Cancer Classification through Gene Selection Using the Social Spider Optimization Algorithm

Chahira Cherif[1], Mohammed Maiza[2], Samira Chouraqui[3], Abdelmalik Taleb-Ahmed[4]
[1]LRIIR, Faculty of Medicine, University of Oran 1, Ahmed Ben bella, Algeria
[2]Faculty of Exact and Applied Sciences, University of Oran 1, Ahmed Ben bella, Algeria
[3]Faculty of Mathematics and Computer Science, University of Sciences and Technology of Oran, Algeria
[4]Laboratory of IEMN, CNRS, Centrale Lille, UMR 8520, Univ. Polytechnique Hauts-de-France, Valenciennes, France
E-mail: cherif.chahira@univ-oran1.dz, maiza.mohammed@univ-oran1.dz, samira.chouraqui@univ-usto.dz, abdelmalik.taleb-ahmed@uphf.fr

Keywords: Microarray data, gene selection, social spider optimization, machine learning classifiers, mutual information, cancer classification

Received: May 6, 2025

Cancer is a leading cause of global mortality, underscoring the need for advanced diagnostic tools to enable early and accurate detection. Microarray technology allows for the simultaneous analysis of thousands of genes, offering valuable insights into cancer biology. However, the high dimensionality of microarray data presents significant challenges for classification tasks. In this study, we propose a novel approach that integrates the Social Spider Optimization (SSO) algorithm with mutual information-based feature selection to identify the most discriminative genes for cancer classification. We evaluate the performance of four machine learning classifiers—Decision Tree (DT), K-Nearest Neighbors (K-NN), Neural Networks (NN), and Support Vector Machines (SVM)—with and without feature selection. Our results demonstrate that the SSO algorithm significantly enhances classification accuracy, with SVM achieving near-perfect performance on leukemia and lymphoma datasets when combined with Max-Relevance Min-Redundancy (MRMR) feature selection. This hybrid approach provides a robust solution for cancer diagnosis by addressing key challenges such as data redundancy and computational complexity.

Povzetek: Za klasifikacijo raka so uporabili optimizacijo (SSO), združeno z merili vzajemne informacije (MIM, JMI, MRMR), za izbiro najbolj diskriminativnih genov in zmanjšanje redundance. Na zbirkah Colon, Prostate, Leukemia, Lymphoma z DT, K-NN, NN, SVM kombinacija SSO+MRMR doseže odlične rezultate (levkemija/limfom) ter zniža računsko zahtevnost.

1 Introduction

Cancer is a complex and heterogeneous disease characterized by uncontrolled cell growth and proliferation. Early and accurate diagnosis is critical for effective treatment and improved patient outcomes. Recent advances in molecular biology, particularly microarray technology, have revolutionized cancer research by enabling the simultaneous measurement of gene expression levels across thousands of genes [1]. These high-throughput datasets provide unprecedented opportunities to identify molecular signatures associated with specific cancer types [2]. However, the high dimensionality of microarray data—where the number of features (genes) far exceeds the number of samples—poses significant challenges for classification tasks. This ”curse of dimensionality” can lead to overfitting, increased computational complexity, and reduced model interpretability [3].

Feature selection is a crucial step in microarray data analysis, as it helps identify biologically relevant genes while minimizing noise and redundancy. Conventional feature selection approaches are typically classified into three main categories: filter, wrapper, and embedded methods [4]. Filter techniques, such as mutual information-based selection, rank genes based on statistical criteria without involving a predictive model. Wrapper methods employ a specific machine learning algorithm to evaluate the performance of different feature subsets. Embedded approaches integrate feature selection directly into the classifier's training process, optimizing both model accuracy and feature relevance. Despite their effectiveness, these methods often suffer from limitations such as local optima convergence and high computational complexity, particularly in high-dimensional spaces [5].

Metaheuristic optimization algorithms, inspired by natural phenomena, have emerged as powerful tools for addressing complex feature selection problems. Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) are among the most widely used metaheuristics in this context [6, 7, 8]. However, these methods may still struggle with premature convergence or parameter sensitivity, limiting their applicability to ultra-

high-dimensional datasets.

To overcome these limitations, we propose the Social Spider Optimization (SSO) algorithm, a novel metaheuristic inspired by the cooperative foraging behavior of social spiders. SSO leverages vibration-based communication among spiders to dynamically adjust search intensity, balancing exploration and exploitation in the feature space. This unique mechanism allows SSO to efficiently navigate high-dimensional datasets and identify optimal gene subsets without extensive parameter tuning [8].

In this study, we integrate SSO with mutual information-based feature selection criteria—Mutual Information Maximization (MIM), Joint Mutual Information (JMI), and Max-Relevance Min-Redundancy (MRMR)—to enhance cancer classification accuracy [9]. We evaluate the performance of four classifiers (DT, K-NN, NN, SVM) on four cancer datasets (Colon Cancer, Prostate Tumor, Leukemia, and Lymphoma). The microarray datasets were subjected to rigorous preprocessing to ensure data quality. Our results demonstrate that the SSO algorithm significantly outperforms traditional feature selection methods, achieving superior classification accuracy and computational efficiency [10].

The remainder of this paper is structured as follows: First, we present the methodology, detailing the SSO algorithm, feature selection approaches, and classification models. Next, we discuss the experimental results and comparative analysis. Then, we examine the advantages and limitations of the proposed approach. Finally, we conclude the paper and outline future research directions.

2 The social spider optimization (SSO)

The Social Spider Optimization (SSO) algorithm is a nature-inspired metaheuristic that mimics the cooperative foraging behavior of social spiders to solve complex optimization problems. In cancer genomics, SSO excels at selecting highly discriminative genes for classification tasks [11]. The algorithm evaluates candidate gene subsets using a fitness function, where a high score indicates an optimal subset that maximizes classification accuracy while minimizing redundant features [12, 13].

This fitness function is defined as :

$$\text{Fitness}(S) = \alpha \cdot \text{Accuracy}(S) + (1-\alpha) \cdot \left(1 - \frac{|S|}{N}\right) \quad (1)$$

Where:

- S represents a candidate gene subset.
- Accuracy(S) denotes the classification performance using features in S.
- $|S|$ is the cardinality of the selected subset.
- N is the total number of available genes.
- $\alpha \in [0, 1]$ controls the trade-off between accuracy and feature reduction.

The search process in SSO is guided by vibrations, which simulate the collective behavior of a spider colony. Each spider (representing a candidate solution) updates its position based on vibrations from fitter neighbors. This mechanism balances exploitation (moving toward high-quality solutions) and exploration (maintaining population diversity to avoid premature convergence) [14]. The result is an adaptive search strategy that efficiently navigates high-dimensional genomic data.

The position update for each spider i at iteration t is calculated as :

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \left(\sum_{j \in \mathcal{N}_i} \frac{\boldsymbol{x}_j^t - \boldsymbol{x}_i^t}{\|\boldsymbol{x}_j^t - \boldsymbol{x}_i^t\|} \cdot \phi_j \right) + \epsilon \quad (2)$$

Where

- $\boldsymbol{x}_i^t$ represents the current position of spider i.
- $\mathcal{N}_i$ is the set of neighboring spiders.
- ϕ_j is the vibration intensity from spider j (proportional to its fitness).
- ϵ is a small random perturbation that encourages exploration.

