ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 12 N. 1 (2023), e31187
eISSN: 2255-2863
DOI: https://doi.org/10.14201/adcaij.31187

Stable Feature Selection using Improved Whale Optimization Algorithm for Microarray Datasets

Dipti Theng[a], Kishor K Bhoyar[b]

[a]Computer Technology Department, YCCE, Nagpur, Maharashtra
[b]Computer Science and Engineering Department, YCCE, Nagpur, Maharashtra
deepti.theng@gmail.com, kbhoyar@gmail.com

KEYWORDS

feature selection; stability of feature selection; whale optimization algorithm; marine predator algorithm; grey wolf optimization; microarray datasets; high dimensional datasets

ABSTRACT

A microarray is a collection of DNA sequences that reflect an organism's whole gene set and are organized in a grid pattern for use in genetic testing. Microarray datasets are extremely high-dimensional and have a very small sample size, posing the challenges of insufficient data and high computational complexity. Identification of true biomarkers that are the most significant features (a very small subset of the complete feature set) is desired to solve these issues. This reduces over-fitting, and time complexity, and improves model generalization. Various feature selection algorithms are used for this biomarker identification. This research proposed a modification to the whale optimization algorithm (WOAm) for biomarker discovery, in which the fitness of each search agent is evaluated using the hinge loss function during the hunting for prey phase to determine the optimal search agent. Also compared the results of the proposed modified algorithm with the original whale optimization algorithm and also with contemporary algorithms like the marine predator algorithm and grey wolf optimization. All these algorithms are evaluated on six different high-dimensional microarray datasets. It has been observed that the proposed modification for the whale optimization algorithm has significantly improved the results of feature selection across all the datasets. Domain experts trust the resultant biomarker/ associated genes by the stability of the results obtained. The chosen feature set's stability was also evaluated during the research work. According to the findings, our proposed WOAm has superior stability compared to other algorithms for the CNS, colon, Leukemia, and OSCC. datasets.

1. Introduction

Bioinformatics data analysis has piqued the interest of both researchers and industry with its ability to simultaneously quantify the activities and interactions of thousands of genes. An important problem in bioinformatics research is the selection of true biomarkers (a most relevant feature subset) for disease identification/classification using gene expression data (Theng and Bhoyar, 2022). Such, mRNA microarray datasets are very difficult to analyze due to thousands of features (genes) and fewer samples i.e. high dimensional and small sample size datasets. Feature reduction by selecting significant genes can make this analysis easier and more fruitful. Reduced over-fitting, enhanced time complexity, model generalization, comprehensibility, and efficiency are all advantages of selecting a subset of the most significant features for creating learning models. Feature selection is a multi-objective optimization task that seeks the (near) optimized set of features based on a set of criteria (target). Genomic microarray datasets are very high in dimension ranging to tens of thousands of features, most of which are irrelevant to the task at hand. Such huge feature space increases the computational time for feature selection beyond reasonable limits. This implies that standard brute force methods are ineffective and new effective search techniques must be adopted instead. A metaheuristic is a problem-independent mathematical model which would provide a better approach with an optimization process, specifically when there is little information and computing capability. Nature-inspired metaheuristic algorithms are especially prevalent for optimization problems that require a lot of processing power, have a lot of complexity, and suffer from premature convergence (Abu et al., 2022; Agrawal et al., 2021; Mahendru and Agarwal, 2019). A new initiation approach, new fitness function, novel operator, modified population structure, updated mechanism, new encoding schemes, multi-objectives, state flipping, and parallelism are some of the techniques used to improve the efficiency of nature-inspired algorithms in optimization (Simumba et al., 2022).

