763
Leukemia	AVG	3	4.96667	3.26667	3.56667	4.06667	7.43333	6.56667
	STD	0	1.90251	0.44978	0.72793	1.04826	1.27802	0.89763
Lung_Cancer	AVG	4.56667	8.4	9	6.93333	12.1	7.43333	6.56667
	STD	1.07265	2.45792	4.32315	2.70291	3.92472	1.27802	0.89763
Lymphoma	AVG	2	3	2.03333	2.73333	2.93333	7.43333	6.56667
	STD	0	0.78784	0.18257	0.52083	0.73968	1.27802	0.89763
MLL	AVG	6.2	9.03333	11.2	8.03333	12.13333	7.43333	6.56667
	STD	3.11171	2.82212	1.93694	2.73525	2.90936	1.27802	0.89763

Ovarian	AVG	3	3.3	3	3.03333	3.06667	7.43333	6.56667
	STD	0	0.65126	0	0.18257	0.25371	1.27802	0.89763
SRBCT	AVG	5.83333	7.86667	6.2	6.86667	7.43333	7.43333	6.56667
	STD	0.53067	1.61316	0.48423	0.97320	1.27802	1.27802	0.89763

Table 16. Performance of MPAC and other modified versions in terms of fitness value.

Datasets	Measure	MPAC	PSOC	WSOC	AOAC	BATC	EOC	GWOC
CNS	AVG	0.05103	0.06875	0.04791	0.05652	0.04941	0.05394	0.05936
	STD	0.00012	0.01935	0.00638	0.00892	0.00552	0.00849	0.01275
Colon	AVG	0.03605	0.07484	0.06790	0.06334	0.05756	0.05426	0.06953
	STD	0.01037	0.01910	0.00538	0.01606	0.01530	0.02112	0.01368
Leukemia_3c	AVG	0.00081	0.00177	0.00130	0.00129	0.00129	0.00123	0.00141
	STD	0.00005	0.00059	0.00044	0.00046	0.00035	0.00037	0.00051
Leukemia_4c	AVG	0.01431	0.02206	0.00820	0.01952	0.00956	0.01551	0.01907
	STD	0.00329	0.01300	0.00666	0.01137	0.00663	0.01190	0.01159
Leukemia	AVG	0.00060	0.00223	0.00065	0.00071	0.00081	0.00065	0.00163
	STD	0.00000	0.00382	0.00009	0.00015	0.00021	0.00009	0.00323
Lung_Cancer	AVG	0.01077	0.01607	0.01262	0.01349	0.01188	0.01313	0.01323
	STD	0.00025	0.00285	0.00178	0.00300	0.00212	0.00240	0.00282
Lymphoma	AVG	0.00040	0.00060	0.00041	0.00055	0.00059	0.00049	0.00054
	STD	0.00000	0.00016	0.00004	0.00010	0.00015	0.00010	0.00011
MLL	AVG	0.00713	0.02196	0.00607	0.01492	0.00632	0.01315	0.01830
	STD	0.00624	0.00919	0.00669	0.00925	0.00583	0.00887	0.00630
Ovarian	AVG	0.00060	0.00066	0.00060	0.00061	0.00061	0.00060	0.00061
	STD	0.00000	0.00013	0.00000	0.00004	0.00005	0.00000	0.00004
SRBCT	AVG	0.00117	0.00157	0.00124	0.00137	0.00149	0.00133	0.00130
	STD	0.00011	0.00032	0.00010	0.00019	0.00026	0.00015	0.00019

Comparison with state-of-the-art methods

The effectiveness of the proposed MPAC was further validated by benchmarking it against twelve well-established gene selection algorithms from the literature: mRMR-MBAO [16], SU-RSHSA [44], mRMR-DBH [19], ISFLA [45], IBCFPA [46], MIM-MFOA [47], BCROSAT [14], rMRMR-HBA [30], TOPSIS-Jaya [48], SARA [49], IG-MBKH [15], and MRMR-BA [18]. **Table 17** summarizes the outcomes of multiple independent executions for each method, reporting both classification accuracy (ACC) and the count of selected genes (#G).