Finally, the selected genes are fed into machine learning classifiers to predict cancer types.

The optimization for a DT classifier focuses on finding the best splits at each node to minimize a loss function, often based on Information Gain or Gini impurity (Minimize the impurity measure at each split):

$$\min_{\text{split}} \left(I(D) - \sum_j \frac{N_j}{N} I(D_j) \right) \quad (3)$$

Where:

- $I(D)$: Impurity of the parent node.
- N: Total number of samples in the parent node.
- N_j: Number of samples in child node j.
- $I(D_j)$: Impurity of child node j.

The optimization for K-NN is expressed as:

$$y_{\text{pred}} = \text{argmax}_{y_k} \sum_{i=1}^{K} I(y_i, y_k) \quad (4)$$

Where:

- y_{pred}: Predicted class label for the new point x.
- argmax_{y_k}: The class label y_k that maximizes the sum across the classes.

– K: Number of closest neighbors considered for the classification.

– $I(y_i, y_k)$: Indicator function that equals 1 if the class label of the i-th neighbor y_i matches the predicted class label y_k, and 0 otherwise.

The optimization for NN involves minimizing a loss function that quantifies the difference between the predicted outputs of the network and the actual target values. Here's a detailed formulation:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i) \tag{5}$$

Where:

– $L(\theta)$: The overall loss of the neural network, dependent on parameters θ.

– N: The total number of samples in the dataset.

– $\mathcal{L}(y_i, \hat{y}_i)$: The loss for the i-th sample, measuring how well the predicted output $\hat{y}_i$ aligns with the true target y_i.

The SVM optimization problem is formulated as :

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)) \tag{6}$$

Where

– $\boldsymbol{w}$ is the weight vector.

– C is a regularization parameter.

– y_i are the class labels.

We evaluated the pipeline using mean ± standard deviation (SD) and 95% confidence intervals (CI) for F1-score, precision, recall, and accuracy over 10 randomized runs. To ensure robustness, we combined 10-fold cross-validation with a 70-30 train-test split, mitigating overfitting risks. The results, averaged across folds, demonstrate that SSO's biologically inspired optimization enhances both accuracy and interpretability in cancer classification [19, 20].

2.1 Runtime analysis

The runtime of the SSO algorithm depends on several factors:

– **Population Size (P)**: The number of spiders (candidate solutions) in the population. A larger population increases diversity but also computational overhead.

– **Number of Iterations (T)**: The maximum number of iterations the algorithm runs before convergence.

– **Feature Dimensionality (N)**: The total number of genes (features) in the dataset. High-dimensional data require more computations per spider.

– **Fitness Evaluation Cost (FEC)**: The cost of evaluating the fitness function for each spider, which involves training and testing a classifier on the selected gene subset.

The overall runtime can be approximated as :

$$\text{Runtime} = O(T \times P \times (N + \text{FEC})) \tag{7}$$

2.2 Computational complexity

The computational complexity of SSO is primarily determined by :

– **Position Update** : For each spider, the position update involves calculating vibrations from neighboring spiders. If each spider interacts with k neighbors, the complexity per spider per iteration is : $O(k \times N)$, where N is the dimensionality of the feature space. For the entire population, this becomes $O(P \times k \times N)$.

– **Fitness Calculation** : The fitness function involves training a classifier on the selected gene subset. Assuming the worst case where all features are selected, the complexity is dominated by the classifier's training time.

However, in practice, SSO selects a small subset of genes $d \ll N$, reducing this to $O(n^2 \times d)$.

– **Total Complexity** : Combining the above, the per-iteration complexity is :

$$O(P \times k \times N) + O(P \times n^2 \times d) \tag{8}$$

Over T iterations, The total complexity becomes:

$$O\left(T \times P \times \left(k \times N + n^2 \times d\right)\right) \tag{9}$$

3 Gene subset selection

To enhance the relevance and informativeness of the genetic data, we focused on a streamlined subset of features. This selective approach facilitates the development of accurate and robust classification models while mitigating challenges associated with high-dimensional genomic data. Gene expression datasets typically encompass thousands of features (genes), which can introduce computational inefficiencies, increased resource demands, and a heightened risk of overfitting. Thus, feature selection is essential to reduce data complexity and improve model interpretability [21].

Our goal is to retain only the most discriminative and biologically significant genes for cancer classification. By identifying and preserving genes that maximize inter-class distinction while eliminating redundant or non-informative features, we enhance model performance—boosting accuracy, recall, and generalizability [22].

In this work, we evaluate feature importance using mutual information as a key relevance metric [23, 24], ensuring that selected genes contribute meaningfully to classification while maintaining biological interpretability.

4 Mutual information

We employ mutual information (MI) to assess the statistical dependence between gene expression features and cancer class labels [25]. MI provides a robust measure of how much knowledge of a particular gene's expression reduces uncertainty about the cancer classification [26, 27]. For our high-dimensional genomic data, we implement empirical estimation methods specifically adapted to maintain accuracy in this challenging context.

We estimate MI empirically using methods adapted for high-dimensional data.

$$I(X;Y) = \sum_{x}\sum_{y} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \tag{10}$$

Where:

- $p(x,y)$ is the joint probability distribution of X and Y.
- $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

This equation quantifies the shared information between variables X and Y, measuring their mutual dependence. MI equals zero when X and Y are statistically independent, indicating no shared information between them [28].

$$I(X;Y) = 0 \quad \text{if} \quad p(x,y) = p(x)\cdot p(y) \tag{11}$$

This means that if the joint probability distribution of X and Y equals the product of their marginal distributions, then the MI is zero, indicating no dependency between the two variables.

Mutual information is linearly related to the entropies of the variables according to the following equations:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{12}$$

Where:

- $H(X)$ is the entropy of variable X.
- $H(Y)$ is the entropy of variable Y.
- $H(X,Y)$ is the joint entropy of variables X and Y.

This relationship demonstrates that MI can be understood as the reduction in uncertainty about one variable given knowledge of the other.

5 Mutual information for feature selection

Mutual information (MI) is a robust statistical measure for quantifying dependency between random variables. In feature selection, MI assesses the mutual dependence between candidate features (explanatory variables) and the target variable (predicted outcome). Features with higher MI values are prioritized, as they provide more predictive information about the target.

The scientific community has developed multiple MI-based selection criteria. In this study, we focus on three prominent methods proven effective in prior research. Their advantages and implementation details are discussed in subsequent sections.

5.1 Mutual information maximization (MIM)

MIM is a principled feature selection method that maximizes the mutual information (MI) between input features and the target variable. Grounded in information theory, MIM selects features that provide the highest information gain about the target, thereby improving predictive model performance [29].

By retaining only the most informative features and discarding non-informative ones, MIM enhances model efficiency and generalization, particularly in high-dimensional datasets where feature relevance varies significantly. The formulation for MIM can be expressed as:

$$\max_{F'\subseteq F} I(X;Y) \tag{13}$$

Where:

- F' is the subset of features selected from the original feature set F.
- $I(X;Y)$ is the MI between the selected features X and the target variable Y.