This research proposed a modification to the whale optimization algorithm (WOA_m) for biomarker discovery, in which the hinge loss function is used to assess the fitness of each search agent during the hunting for prey phase to determine the optimal search agent. One of the advantages of using the WOA_m algorithm for feature selection in microarray datasets is its ability to handle high-dimensional data and a large number of features. The WOA_m algorithm is robust to the curse of dimensionality and can effectively search for informative features among a large number of features in microarray datasets. Additionally, the WOA_m algorithm is not sensitive to the initial solution and can converge to the optimal solution in a reasonable number of iterations. Also evaluated and tested the proposed algorithm's results against the original whale optimization algorithm and also with contemporary algorithms like whale optimization, improved whale optimization, grey wolf optimization, and marine predator optimization. All these algorithms are evaluated on six different high-dimensional microarray datasets. From the results obtained, it is observed that the proposed modification for the whale optimization algorithm has significantly improved the results of feature selection across all the datasets. The performance of the feature subset selected by all algorithms is tested for stability. The stability issue in feature selection has received much attention recently.

Because there are so many attributes and very few samples, whenever a new sample is introduced to the training data, the selection frequently yields a distinct subset producing an unstable feature set. As the instability of feature selection is closely linked to data variance, researchers believe that limiting data variation improves selection stability. Stability refers to the algorithm's output for feature set selection being unaffected by minor changes to the training set. Domain experts trust the resultant biomarker/ associated genes by the stability of the results obtained. The most significant parameters to

analyze and compare the performances of many selection strategies are stable feature relevance (precision), accuracy, bias, and complexity to evaluate the relationship between stability scores (Shukla et al., 2020). The stable feature subset contains the true relevant biomarkers that researchers rely on for bioinformatics studies. To ensure accurate biomarker selection, the stability is also evaluated for the feature set that was chosen using the modified WOA_m.

Table 1 below elaborates the list of symbols used in the manuscript.

Table 1. Definition of symbols used

Symbol	Description
$X(t)$	The position vector of a particular solution (search agent) at the current iteration.
$X^*(t)$	The position vector of the best solution.
A	Coefficient vector where t is the current iteration.
C	Coefficient vector where t is the current iteration.
D	Distance between two position vectors, $C \cdot X^*(t)$ and $X(t)$.
μ	The constant and has a value of1.07
$\vec{D}$	The distance vector between two position vectors
$\overrightarrow{X_{rand}}$	A randomly selected position vector
$\vec{A}$	A coefficient vector used to control the update rate of the distance vector.
L	The hinge loss, which quantifies the misclassification error or the deviation of the predicted output from the true label
y	The true label of the binary classification problem, which can take the values +1 or -1. In the WOA context, it represents the actual class label of the training instance.
$f(x)$	The predicted score or output of the classification model for the input instance x. In the WOA context, it represents the output of a search agent (whale) in finding a good solution.
Max_{itern}	The maximum number of iterations or epochs for which the algorithm will run.
$R_i (i = 1,2, ... 50)$	Initialization of random search agents. Each R_i represents a search agent.
R^*	The best search agent among all the search agents based on their fitness evaluation.
t	The current iteration or epoch number
R	The position matrix of the search agents. It contains the positions of all the search agents.

2. Literature review

This section summarizes the most popular Metaheuristics algorithms for feature selection and discusses the research gap. This review assisted in determining the scope of modification and improvement required for efficient and stable feature selection. Some of the most popular metaheuristics algorithms for feature selection are:

- Genetic Algorithm (GA) (Tan et al., 2006; Xie et al., 2022): GA is a biologically inspired optimization algorithm that mimics the process of natural selection. It has been successfully demonstrated to identify informative features and has been widely utilized for feature selection in a variety of fields.

- Particle Swarm Optimization (PSO) (Wang and Jia, 2022; Alrefai and Ibrahim, 2022): PSO is a population-based optimization algorithm inspired by the movement of birds and fish. It has been used to solve feature selection issues, and it is known to be successful at locating informative features.
- Ant Colony Optimization (ACO) (Fahrudin et al., 2016; Shojaee et al., 2022): ACO is an optimization algorithm that mimics the behavior of ants searching for food. It is widely used to solve feature selection issues and has proven to be successful at locating informative features.
- Simulated Annealing (SA) (Pashaei and Pashaei, 2022; Perez and Marwala, 2012): SA is a probabilistic optimization algorithm that mimics the process of annealing in metallurgy. It is being tested on feature selection problems and proved to be efficient at identifying relevant features.
- Artificial Bee Colony (ABC) (Zhong et al, 2023; Silva and Gertrudes, 2022): ABC is a metaheuristic algorithm inspired by the intelligent behavior of honeybees in foraging. It has been widely used for feature selection in various domains and has been shown to be effective in identifying informative features.
- Whale Optimization Algorithm (WOA) (Septian and Utami, 2022; Tawhid and Ibrahim, 2020): WOA is a metaheuristic algorithm inspired by the hunting strategy of humpback whalesIt is being applied to solve feature selection challenges and has been successful in identifying important features.
- Cuckoo search (CS) (Alzaqebah et al., 2021; Aziz, 2022): CS is a metaheuristic algorithm inspired by the brood parasitism of some cuckoo species. It is showed effectively to identify informative features and is frequently used for feature selection in a variety of fields.