As shown in **Table 17**, MPAC outperformed the competing techniques across multiple datasets in classification accuracy. Remarkably, it achieved 100% accuracy on SRBCT, Lung Cancer, Leukemia_3c, and Leukemia—either equaling or surpassing all other methods. MPAC also demonstrated particularly strong results on the CNS, Leukemia_4c, and MLL datasets, confirming its consistent effectiveness across varied gene expression patterns.

Regarding the number of selected genes, MPAC demonstrated a clear advantage, frequently identifying smaller gene subsets while maintaining excellent classification accuracy. This advantage is especially evident in datasets such as Lung Cancer and Leukemia_3c, where MPAC not only achieved the highest classification accuracy but also did so with considerably fewer genes than competing approaches. The effective trade-off between reduced gene count and strong predictive performance makes MPAC a highly suitable option for gene selection in genomic research.

Moreover, in the context of gene selection efficiency, the findings indicate that MPAC consistently achieves optimal classification accuracy with notably fewer genes. For example, on the Leukemia_3c dataset, although methods like TOPSIS-Jaya, rMRMR-HBA, and IG-MBKH select a small number of genes (6.6, 7.3, 8.8, respectively) while achieving 100% classification accuracy, MPAC surpasses them by selecting an average of only 4.06667 genes while maintaining the same perfect accuracy. Likewise, for the Lung Cancer dataset, in comparison to mRMR-MBAO, IBCFPA, TOPSIS-Jaya, and IG-MBKH—which select larger gene sets (23.833, 82.2, 9.9, 23.80) and obtain lower accuracies (98.54%, 94.44%, 94.24%, 96.12%)—MPAC dramatically lowers the gene count to 4.56667 while attaining flawless 100% classification accuracy.

These outcomes highlight how the hybrid design of MPAC improves both gene reduction efficiency and overall classification effectiveness across multiple datasets. This marked enhancement stems from the powerful combination of the MPA framework with sophisticated computational strategies, which refines the feature selection procedure and ensures the identification of the most informative yet minimal gene panels required for precise cancer classification. As summarized in **Table 17**, MPAC's exceptional capabilities underscore its promise as a top-performing bioinformatics technique for building accurate, streamlined diagnostic tools.

Table 17. A comparison between the proposed rMRMR-MPA-GWO and other state-of-the-art methods. The results are reported as classification accuracy (Average number of genes).

Datasets	Metric	ISFLA	SARA	TOPSIS-Jaya	rMRMR-HBA	IG-MBKH	MRMR-BA	MPAC	mRMR-MBAO	SU-RSHSA	mRMR-DBH	IB
CNS	# G	41.1	–	8.7	19.2	14.70	19.2	7.63333	21.37	13.15	39.75	2
Empty Cell	ACC	77.46	–	96.22	94.22	90.34	94.22	0.95	88.57	89.36	97.19	8
Ovarian	# G	33.3	6	18.50	3.8	3.40	3.83	3	5.166	20.47	2.66	4
Empty Cell	ACC	97.29	99.15	99.52	100	100	100	100	100	99.61	100	9
SRBCT	# G	43.10	5	15.80	12.8	6.3	12.83	5.83333	7.5	8.37	–	4
Empty Cell	ACC	93.72	99.81	100	100	100	100	100	100	97.98	–	9
Lung Cancer	# G	40.3	5	9.9	–	23.80	–	4.56667	23.833	–	–	8
Empty Cell	ACC	89.56	90.22	94.24	–	96.12	–	0.99004	98.54	–	–	9
Leukemia_4c	# G	32.2	–	19.50	11	15.80	11.03	5.53333	21.33	12.02	–	4
Empty Cell	ACC	90.91	–	99.72	99.77	99.44	99.77	0.98667	98.70	97.11	–	9
MLL	# G	40.7	–	12.90	19	11.10	19.03	6.2	11.28	7.83	5.25	4

Empty Cell	ACC	92.59	–	99.62	79.86	99.72	79.86	0.99405	99.64	99.94	100	9
Leukemia	# G	35.8	7	16.10	5.2	4.2	5.23	3	8.71	21.64	4	2
Empty Cell	ACC	96.34	97.65	100	100	100	100	100	100	100	100	9
Leukemia_3c	# G	40.0	7	6.6	7.3	8.80	7.27	4.06667	12.28	10.72	–	4
Empty Cell	ACC	94.00	98.02	100	100	100	100	100	100	100	–	9

Conclusion

The challenge of detecting cancer-related gene expression signatures in microarray data remains complex owing to extreme dimensionality, intricate patterns, and inherent noise. Pinpointing a compact set of biomarker genes is essential for constructing a robust cancer classification framework that supports clinical diagnosis, especially for early detection.