5.2 Joint mutual information (JMI)

JMI extends traditional MI-based feature selection by evaluating the joint predictive power of feature subsets. Rather than assessing features individually, JMI maximizes their combined MI with the target, capturing synergistic interactions while minimizing redundancy [30]. This approach is especially effective for high-dimensional data, where features often exhibit complex dependencies. The formulation for JMI can be expressed as:

$$\max_{F'\subseteq F} I(F';Y) \tag{14}$$

Where:

- $I(F';Y)$ is the MI between the selected features F' and the target variable Y.

5.3 Max relevance min redundancy (MRMR)

MRMR selects features that are maximally relevant to the target variable while minimizing redundancy among them. This criterion is particularly advantageous in high-dimensional settings, where reducing feature correlations improves model efficiency without compromising accuracy

[31]. MRMR achieves this balance by maximizing relevance (MI with the target) and penalizing redundant (intercorrelated) features, ensuring a diverse and informative feature set. The complete optimization problem is expressed as:

$$\max_{F' \subseteq F} \left(I(F';Y) - \frac{1}{|F'|^2} \sum_{f_i, f_j \in F'} I(f_i; f_j) \right) \quad (15)$$

where:

- F' is the subset of features selected from the original feature set F.
- $I(F', Y)$ is the MI between the selected features F' and the target variable Y.
- $I(f_i; f_j)$ is the MI between the features f_i and f_j.
- $\mid F' \mid$is the number of features in the subset F'.

6 Feature selection with SSO

After completing feature extraction and MI-based feature selection, the final stage involves building and evaluating classification models. In machine learning, classification follows a standard two-phase process: training and testing. During the training phase, the algorithm learns patterns from labeled training data to construct a predictive model [32]. The testing phase evaluates the model's performance on unseen data to assess its generalization capability and determine its readiness for real-world deployment. During this stage, the trained model undergoes rigorous evaluation to measure its predictive accuracy and overall effectiveness. This critical step ensures that the model meets the required performance thresholds before deployment.

For the classification task, we employed four well-established supervised learning algorithms: DT, K-NN, NN, and SVM. These methods were selected for their complementary strengths in handling diverse data characteristics and their proven effectiveness in similar classification tasks.

The Social Spider Optimization (SSO) algorithm was implemented to optimize gene selection by simulating the collective foraging behavior of social spiders, which dynamically adjust their search patterns based on vibratory communication within their colony.

In this approach, each spider in the population represents a candidate subset of genes, initialized randomly to ensure diversity in the search space. The fitness of each spider, corresponding to the quality of the gene subset, was evaluated using MI as the objective function, quantifying the statistical dependence between the selected genes and the target class labels. The algorithm leverages a unique vibration-based communication mechanism, where spiders share information about promising regions of the feature space through simulated vibrations, allowing the population to collectively balance exploration (global search for diverse gene combinations) and exploitation (local refinement of high-fitness subsets).

This adaptive behavior enables SSO to efficiently navigate the high-dimensional microarray data, avoiding local optima while converging toward highly discriminative gene subsets. The iterative process continues until convergence criteria are met, yielding an optimal set of genes that maximizes classification performance.

Compared to traditional metaheuristics like Genetic Algorithms or Particle Swarm Optimization, SSO demonstrates superior efficiency in feature selection due to its self-organizing nature, reduced parameter sensitivity, and ability to maintain population diversity throughout the search process.

The integration of SSO with MI criteria further enhances its biological relevance, as it prioritizes genes with strong functional associations to cancer phenotypes while minimizing redundancy. This hybrid approach addresses key limitations of conventional methods, such as premature convergence and computational inefficiency, making it particularly suited for high-dimensional genomic datasets where traditional techniques often struggle.

7 Proposed approach for cancer classification

The global healthcare community faces a critical challenge in addressing cancer, necessitating cutting-edge methods for precise diagnosis and classification. The proposed approach leverages SSO to enhance cancer classification accuracy through optimized gene selection.

The workflow begins with collecting a gene expression dataset categorized by cancer type, followed by preprocessing steps such as normalization and missing value imputation to ensure data quality. Next, the SSO algorithm identifies the most discriminative genes, mimicking the collaborative behavior of social spiders to efficiently explore the high-dimensional gene space. This step reduces redundancy and improves computational efficiency.

The selected gene subset is then analyzed using detection algorithms to identify cancer-specific patterns or anomalies. Finally, classification algorithms predict cancer types, with SSO-optimized features ensuring higher accuracy compared to traditional methods.

By integrating SSO-based gene selection with detection and classification algorithms, this approach provides a robust and scalable solution for precise cancer classification. The proposed framework is illustrated in Figure 1.

The proposed framework introduces a structured approach to enhance cancer classification accuracy using advanced computational techniques. The process begins with a cancer-labeled gene expression dataset containing genomic profiles of various tumor types. This raw biological data undergoes preprocessing to normalize values, handle missing data, and ensure quality for downstream analysis.

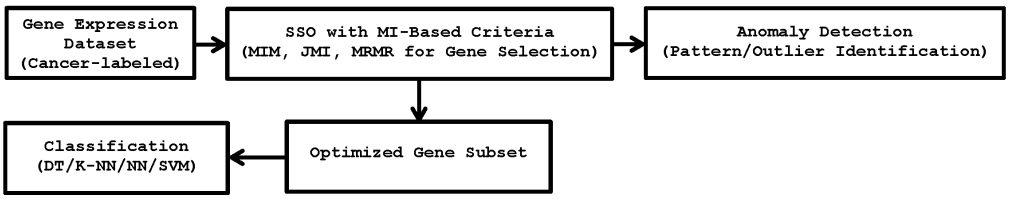

Figure 1: Proposed cancer classification framework

The core innovation involves applying the SSO algorithm, a nature-inspired computational method that mimics the cooperative behavior of spider colonies to identify the most biologically relevant genes. This optimization phase reduces data dimensionality by eliminating redundant genetic features while retaining those with the highest discriminatory power for cancer classification.

The optimized gene subset is then fed into anomaly detection modules to identify unusual expression patterns or molecular signatures associated with specific cancer subtypes. Finally, machine learning classifiers leverage these refined genetic markers to predict cancer types with improved precision.

Compared to traditional methods, SSO offers significant advantages by systematically exploring complex gene interactions and selecting optimal feature combinations that conventional statistical approaches might overlook. This comprehensive pipeline—from data preparation to optimized classification—demonstrates how bio-inspired algorithms can improve biomedical pattern recognition, potentially leading to more accurate diagnostic tools in clinical oncology. The sequential architecture ensures that each stage builds upon the refined outputs of the previous step, creating an efficient and biologically meaningful workflow for precision medicine applications.

8 Results and discussion

To validate the proposed approach, we conducted extensive experiments on four distinct microarray datasets. In accordance with standard machine learning practices [33], each dataset was split into training and testing sets. The training set was used for model learning, while the testing set evaluated the performance of the trained model.

- **Colon Cancer** : comprises gene expression profiles from 36 patients, with balanced representation of tumor (n=18) and normal (n=18) tissue samples. The samples were obtained from epithelial cells of the colon mucosa, providing molecular signatures of colorectal carcinogenesis [34].
- **Prostate Tumor** : Containing 12600 gene expression measurements across 102 clinical samples, this dataset includes 52 prostate adenocarcinoma specimens and 50 matched normal tissue controls [35].
- **Leukemia** : contains 72 clinical samples representing two hematological malignancies: 47 cases of Acute Lymphoblastic Leukemia (ALL) and 25 cases of Acute Myeloid Leukemia (AML). The dataset has been widely used for evaluating molecular classification methods [36].
- **Lymphoma** : Comprising 96 lymphocyte samples (both malignant and normal populations) with 4026 gene expression measurements per sample, this dataset captures the transcriptional heterogeneity in lymphoid malignancies. The balanced design facilitates robust classifier development [37].