These metaheuristics algorithms have been applied to various feature selection problems and are effective in identifying informative features. However, it is important to note that the performance of these algorithms can vary depending on the specific dataset. Of all these metaheuristic algorithms, the WOA algorithm is best suitable for feature selection in microarray datasets because of its ability to handle high-dimensional data and the large number of features.

In Whale Optimization Algorithm (WOA) (Mirjalili and Lewis, 2016), involves three operations that represent the humpback whale's bubble-net hunting technique, prey tracking, and prey encirclement. It possesses the capacity to obtain a globally optimal solution while avoiding local optima. WOA has a limitation of slow convergence and easy localization. Marine Predator Optimization (MPO) (Faramarzi et al, 2022), obeys the natural rules guiding the optimal foraging method and the prey-to-predator ratio in maritime ecosystems. High-performance optimizer MPO outperforms CMA-ES, SSA, CS, GSA, PSO, and GA by a considerable margin. However, it suffers from time consumption to solve optimization. Grey Wolf Optimization (GWO) (Mirjalili et al., 2014), mimics the natural leadership structure and foraging strategy of grey wolves. In resolving numerous issues from the real world, GWO exhibits great performance and competitive outcomes. Low solution precision, poor local search performance, and sluggish convergence rate are limitations of GWO. An improved Whale Optimization Algorithm (iWOA) was proposed, which implements the exploitation phase using Singer's equation (Khaire and Dhanalakshmi, 2022). It has shown good stability for the selected feature subset. However, it is still convergent slowly.

Table 2. Review Summary

Sr. No.	Algorithm	Technique	Advantages	Limitations
1	Whale Optimization Algorithm (WOA)	Involves three operations that represent the humpback whale's bubble-net hunting technique, prey tracking, and prey encirclement.	Possesses the capacity to obtain a globally optimal solution while avoiding local optima.	Slow convergence and easy localization.
2	Marine Predator Optimization (MPO)	Obeys the natural rules guiding the optimal foraging method and the prey-to-predator ratio in maritime ecosystems.	High-performance optimizer MPO outperforms CMA-ES, SSA, CS, GSA, PSO, and GA by a considerable margin.	Time consumptions to solve optimization.
3	Grey Wolf Optimization (GWO)	Mimics the natural leadership structure and foraging strategy of grey wolves.	In resolving numerous issues from the real world, GWO exhibits great performance and competitive outcomes.	Low solution precision, poor local search performance, and sluggish convergence rate.
4	Improved Whale Optimization Algorithm (iWOA)	Exploitation phase using Singer's equation	Achieves good stability of feature subset	Slow convergence

Mirjalili et al. ((Mirjalili and Lewis, 2016) have developed a model meta-heuristic method called the Whale Optimization algorithm. WOA is a swarm-based approach that mimics the bubble-net attack strategy used by humpback whales to catch their prey. The largest mammals in the world are believed to be whales. Spindle cells in their brains are what give them intelligence. The whales can create their vernacular since they can live in groups. The humpback whale is one of seven different species of whales. A unique hunting technique used by humpback whales is known as the bubble-net feeding approach. By forming specific bubbles along a spiraling or nine-shaped pattern, this forging action is accomplished.