This research introduces an innovative gene selection strategy that integrates a filter-wrapper framework to effectively tackle the gene selection challenge in cancer classification. The proposed technique aims to extract a compact set of genes rich in biologically relevant signals capable of distinguishing cancerous from normal samples. It leverages rMRMR as the filtering stage and applies a hybrid optimizer that merges MPA with a crossover operator, termed MPAC. This integrated model strengthens the exploration capacity of MPA agents, enabling them to discover optimal biomarker gene subsets for more reliable cancer categorization.

The proposed method underwent thorough validation through extensive experiments on nine well-established microarray datasets, structured as a three-stage comparative evaluation. In the first stage, standard optimization algorithms were pitted against their crossover-enhanced hybrid counterparts, with the modified versions typically outperforming or matching the originals across most datasets and performance metrics. The second stage involved head-to-head comparisons among hybrid optimization variants, in which MPAC consistently delivered the strongest or highly competitive results across most datasets. In the final stage, benchmarking MPAC against prominent state-of-the-art gene selection techniques confirmed that it achieved superior performance in the majority of evaluated aspects.

Although the evaluation of the MPAC approach was conducted using nine well-established microarray datasets that differ in scale and complexity, the validation remains somewhat restricted in scope. While these datasets provide valuable insights and support fair benchmarking against prior techniques, extending the assessment to a larger, more varied dataset collection would better confirm the broad applicability of the proposed method.

MPAC's strong performance in gene selection underscores its promise as a practical, powerful tool for discovering key cancer biomarkers. Given the encouraging outcomes achieved in tackling gene selection challenges, several exciting avenues for future investigation could further strengthen and expand the framework's capabilities:

- Although the k-nearest neighbor classifier was employed in this work, investigating alternative classification methods may yield better results. Promising candidates include Convolutional Neural Networks (CNNs), which have demonstrated effectiveness in analyzing gene expression profiles, Deep Belief Networks (DBNs), valued for their ability to capture probabilistic patterns, and Support Vector Machines (SVMs), widely appreciated for their reliability in high-dimensional environments. Incorporating such classifiers could enhance overall predictive performance.
- The MPAC framework could benefit from adopting cutting-edge optimization algorithms. Introducing and testing new optimization operators could refine the MPA component, thereby increasing its ability to handle the diverse and demanding optimization tasks inherent in complex gene selection problems.

- The gene selection strategies developed here hold potential for application well beyond cancer classification in bioinformatics. Promising extensions include drug discovery, where accurate gene prioritization is essential; sequence alignment, a cornerstone of genomic analysis; and protein structure prediction, which plays a vital role in elucidating biological processes and advancing therapeutic development.
- This study assessed the effectiveness of the identified gene subsets using individual benchmark datasets. In future efforts, emphasis could be placed on confirming the stability and transferability of these selected biomarkers by testing them on independent datasets representing the same cancer type. Such cross-validation would enhance the clinical significance of the results and assess the reliability of the selected features across studies.
- The present MPAC implementation lacks a detailed examination of how its parameter choices affect outcomes. Subsequent research could investigate the influence of variations in critical parameters—including population size, iteration count, and crossover mechanism—on both the effectiveness and consistency of the gene selection procedure. These analyses would provide valuable insight into the method's robustness and replicability.
- While the current investigation focused on microarray-based gene expression data, the MPAC approach is inherently flexible and suitable for other high-dimensional biological datasets. With appropriate data preparation and scaling, future studies could adapt and apply the framework to additional omics fields, such as RNA-seq, proteomics, and epigenomics, where effective feature selection remains a major challenge.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Mar 2023