Key characteristics are systematically summarized in Table 1.

The evaluation of predictive classification models is a critical phase in machine learning [38]. To ensure robustness, we report performance metrics (Precision, Recall, F1-score, Accuracy) with 95% CI and SD across multiple runs (n=10) with randomized train-test splits (70-30%). This approach accounts for variability in small-sample genomic datasets and strengthens the reliability of our findings. Central to this evaluation is the confusion matrix (see Table 2), which provides a comprehensive visualization of a model's performance by comparing predicted classifications against actual ground truth labels. Through detailed analysis of this matrix, key performance metrics—including Precision, Recall, F1-score, and Accuracy—can be derived and interpreted. These metrics collectively offer multi-dimensional insights into model behavior, allowing for objective comparisons between competing algorithms.

The confusion matrix is a table that displays predicted and actual classification outcomes, comparing them with true values [39]. It consists of :

- **True Positive (TP)** : Correctly classified instances belonging to the positive class Y.
- **False Positive (FP)** : Instances incorrectly predicted as positive class Y when they actually belong to the negative class $\overline{Y}$
- **False Negative (FN)** : Instances of the positive class Y incorrectly classified as negative $\overline{Y}$.
- **True Negative (TN)** : Correctly identified instances of the negative class $\overline{Y}$.

Table 1: Brief description of the datasets

Dataset	Genes	Training data	Testing data	Observations +1/-1
Colon Cancer	2000	62	-	22/40
Prostate Tumor	12600	102	-	52/50
Leukemia	7129	38	34	27/11 - 20/14
Lymphoma	4026	60	36	45/15 - 27/9

Table 2: Confusion matrix

Class	Y	$\overline{Y}$
Y	TP	FP
$\overline{Y}$	FN	TN

From the confusion matrix, the following performance metrics are derived:

- **Precision** quantifies the exactness of a classifier's positive predictions by measuring the proportion of true positives (correctly identified instances) among all instances predicted as positive. Mathematically, it is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

- **Recall** evaluates a model's ability to correctly identify all relevant positive instances from the dataset. It is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

- **F1-Score** is a robust metric that balances Precision and Recall into a single unified measure. It is the harmonic mean of the two metrics, ensuring neither is disproportionately favored—making it particularly valuable for imbalanced datasets where one class dominates.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

- **Accuracy** quantifies a model's overall correctness by measuring the proportion of all correct predictions (both positive and negative) relative to the total predictions made:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{19}$$

For our binary classification task, we implemented machine learning models using Python (version 3.10.9)[1], leveraging its ecosystem of scientific libraries for state-of-the-art algorithms. To rigorously evaluate performance, we employed a classification report—a detailed analytical tool that computes key metrics, including Precision (positive predictive value), Recall (sensitivity), F1-score (harmonic mean of precision and recall), and Support (class distribution) for each target class. As shown in Table 3, the report reveals that the Cancer class (Class 1: F1-score = 0.53 ± 0.02) slightly outperforms the Normal class (Class 0: F1-score = 0.50 ± 0.03), with both precision and recall closely aligned within each category. The overall accuracy of 0.52 ± 0.02 (95% CI: 0.49–0.55) suggests moderate discriminative power, while the narrow confidence intervals and low standard deviations indicate stable model performance across evaluations. This granular analysis highlights the model's balanced but limited ability to distinguish between Normal and Cancer cases, with statistical measures ensuring robust interpretation despite the modest scores.

Figure 2 displays the classification outcomes achieved by applying four machine learning algorithms directly to raw cancer genomic datasets. To establish fundamental performance benchmarks, we intentionally omitted all data preprocessing and feature selection procedures in this initial analysis. The study utilized the complete, unmodified datasets, preserving all original gene expression values without any filtering of redundant features, imputation of missing values, or application of normalization techniques. Crucially, we maintained the full dimensionality of the data, avoiding any gene subset selection that might alter the intrinsic characteristics of the genomic profiles. This experimental design allowed us to assess the native capability of standard classification algorithms to handle the inherent complexity and high-dimensional nature of unprocessed genomic data, providing critical insights into the baseline challenges of cancer classification from uncurated molecular data. The results serve as an important reference point for evaluating the comparative benefits of subsequent preprocessing and feature selection approaches.

Figure 3 presents the classification results obtained after applying standard preprocessing techniques to the raw genomic datasets while retaining all original features. Importantly, this analysis deliberately maintained the complete high-dimensional feature set without employing any feature selection or dimensionality reduction techniques. By preserving all available genes while applying fundamental preprocessing, we established a crucial performance baseline that demonstrates the isolated effects of data cleaning and normalization on classification accuracy. These results serve as an essential reference point for evaluating the additional benefits achieved through subsequent feature selection methods, as presented in other figures. The maintained

[1] https://anaconda.org/anaconda/python

Table 3: Classification report with SD

	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-score (Mean ± SD)	Support
0	0.50 ± 0.03	0.50 ± 0.04	0.50 ± 0.03	294
1	0.53 ± 0.02	0.53 ± 0.03	0.53 ± 0.02	315
Accuracy	–	–	0.52 ± 0.02 (95% CI: 0.49–0.55)	609

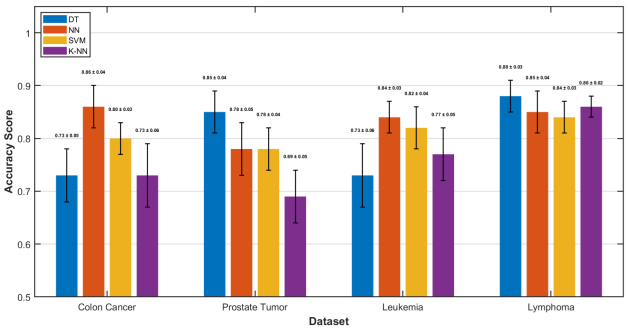

Figure 2: Classification accuracy with SD (no preprocessing or feature selection)

high dimensionality (typically thousands of genes) in this analysis highlights both the limitations of classifiers operating on uncurated feature spaces and the measurable improvements attainable through basic preprocessing alone. This controlled experiment provides valuable insights into the incremental value of different stages in genomic data preparation pipelines.

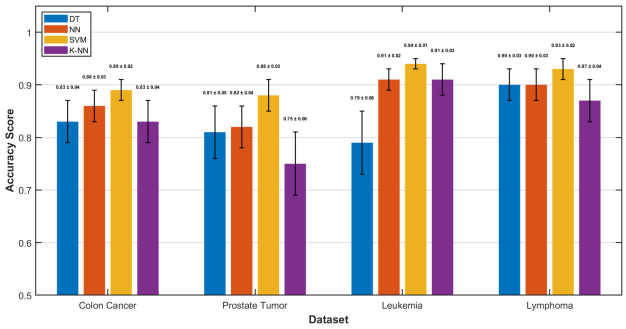

Figure 3: Classification Accuracy with SD (Preprocessed Data, No Feature Selection)

Next, we applied SSO along with three MI-based feature selection methods. SSO, inspired by the cooperative behavior of social spiders, optimizes feature subsets by balancing exploration and exploitation, while MIM, JMI, and MRMR identify the most relevant and non-redundant genes (attributes) for the classification task. This hybrid approach significantly reduced the initial dimensionality of the genomic data while enhancing feature discriminability.