The WOA consists of the following components:

- A population of "whales" that represent potential solutions to a problem. Each whale has a unique position in the search space, also known as its "song".
- An objective function that is used to evaluate the fitness of each whale, based on its position in the search space.
- An update mechanism that modifies the position of each whale based on its position, the best position of its neighbors, and the global best position found so far.
- A stopping criterion that determines when the algorithm terminates, such as reaching a maximum number of iterations or achieving a satisfactory level of optimization.
- A random walk mechanism that is used to update the position of each whale based on the best positions found by its neighbors and the global best position. The probability of moving in a certain direction is related to the fitness of each position.
- A control mechanism that manages the overall execution of the algorithm, including initializing the population, evaluating the fitness of each whale, updating the positions, and checking for the stopping criterion.

The algorithm is designed to mimic the behavior of humpback whales in finding their prey by using a singing and searching strategy and that's why it is called Whale Optimization Algorithm.

The mathematical model for the WOA algorithm consists of an update rule that determines how the position of each whale (i.e., the solution candidate) is updated at each iteration. The update rule is inspired by the hunting behavior of whales. The humpback whale locates its prey and circles it. They believe that the best candidate option currently available is close to the ideal solution. The other agents attempt to update their positions in favor of the best search agent after allocating the best candidate solution, as shown in the equation below.

$$D = \left| C \cdot X^{*}(t) - X(t) \right| \tag{1}$$

$$X\,(t+1) = X^{*}(t) - A \cdot D \tag{2}$$

In iWOA, the authors have proposed the use of the singer function to update the position of search agents. The rest of the steps are the same as the original Whale Optimization Algorithm (WOA). The proposed equation for search agent position updating in the Encircling Prey phase. The other search agents will attempt to adjust their locations toward the best search agent after the greatest search agent has been identified. Singer function is used to express this behavior as

$$X\,(t+1) = \mu * (7.86 * X^{*}(t) - 23.31 * X(t)^{2} + 28.75 * X(t)^{3} - 13.308 * X(t)^{4}) \tag{3}$$

Singer's function adds unpredictability to the search agents' initial positions, enabling them to look for the target prey across a wider search field.

The literature review examined popular metaheuristic algorithms for feature selection, including Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Simulated Annealing (SA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), and Cuckoo Search (CS). It was found that these algorithms have been successfully applied to various feature selection problems and have demonstrated the ability to identify informative features. For example, GA has been used in cancer microarray data analysis (Tan et al., 2006), while PSO has shown effectiveness in gene expression data analysis (Wang and Jia, 2022; Alrefai and Ibrahim, 2022). ACO has been applied to feature selection in medical diagnosis (Fahrudin et al., 2016), and SA has been tested on bioinformatics datasets (Pashaei and Pashaei, 2022; Perez and Marwala, 2012). ABC has been widely used in different domains for feature selection (Zhong et al., 2023; Silva and Gertrudes, 2022). WOA has shown promise in handling high-dimensional microarray datasets (Septian and Utami, 2022; Tawhid and Ibrahim, 2020), while CS has been applied to various fields for feature selection (Alzaqebah et al., 2021; Aziz, 2022). The review also discussed the limitations of these algorithms, such as slow convergence and easy localization in WOA, time consumption in MPO, and limitations in solution precision and local search performance in GWO. Additionally, the improved Whale Optimization Algorithm (iWOA) was introduced, which showed good stability for the selected feature subset (Khaire and Dhanalakshmi, 2022). Overall, the literature review provides a comprehensive overview of these algorithms, their applications in feature selection, and their respective advantages and limitations.

3. Proposed Algorithm Modification

Modification in the existing whale optimization algorithm (WOA) is proposed in the prey search phase. The mathematical model for prey search is given below in equation 4.

$$\vec{D} = \overrightarrow{X_{rand}} - \vec{A} \cdot \vec{D} \tag{4}$$

Proposed improvement for the whale optimization algorithm

It is proposed to make modifications to the existing whale optimization algorithm (WOA) during the phase of looking for prey (prey searching). Humpback whales follow random searching (hunting for prey) concerning the position of one another. To drive the search agent to go further from a reference whale, 'A' with random values larger than 1 or less than 1 must be used. In the proposed approach, the best search agent is determined by comparing the fitness of each search agent using the hinge loss function. The hinge loss function is a type of loss function that is used in machine learning to calculate the error of a model. It is often used in classification tasks, and it is defined as the maximum of 0 and the difference between the true label and the predicted label.