Revised: 11 May 2023

Accepted: 28 Jun 2023

Published online: 20 September 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(5439):531–7.
- Reddy GD, Brahmaiah V, Kavitha B, Nethagani J, Palaparthi EC, Lebaka RR, et al. Role of epigenetics variation with focus on DNA methylation in triple-negative breast cancer patients. *Arch Med Heal Sci*. 2023;11(1):96–110.
- Zheng X, Zhang C. Gene selection for microarray data classification via dual latent representation learning. *Neurocomputing*. 2021;461:266–80.
- El Akadi A, Amine A, El Ouardighi A, Aboutajdine D. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowl Inf Syst*. 2011;26(3):487–500.
- Lai C-M, Yeh W-C, Chang C-Y. Gene selection using information gain and improved simplified swarm optimization. *Neurocomputing*. 2016;218:331–8.
- Deepa R, Arunkumar S, Jayaraj V, Sivasamy A. Healthcare's new frontier: AI-driven early cancer detection for improved well-being. *AIP Adv*. 2023;13(11).
- Al-Raeei M. Shaping the future of cancer treatment: The commitment of medical physicists. *AIP Adv*. 2024;14(9):090703.
- Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell*. 1997;19(2):153–8.
- Almazrui H, Alshamlan H. A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data. *IEEE Access*. 2022;10:71427–49.
- Li BQ, Hu LL, Chen L, Feng KY, Cai YD, Chou KC. Prediction of protein domain with mRMR feature selection and analysis. *PLoS One*. 2012;7(6):e39308.
- Alshamlan HM, Badr GH, Alohalı YA. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. *Comput Biol Chem*. 2015;56:49–60.
- Zhu Z, Ong YS. Memetic algorithms for feature selection on microarray data. In: *International Symposium on Neural Networks*. Springer; 2007. p. 1327–35.
- Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles. *Appl Soft Comput*. 2017;50:124–34.
- Yan C, Ma J, Luo H, Patel A. Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. *Chemom Intell Lab Syst*. 2019;184:102–11.
- Zhang G, Hou J, Wang J, Yan C, Luo J. Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. *Interdiscip Sci Comput Life Sci*. 2020;12(3):288–301.
- Pashaei E. Mutation-based Binary Aquila optimizer for gene selection in cancer classification. *Comput Biol Chem*. 2022;101:107767.
- Chuang L-Y, Yang C-H, Li J-C, Yang C-H. A hybrid BPSO-CGA approach for gene selection and classification of microarray data. *J Comput Biol*. 2012;19(1):68–82.
- Alomari OA, Khader AT, Al-Betar MA, Abualigah LM. Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm. *Int J Data Min Bioinform*. 2017;19(1):32–51.

Pashaei E. Gene selection using hybrid dragonfly black hole algorithm: A case study on RNA-seq COVID-19 data. *Anal Biochem*. 2021;627:114242.

Alomari OA, Makhadmeh SN, Al-Betar MA, Alyasseri ZAA, Doush IA, Abasi AK, et al. Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators. *Knowl-Based Syst*. 2021;223:107034.

Faramarzi A, Heidarinejad M, Mirjalili S, Gandomi AH. Marine predators algorithm: A nature-inspired metaheuristic. *Expert Syst Appl*. 2020;152:113377.

Houssein EH, Hassaballah M, Ibrahim IE, Abdelminaam DS, Wazery YM. An automatic arrhythmia classification model based on improved marine predators algorithm and convolutional neural networks. *Expert Syst Appl*. 2022;187:115936.

Wang N, Wang J-S, Zhu L, Wang H, Wang G. A novel dynamic clustering method by integrating marine predators algorithm and particle swarm optimization algorithm. *IEEE Access*. 2020;9:3557–69.

Abdel-Basset M, Mohamed R, Elhoseny M, Bashir AK, Jolfaei A, Kumar N. Energy-aware marine predators algorithm for task scheduling in IoT-based fog computing applications. *IEEE Trans Ind Inform*. 2020;17(7):5068–76.

Abd Elaziz M, Ewees AA, Yousri D, Alwerfali HSN, Awad QA, Lu S, et al. An improved marine predators algorithm with fuzzy entropy for multi-level thresholding: Real world example of COVID-19 CT image segmentation. *IEEE Access*. 2020;8:125306–30.

Abdel-Basset M, Mohamed R, Elhoseny M, Chakraborty RK, Ryan M. A hybrid COVID-19 detection model using an improved marine predators algorithm and a ranking-based diversity reduction strategy. *IEEE Access*. 2020;8:79521–40.