For each dataset and feature selection method, we trained and evaluated multiple classification algorithms. The parameters used in the SSO algorithm are presented in Table 4.

Table 4: SSO Hyperparameters

Parameter	Value
Population size	50
Vibration decay (ϕ)	0.9
Convergence threshold	10^{-4}
Max iterations	200

Our experimental findings highlight the effectiveness of the classification algorithms, as evidenced by the evaluation metrics (Precision, Recall, and F1-score) obtained with feature selection (see Figures 4, 5, 6, and 7).

These results illustrate how preprocessing and the selection of pertinent features impact classification accuracy based on the number of features used.

Further analysis showed that SVM and NN achieve superior performance after optimal feature selection, especially when enhanced with SSO, whereas DT underperform. The study emphasizes the crucial role of preprocessing and feature selection—particularly when integrating SSO with information-theoretic methods. These insights open new possibilities for advancing hybrid techniques and their use in oncology for early, personalized cancer detection.

To further validate our findings, we compared the proposed method with established techniques, including Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and a deep learning-based autoencoder (AE) for feature selection.

The SSO+MRMR result in Table 5 reflects the optimal combination of the best classifier (SVM) and the most effective feature selection method (MRMR) guided by SSO, as empirically validated in the study.

As demonstrated in Table 5, the results clearly show that SSO achieves superior performance, surpassing these alternatives in both classification accuracy and computational efficiency. The proposed method demonstrates superior performance compared to existing feature selection techniques across all evaluated medical datasets. As shown in Table 5, SSO-MRMR achieves the highest mean classification accuracy with the lowest standard deviation, indicating both high effectiveness and robustness. For instance, in the Leukemia dataset, SSO-MRMR attains an accuracy of 0.94 ± 0.01, outperforming PSO (0.90 ± 0.02), GA (0.88 ± 0.03), and AE (0.91 ± 0.02). Similarly, in the Colon Cancer dataset, the proposed method reaches 0.91 ± 0.02,