In the context of the whale optimization algorithm (WOA), the hinge loss function could be used to find the best search agent (i.e., the whale that is most effective at finding a good solution) by comparing the performance of the different search agents on a classification task. The search agent with the lowest hinge loss would be considered the best. To use the hinge loss function for this purpose, defined a classification task and a dataset to use for training and evaluation. After training and evaluating each search agent using the hinge loss function, and compared the results to determine which search agent performed the best. Additionally, unpredictability to the search agents' initial position is added using the singer function.

The equation that describes hinge loss, a binary classification loss function, is:

$$L = max\,(0,1 - y * f(x)) \tag{5}$$

Hinge Loss is selected because of its unique property as it not only penalizes the wrong predictions but also the right predictions that are not confident. This property of the hinge loss makes it more suitable for biomarker identification where confidence is important for the selected biomarkers. It helps identify the search agent that achieves the least misclassification error, leading to improved accuracy in the classification problem. This enhances the exploitation ability of WOA resulting in fast convergence. The hinge loss function has several other properties that make it a suitable choice for the search agent in WOA:

It is convex: The hinge loss function is a convex function, which means that it has a single global minimum and that any local minimum is also a global minimum. This property makes it easier to optimize, and it ensures that the optimization algorithm will converge to the global minimum.

It is differentiable: The hinge loss function is differentiable, which means that its gradient can be calculated. This property is important for optimization algorithms that use gradient-based methods, such as WOA, as it allows the algorithm to update the solution based on the gradient information.

It is sensitive to misclassification: The hinge loss function penalizes misclassification more than classification errors, which makes it well-suited for binary classification problems, where the goal is to accurately separate the samples into two classes.

It is a good balance between simplicity and effectiveness: The hinge loss function is relatively simple, but it is still effective at capturing the underlying relationship between the input and output in binary classification problems.

Overall, the hinge loss function is a suitable choice for the search agent in WOA because it is a well-suited loss function for binary classification problems, which is convex, differentiable, sensitive to misclassification, and has a good balance between simplicity and effectiveness.

Mathematical Model for modified WOA (WOAm):

Algorithm_WOA$_m$:
Epochs initialization as Max_itern and random search agents initialization to R_i (i = 1, 2... 50)
To determine which search agent is the best, evaluate each agent's fitness (R*)
while (t < Max_itern)
    Updating position matrix (R)
    for every search agent
        Update E, r, C, c, and w
            if$_1$(r < 0.5)
                if$_2$ (|C| < +1)
                    Updating the present leading search agent
                else if$_2$ (|C| > +1)
                    By equation (5), choose a random search agent (R_{rand})
                    By equation (4), the best search agent to be updated
                end if$_2$ else if$_1$ (r ≥ 0.5)
                Updating the present leading search agent
            end if$_1$
    end for
  Evaluate each search agent's fitness value
  Update R* t = t+1
 end while
return R* and fitness score of R*

4. Implementation and Results Discussion

Metaheuristics algorithms like Whale Optimization, improved Whale Optimization, Grey Wolf Optimization, Marine Predator, and proposed modified Whale Optimization algorithms are implemented on six different microarray datasets. Proposed modifications and discussed their results in this section. Implementation is done using the Python 3.7.12 version in Google Colab. The datasets used for the experimentation are described in Table 3. All datasets are microarray sequences from the bioinformatics domain. These six datasets are six types of cancer each having a class label as 1 or 0 indicating cancerous and non-cancerous samples respectively. Except for OSCC, which was obtained from the NCBI public repository, other datasets were retrieved from

http://csse.szu.edu.cn/staff/zhuzx/Datasets.html (National Center for Biotechnology Information). Every dataset requires dimensionality reduction or feature selection since they have a large set of features (high dimension) and few samples.