Yousri D, Babu TS, Beshr E, Eteiba MB, Allam D. A robust strategy based on marine predators algorithm for large scale photovoltaic array reconfiguration to mitigate the partial shading effect on the performance of PV system. *IEEE Access*. 2020;8:112407–26.

Fu Q, Li Q, Li X. An improved multi-objective marine predator algorithm for gene selection in classification of cancer microarray data. *Comput Biol Med*. 2023;160:107020.

Osama S, Ali AA, Shaban H. Gene selection based on recursive spider wasp optimizer guided by marine predators algorithm. *Neural Comput Appl*. 2024;36(28):17327–44.

Alomari OA, Khader AT, Al-Betar MA, Awadallah MA. A novel gene selection method using modified MRMR and hybrid bat-inspired algorithm with β -hill climbing. *Appl Intell*. 2018;48(11):4429–47.

Kononenko I. Estimating attributes: analysis and extensions of RELIEF. In: *European Conference on Machine Learning*. Springer; 1994. p. 171–82.

Su C-T, Hsu J-H. An extended chi2 algorithm for discretization of real value attributes. *IEEE Trans Knowl Data Eng*. 2005;17(3):437–41.

Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat*. 1951;22(1):79–86.

Duval B, Hao J-K, Hernandez JC. A memetic algorithm for gene selection and molecular classification of cancer. In: *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (GECCO)*; 2009. p. 201–8.

Dash M, Liu H. Feature selection for classification. *Intell Data Anal*. 1997;1(3):131–56.

Li X, Yin M. Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Trans Nanobioscience*. 2013;12(4):343–53.

Zhu Y, Li W, Li T. A hybrid artificial immune optimization for high-dimensional feature selection. *Knowl-Based Syst*. 2023;260:110111.

Hijazi NM, Faris H, Aljarah I. A parallel metaheuristic approach for ensemble feature selection based on multi-core architectures. *Expert Syst Appl*. 2021;182:115290.

Yang X, Zhao D, Yu F, Heidari AA, Bano Y, Ibrohimov A, et al. Boosted machine learning model for predicting intradialytic hypotension using serum biomarkers of nutrition. *Comput Biol Med*. 2022;147:105752.

Abualigah L, Diabat A, Mirjalili S, Abd Elaziz M, Gandomi AH. The arithmetic optimization algorithm. *Comput Methods Appl Mech Eng*. 2021;376:113609.

Yang X-S. A new metaheuristic bat-inspired algorithm. In: *Nature-Inspired Cooperative Strategies for Optimization*. Springer; 2010. p. 65–74.

Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: *Proceedings of the Sixth International Symposium on Micro Machine and Human Science (MHS'95)*. IEEE; 1995. p. 39–43.

Braik M, Hammouri A, Atwan J, Al-Betar MA, Awadallah MA. White shark optimizer: A novel bio-inspired meta-heuristic algorithm for global optimization problems. *Knowl-Based Syst*. 2022;243:108457.

Shreem SS, Ahmad Nazri MZ, Abdullah S, Sani NS. Hybrid symmetrical uncertainty and reference set harmony search algorithm for gene selection problem. *Mathematics*. 2022;10(3):374.

Hu B, Dai Y, Su Y, Moore P, Zhang X, Mao C, et al. Feature selection for optimized high-dimensional biomedical data using an improved shuffled frog leaping algorithm. *IEEE/ACM Trans Comput Biol Bioinform*. 2016;15(6):1765–73.

Yan C, Ma J, Luo H, Zhang G, Luo J. A novel feature selection method for high-dimensional biomedical data based on an improved binary clonal flower pollination algorithm. *Hum Hered*. 2019;84(1):34–46.

Dabba A, Tari A, Meftali S, Mokhtari R. Gene selection and classification of microarray data method based on mutual information and moth flame algorithm. *Expert Syst Appl*. 2021;166:114012.

Chaudhuri A, Sahu TP. A hybrid feature selection method based on binary Jaya algorithm for micro-array data classification. *Comput Electr Eng*. 2021;90:106963.

Baliarsingh SK, Muhammad K, Bakshi S. SARA: A memetic algorithm for high-dimensional biomedical data. *Appl Soft Comput*. 2021;101:107009.