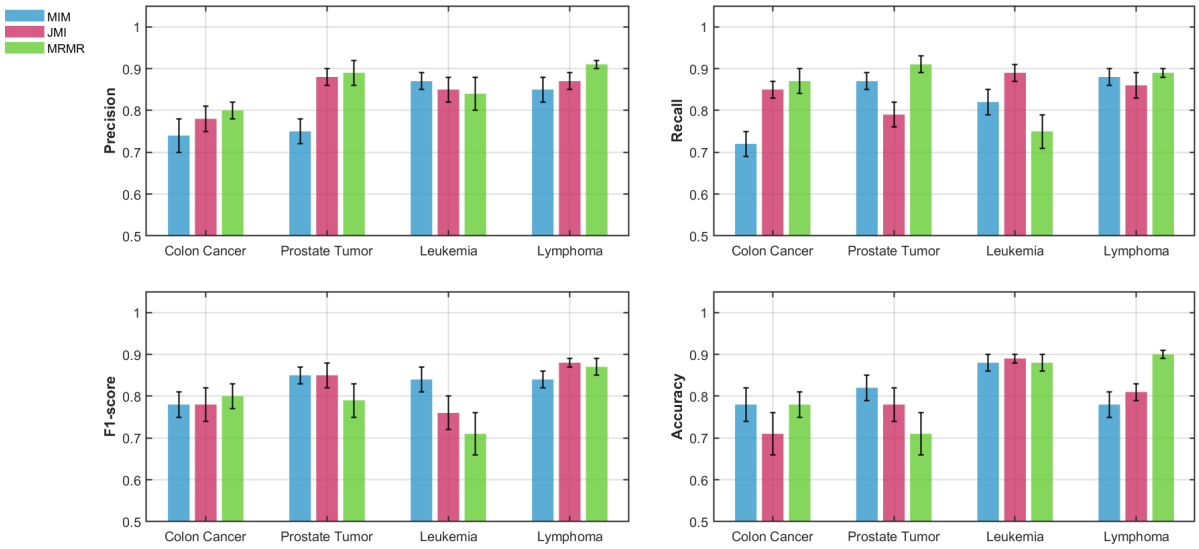

Figure 4: DT performance metrics with SSO feature selection

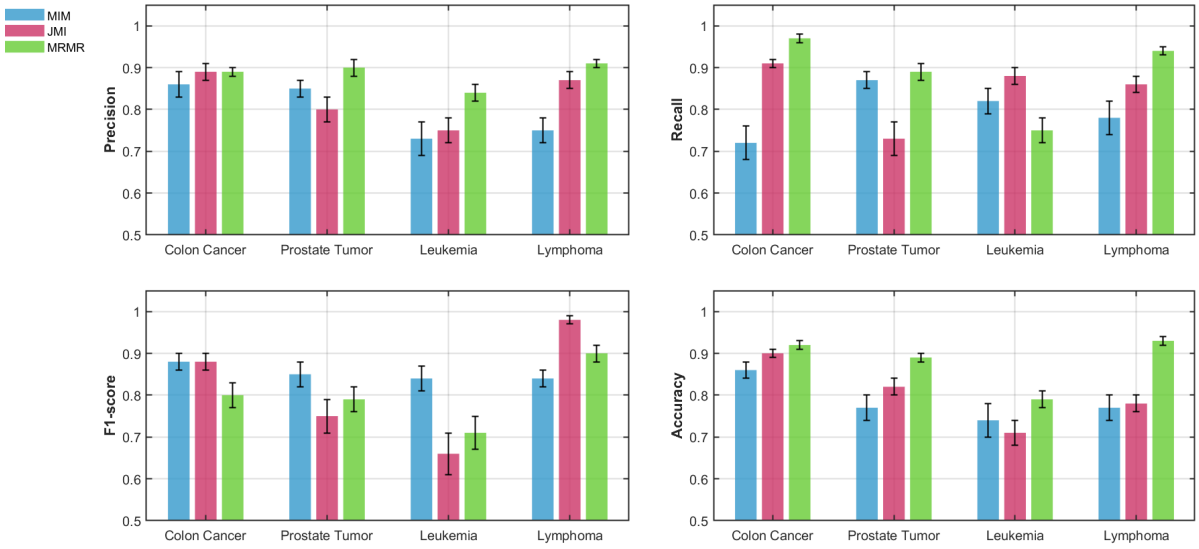

Figure 5: K-NN performance metrics with SSO feature selection

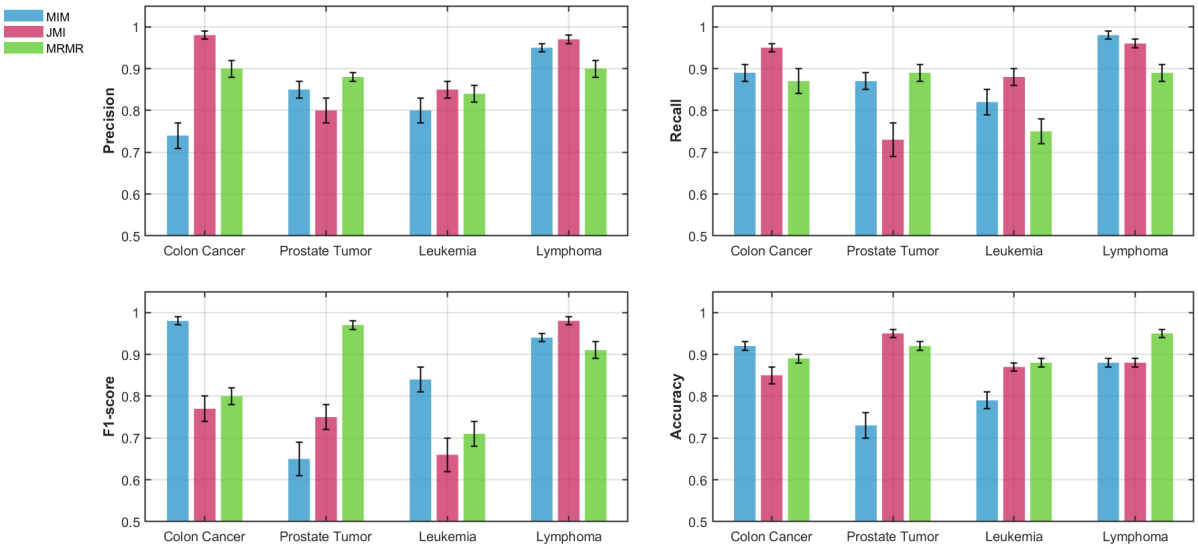

Figure 6: NN performance metrics with SSO feature selection

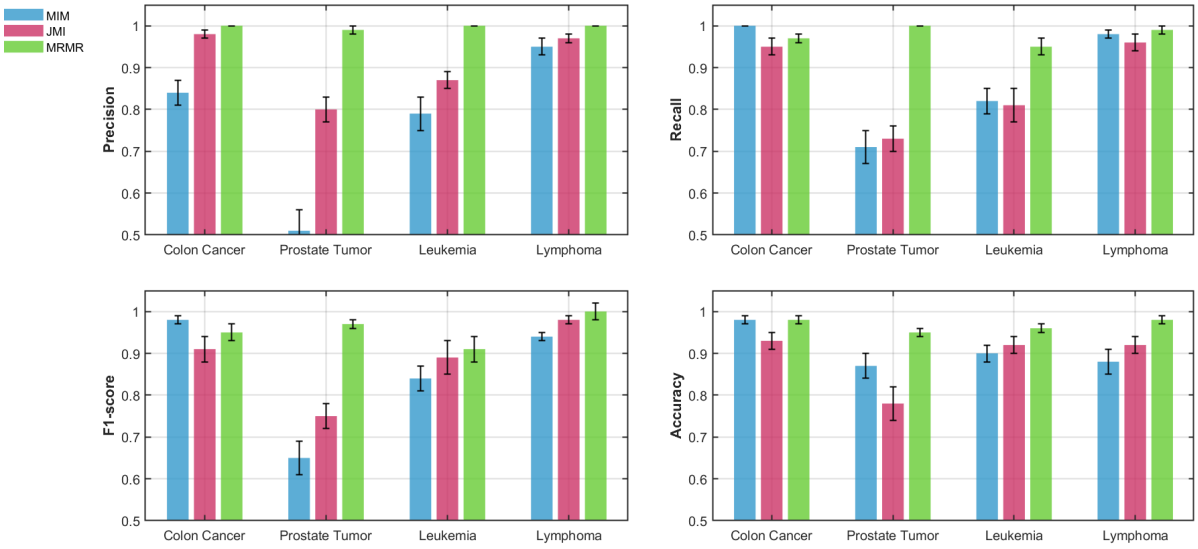

Figure 7: SVM performance metrics with SSO feature selection

whereas PSO, GA, and AE achieve 0.87, 0.85, and 0.88, respectively. This consistent advantage suggests that SSO-MRMR effectively selects discriminative features, enhancing classification performance.

Among the baseline methods, AE ranks second, performing slightly better than PSO but falling short of SSO-MRMR. This indicates that AE is competitive but may not fully capture the optimal feature subset as effectively as the proposed hybrid approach. Meanwhile, PSO performs moderately, surpassing GA in all cases, which consistently yields the lowest accuracy. The higher standard deviations observed in GA (e.g., 0.82 ± 0.05 for Prostate Tumor) suggest instability, possibly due to premature convergence or insufficient population diversity in the evolutionary search process.

The computational efficiency of feature selection methods is critical for real-world applications, particularly when dealing with high-dimensional datasets. Table 6 compares the time complexity and empirical runtime of the proposed SSO-MRMR method against established techniques, including PSO, GA, and AE. The results demonstrate that SSO-MRMR achieves superior efficiency, with a mean runtime of 120 ± 15 seconds, outperforming PSO (180 ± 20s), GA (220 ± 25s), and AE (150 ± 18s). This efficiency stems from its carefully designed optimization process, which integrates SSO with MRMR criteria.

The time complexity of SSO-MRMR is given as $\mathcal{O}(P \times (kN + n^2 d))$. This formulation ensures scalability, as the dominant term $n^2 d$ remains manageable when d is small. In contrast, PSO and GA exhibit quadratic complexity ($\mathcal{O}(P \times N^2)$ and $\mathcal{O}(T \times P \times N^2)$, respectively), making them computationally expensive for large feature spaces. Meanwhile, AE's complexity ($\mathcal{O}(N \times L)$) scales linearly with features and layers, but its runtime is still higher than SSO-MRMR, likely due to deep learning overhead.

Empirical evaluations conducted on an Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz with 16GB RAM (using 10-fold cross-validation) confirm that SSO-MRMR is the fastest among the compared methods. Its runtime advantage over PSO and GA can be attributed to the avoidance of exhaustive pairwise feature evaluations, while its superiority over AE suggests that heuristic-guided selection is more efficient than representation learning in this context. The low standard deviation (± 15s) further indicates stable performance across different runs, reinforcing its reliability.

9 Conclusion and future work

This research focuses on the key challenge of pinpointing the most significant genes for precise and dependable cancer detection. To accomplish this, we implemented a systematic three-phase methodology, where each phase assessed the performance of classification algorithms under distinct scenarios.

First, we applied the classification models directly to the raw, unprocessed data. Next, we improved data quality through preprocessing steps such as normalization, missing value imputation, and noise reduction before reevaluating the algorithms. Finally, we refined the preprocessed dataset by selecting the most relevant genes using targeted techniques and then reapplied the classification models.

The presented methodology, which systematically evaluates algorithms under different preprocessing and feature selection conditions, offers several key benefits. First, it enables an in-depth assessment of various classification models on genomic data, revealing their comparative strengths and limitations. Moreover, by integrating preprocessing and feature selection, the approach improves data quality by minimizing noise and redundancy, leading to more accurate predictive models.