Table 3. Experimental dataset description

Dataset	#Features	#classes	#samples
Breast	24482	2	97
CNS	7130	2	60
Colon	2001	2	62
Leukemia	7130	2	72
Ovarian	15155	2	253
OSCC	41004	2	50

To deal with the imbalanced dataset, the Smote-Tomek algorithm is applied. The SMOTE-Tomek algorithm efficiently manages the imbalanced dataset by combining oversampling and undersampling. Further, the dataset is normalized to bring all the features to the same level. Selected feature subsets are tested by applying a Support Vector Machine (SVM) classification algorithm, for evaluating how effectively selected features contribute to the objective function discrimination. Implementation of the proposed WOA_m for feature selection in microarray datasets typically involves the following steps:

- Preprocessing: This step involves cleaning and normalizing the microarray data, as well as splitting the dataset into a training set and a test set.
- Optimization: This step involves defining a fitness function that evaluates the performance of different feature subsets based on a classification or regression performance metric, such as accuracy or area under the curve (AUC). The iWOA algorithm is then used to optimize this fitness function by iteratively exploring different feature subsets and updating the solution based on the fitness value until it converges to a near-optimal subset of features.
- Evaluation: This step involves evaluating the performance of the classifier trained on the selected feature subset using the test set. This can be done using a variety of performance metrics such as accuracy, precision, recall, and F1-score.

Using the WOA_m algorithm for feature selection in microarray datasets has achieved better performance than other feature selection algorithms in various microarray datasets. A study comparing the WOA_m algorithm to different feature selection algorithms on a microarray dataset for cancer classification reported that the WOA_m algorithm achieved comparable performance to the best algorithm in terms of accuracy and AUC.

It is important to note that the results of using the WOA_m algorithm for feature selection in microarray datasets may be affected by various factors, such as the quality of the data, the number of samples and features, and the specific implementation of the algorithm. It is also essential to consider the stability of the algorithm and use stability analysis techniques, such as the Jaccard index, as proposed by Kalousis, to evaluate the consistency of the feature subsets selected by the WOA_m algorithm across different runs.

Results in Table 4 show the proposed modification in the whale optimization algorithm over the other metaheuristic algorithms.

Table 4. Experimental results for implementation of iWOA, GWO, MPO, WOA, and WOAm

Dataset	**Class**	**Improved Whale Optimization Algorithm (iWOA)**				**Grey Wolf Optimization (GWO)**				**Marine Predator Optimization (MPO)**				**Whale Optimization Algorithm (WOA)**				**Proposed Whale Optimization Algorithm (WOAm)**			
		Prec	**Rec**	**Acc**	**F1**	**Prec**	**Rec**	**Acc**	**F1**	**Prec**	**Rec**	**Acc**	**F1**	**Prec**	**Rec**	**Acc**	**F1**	**Prec**	**Rec**	**Acc**	**F1**
Breast	0	0.7	0.7	0.7	0.69	0.63	0.75	0.65	0.69	0.4	0.53	0.37	0.46	0	0	0.5	0	0.74	0.86	0.78	0.79
	1	0.7	0.7	0.7	0.71	0.67	0.53	0.65	0.59	0.3	0.2	0.37	0.24	0.5	1	0.5	0.67	0.83	0.79	0.78	0.81
CNS	0	1	0.7	0.83	0.8	1	0.64	0.82	0.78	1	0.73	0.86	0.84	1	0.55	0.77	0.71	1	0.85	0.85	0.87
	1	0.7	1	0.83	0.84	0.73	1	0.82	0.85	0.79	1	0.86	0.88	0.69	1	0.77	0.81	0.71	0.91	0.85	0.86
Colon	0	1	0.8	0.84	0.86	0.86	1	0.92	0.92	0.83	0.83	0.83	0.83	0.69	0.82	0.74	0.75	0.91	0.85	0.88	0.86
	1	0.8	1	0.84	0.89	1	0.83	0.92	0.91	0.83	0.83	0.83	0.83	0.8	0.67	0.74	0.73	0.72	0.96	0.88	0.83
Leukemia	0	0.9	0.9	0.89	0.89	0.77	0.67	0.72	0.71	1	0.79	0.89	0.88	0.87	0.93	0.89	0.9	0.9	0.64	0.81	0.88
	1	0.9	0.9	0.89	0.89	0.69	0.79	0.72	0.73	0.81	1	0.89	0.9	0.92	0.86	0.89	0.89	0.71	0.92	0.81	0.9
Ovarian	0	0.6	0.9	0.62	0.7	0.72	0.98	0.8	0.83	0.59	1	0.66	0.74	0.96	0.5	0.74	0.66	1	0.88	0.94	0.93
	1	0.8	0.4	0.62	0.49	0.97	0.61	0.8	0.75	1	0.31	0.66	0.48	0.66	0.98	0.74	0.79	0.94	0.96	0.94	0.94
OSCC	0	0.9	0.9	0.9	0.9	1	1	1	1	1	0.91	0.95	0.95	0.92	1	0.95	0.96	1	1	1	1
	1	0.9	0.9	0.9	0.9	1	1	1	1	0.91	1	0.95	0.95	1	0.9	0.95	0.95	1	1	1	1