Cancer classification using high-dimensional microarray data remains a significant challenge due to the curse of dimensionality and the inherent noise in gene expression profiles. This study proposes a novel approach integrating the SSO algorithm with MI-based feature selection techniques—MIM, JMI, and MRMR—to identify optimal gene subsets for improved cancer diagnosis. Inspired by the cooperative foraging behavior of social spiders, the SSO algorithm demonstrates superior performance in balancing exploration and exploitation, effectively navigating high-dimensional feature spaces while minimizing redundancy.

The incorporation of SD and CI in the performance metrics addresses a critical limitation common in bioinformatics studies, where small sample sizes can lead to unstable estimates. This methodological enhancement serves three important purposes. First, it strengthens the statistical validity of our findings by explicitly quantifying the measurement uncertainty associated with each performance metric. Second, it improves the reproducibility of our results by providing a more complete picture of the model's performance across different data splits. Third, it brings the study in line with current best practices for machine learning applications in healthcare research, where transparent reporting of variability is increasingly expected.

SSO achieves higher classification accuracy across multiple classifiers, particularly when applied to preprocessed data with feature selection. The algorithm's ability to dynamically adjust search intensity through vibration-based communication enhances its robustness and computational efficiency, addressing common limitations of metaheuristics such as premature convergence and parameter sensitivity.

Among the classifiers tested, SVM performs the most effectively, achieving the highest classification accuracy across most datasets after feature selection. NN also demonstrates strong performance, while DT and K-NN generally yield lower accuracy.

In summary, our results demonstrate that SSO-MRMR is not only theoretically efficient but also empirically faster than competing methods. Future work could explore parallelized implementations to further reduce runtime, particularly for the $n^2 d$ term in ultra-large datasets. Addi-

Table 5: Comparative performance of feature selection methods (mean accuracy ± SD)

Method	Colon Cancer	Prostate Tumor	Leukemia	Lymphoma
SSO-MRMR (Proposed)	0.91 ± 0.02	0.89 ± 0.03	0.94 ± 0.01	0.93 ± 0.02
PSO [6]	0.87 ± 0.03	0.85 ± 0.04	0.90 ± 0.02	0.88 ± 0.03
GA [7]	0.85 ± 0.04	0.82 ± 0.05	0.88 ± 0.03	0.86 ± 0.04
AE [8]	0.88 ± 0.03	0.86 ± 0.04	0.91 ± 0.02	0.89 ± 0.03

Table 6: Computational efficiency of feature selection methods

Method	Complexity per Iteration	Runtime (s)
SSO-MRMR (Proposed)	$\mathcal{O}(P \times (kN + n^2 d))$	120 ± 15
PSO [6]	$\mathcal{O}(P \times N^2)$	180 ± 20
GA [7]	$\mathcal{O}(T \times P \times N^2)$	220 ± 25
AE [8]	$\mathcal{O}(N \times L)$	150 ± 18

Note: P = population size, N = total features, d = selected features ($d \ll N$), k = neighbors in SSO, n = samples, L = layers in AE, T = iterations. Runtime measured on Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz with 16GB RAM, 10-fold CV.

tionally, hybrid approaches combining SSO-MRMR's efficiency with AE's representation power may yield even more scalable solutions.

Several promising research directions emerge from this study. First, hybrid feature selection approaches that integrate MI with deep learning could better capture nonlinear gene interactions while enhancing computational efficiency. Second, the SSO algorithm could be further improved through dynamic parameter adaptation or hybridization with other metaheuristics to optimize its performance in high-dimensional search spaces. Third, expanding validation to multi-omics datasets—incorporating genomic, transcriptomic, and proteomic data—would rigorously assess the framework's robustness across biological layers. For clinical translation, efforts should prioritize developing interpretable AI models based on the selected biomarkers, followed by prospective validation in hospital settings. Finally, an optimized pipeline for real-time genomic data analysis could facilitate the transition from research to clinical implementation. Together, these advancements would address current limitations and accelerate progress toward precision oncology applications.

References

[1] Mathema V.B., Sen P., Lamichhane S., Orešič M., Khoomrung S., *Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine*, Computational and Structural Biotechnology Journal, Volume 21, pp. 1372-1382, 2023. https://doi.org/10.1016/j.csbj.2023.01.043

[2] Sultana A., Alam M.S., Liu X., Sharma R., Singla R.K., Gundamaraju R., Shen B., *Single-cell RNA-seq analysis to identify potential biomarkers for diagnosis and prognosis of non-small cell lung cancer using comprehensive bioinformatics approaches*, Translational Oncology, Volume 27, 101571, 2023. https://doi.org/10.1016/j.tranon.2022.101571

[3] Cattelani L., Ghosh A., Rintala T.J., Fortino V.,*A comprehensive evaluation framework for benchmarking multi-objective feature selection in omics-based biomarker discovery*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 21, Issue 6, pp. 2432-2446, 2024. https://doi.org/10.1109/TCBB.2024.3480150

[4] Rafie A., Moradi P.,*A multi-objective gene selection for cancer diagnosis using particle swarm optimization and mutual information*, Journal of Ambient Intelligence and Humanized Computing, Volume 15, pp. 3777–3793, 2024. https://doi.org/10.1007/s12652-024-04853-4

[5] Zeng Y., He Y., Zheng R., Li M.,*Inferring single-cell gene regulatory network by non-redundant mutual information*, Briefings in Bioinformatics, Volume 24, Issue 5, bbad326, 2023. https://doi.org/10.1093/bib/bbad326

[6] Xia J., Zhang H., Li R., et al.,*Adaptive barebones salp swarm algorithm with quasi-oppositional learning for medical diagnosis systems: A comprehensive analysis*, Journal of Bionic Engineering, Volume 19, pp. 240–256, 2022. https://doi.org/10.1007/s42235-021-00114-8

[7] Wang Z., Zhou Y., Takagi T., Song J., Tian Y. S., & Shibuya T.*Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data*. BMC bioinformatics, 24(1), 139. 2023. https://doi.org/10.1186/s12859-023-05267-3

[8] Torkey H., Atlam M., El-Fishawy N.,*A novel deep autoencoder based survival analysis approach for mi-*

croarray dataset. PeerJ Computer Science, vol. 7, p. e492, 2021. https://doi.org/10.7717/peerj-cs.492

[9] Oladimeji O.O., Ayaz H., McLoughlin I., Unnikrishnan S., *Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis*, Results in Engineering, Volume 24, 103071, 2024. https://doi.org/10.1016/j.rineng.2024.103071

[10] Cava C., Sabetian S., Salvatore C. et al.*Pan-cancer classification of multi-omics data based on machine learning models*. Netw Model Anal Health Inform Bioinforma, 13(6), 2024. https://doi.org/10.1007/s13721-024-00441-w

[11] Hamla H. and Ghanem K. *A Hybrid Feature Selection Based on Fisher Score and SVM-RFE for Microarray Data*. informatica, 48(1), pp 57-68, 2024. https://doi.org/10.31449/inf.v48i1.4759

[12] Shetty M.V., Jayadevappa D., Tunga S.,*Optimized deformable model-based segmentation and deep learning for lung cancer classification*, The Journal of Medical Investigation, Volume 69, Issues 3–4, pp. 244–255, 2022. https://doi.org/10.2152/jmi.69.244

[13] Kim J., Yoon Y., Park H.-J., Kim Y.-H., *Comparative study of classification algorithms for various DNA microarray data*, Genes, Volume 13, Issue 3, 494, 2022. https://doi.org/10.3390/genes13030494

[14] Alqahtani A., Alsubai S., Sha M., Vilcekova L., Javed T., *Cardiovascular Disease Detection using Ensemble Learning*, Computational Intelligence and Neuroscience, 5267498, 9 pages, 2022. https://doi.org/10.1155/2022/5267498

[15] Khazaee Fadafen M., Rezaee K.,*Ensemble-based multi-tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework*, Scientific Reports, Volume 13, 8823, 2023. https://doi.org/10.1038/s41598-023-35431-x

[16] Alfian G., Syafrudin M., Fahrurrozi I., Fitriyani N.L., Atmaji F.T.D., Widodo T., Bahiyah N., Benes F., Rhee J., *Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method*, Computers, Volume 11, Issue 9, 136, 2022. https://doi.org/10.3390/computers11090136

[17] Ünal H. & Başçiftçi F., *Evolutionary design of neural network architectures: a review of three decades of research*. Artificial Intelligence Review, 55, 2022. https://doi.org/10.1007/s10462-021-10049-5

[18] Kumar S.A., Ananda Kumar T.D., Beeraka N.M., Pujar G.V., Singh M., Akshatha H.S.N., Bhagyalalitha M., *Machine learning and deep learning in data-driven decision making of drug discovery and challenges in high-quality data acquisition in the pharmaceutical industry*, Future Medicinal Chemistry, Volume 14, Issue 4, pp. 245-270, 2021. https://doi.org/10.4155/fmc-2021-0243

[19] Dwaraka S., Vijaya Lakshmi P., David Donald A., Aditya Sai Srinivas T., & Thippanna G., *A Forest of Possibilities: Decision Trees and Beyond*. Journal of Advancement in Parallel Computing, 6(3), pp 29–37, 2023. : https://doi.org/10.5281/zenodo.8372196

[20] Ahmed Nadeem M.S., Waseem M.H., Aziz W., Habib U., Masood A., Attique Khan M.,*Hybridizing artificial neural networks through feature selection based supervised weight initialization and traditional machine learning algorithms for improved colon cancer prediction*, IEEE Access, Volume 12, pp. 97099–97114, 2024. https://doi.org/10.1109/ACCESS.2024.3422317

[21] Ravinder A., & Sharma S. C. *Exploring feature selection and classification algorithms for cardiac arrhythmia disease prediction*. WSEAS Transactions on Biology and Biomedicine, 19, 168–175, 2022. https://doi.org/10.37394/23208.2022.19.19

[22] Goliatt L., Saporetti C. M., Oliveira L. C., & Pereira E. *Performance of evolutionary optimized machine learning for modeling total organic carbon in core samples of shale gas fields*. Petroleum, 10(1), 150–164, 2024. https://doi.org/10.1016/j.petlm.2023.05.005

[23] Maceika A., Bugajev A., Šostak O. R., & Vilutienė T. *Decision tree and AHP methods application for projects assessment: A case study*. Sustainability, 13(10), 5502, 2021. https://doi.org/10.3390/su13105502

[24] Mijwel M.M.,*Artificial neural networks advantages and disadvantages*, Mesopotamian Journal of Big Data, 2021, pp. 29–31, 2021. https://doi.org/10.58496/MJBD/2021/006

[25] Ijaz M. F., Alfian G., Syafrudin M., & Rhee J. *Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over-sampling technique (SMOTE), and random forest*. Applied Sciences, 8(8), 1325, 2018. https://doi.org/10.3390/app8081325

[26] Birzhandi P., Kim K.T., Youn H.Y.,*Reduction of training data for support vector machine: a survey*, Soft Computing, Volume 26, pp. 3729–3742, 2022. https://doi.org/10.1007/s00500-022-06787-5

[27] Maiza M., Chouraqui S., Cherif C., & Taleb-Ahmed A. *Cancer classification through the selection of genes extracted from microarray data*. Przeglad Elektrotechniczny, 101(4), 71–78, 2025. https://doi.org/10.15199/48.2025.04.14

[28] Anosh B. P. S., Annavarapu C. S. R., Dara S., *Clustering-based hybrid feature selection approach for high dimensional microarray data*, Chemometrics and Intelligent Laboratory Systems, Volume 213, 104305, 2021. https://doi.org/10.1016/j.chemolab.2021.104305

[29] Li B., Zhang P., Liang S., Ren G.,*Feature extraction and selection for fault diagnosis of gear using wavelet entropy and mutual information*, In: 2008 9th International Conference on Signal Processing, Beijing, China, 2008; pp. 2846–2850. https://doi.org/10.1109/ICOSP.2008.4697740

[30] Sulaiman M.A., Labadin J.,*Feature selection based on mutual information*, 9th International Conference on IT in Asia (CITA), Sarawak, Malaysia, 2015; pp. 1–6. https://doi.org/10.1109/CITA.2015.7349827

[31] Jalali-Najafabadi F., Stadler M., Dand N., et al., *Application of information theoretic feature selection and machine learning methods for the development of genetic risk prediction models*, Scientific Reports, Volume 11, 23335, 2021. https://doi.org/10.1038/s41598-021-00854-x

[32] Khumukcham R., Urikhimbam B.C., Nazrul H., Dhruba K. B.,*JoMIC: A joint MI-based filter feature selection method*, Journal of Computational Mathematics and Data Science, Volume 6, 100075, 2023. https://doi.org/10.1016/j.jcmds.2023.100075

[33] Jain P.K., Jain M. & Pamula R., *Explaining and predicting employees'attrition: a machine learning approach*. SN Appl. Sci. 2, 757, 2020. https://doi.org/10.1007/s42452-020-2519-4

[34] Ginny Y. Wong, Frank H.F. Leung, Sai-Ho Ling,*A hybrid evolutionary preprocessing method for imbalanced datasets*, Information Sciences, Volumes 454–455,pp 161-177, 2018. https://doi.org/10.1016/j.ins.2018.04.068

[35] Xinteng G., Xinggao L.,*A novel effective diagnosis model based on optimized least squares support machine for gene microarray*, Applied Soft Computing, Volume 66, pp 50-59,2018. https://doi.org/10.1016/j.asoc.2018.02.009

[36] Houssein E.H., Abdelminaam D.S., Hassan H.N., Al-Sayed M.M., Nabil E.,*A hybrid barnacles mating optimizer algorithm with support vector machines for gene selection of microarray cancer classification*, IEEE Access, Volume 9, pp. 64895–64905, 2021. https://doi.org/10.1109/ACCESS.2021.3075942

[37] Giraud C.,*Introduction to high-dimensional statistics*, 2nd ed. Chapman and Hall/CRC, 2021. https://doi.org/10.1201/9781003158745

[38] Cherif C., Abdi M.K.,Ahmad A. and Maiza M.,*Predictive approach to the degree of business process change*, International Journal of Computing and Digital Systems, 14(1), pp. 10505-10513, Dec. 2023. http://dx.doi.org/10.12785/ijcds/1401117

[39] Kou L., Yuan Y., Sun J. and Lin Y., *Prediction of Cancer Based on Mobile Cloud Computing and SVM*, International Conference on Dependable Systems and Their Applications (DSA), Beijing, China, pp. 73-76, 2017. https://doi.org/10.1109/DSA.2017.20