*Acc: Accuracy, Prec: Precision, Rec: Recall, F1: F1- score.

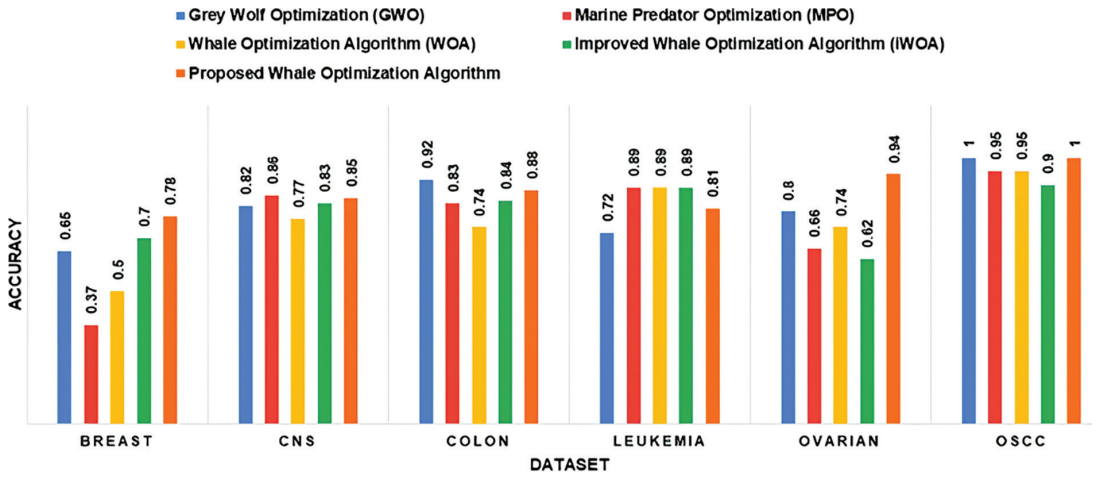

Figure 1. Comparative accuracy analysis of proposed WOA with other Metaheuristics algorithms

The chosen feature set's stability is also evaluated during the research work. The stability issue in feature selection has received much attention recently. Stability means the insensitivity of the result of a feature selection algorithm to small perturbation to the training set. Evaluation of stability, mandates the researchers to measure stability—to know whether they have increased it, or decreased it. Three different forms of outputs are possible from feature selection processes: a feature set, a feature ranking, and a feature weighting. For each form of output, different stability measures are available (Theng and Bhoyar, 2022). In this experiment, the Jaccard index is used for measuring the stability of the feature subset generated. Kalousis (Kalousis et al. 2007) were the first to discuss stability in-depth, independent of the particular feature selection algorithm. The stability index has a value range [min; max] as [0; 1]. Jaccard index is proposed by Kalousis and it is defined by the formula below:

$$\frac{r_{i,j}}{|S_i \cup S_j|} \tag{6}$$

A higher Jaccard index value indicates a higher degree of similarity between the two feature subsets, and a lower Jaccard index value indicates a lower degree of similarity. By comparing the Jaccard index of the feature subsets selected by the proposed feature selection algorithm in different runs, it is possible to determine how consistent the algorithm is in identifying the same or similar features. This can provide valuable information about the stability of the algorithm, which can help in evaluating the performance of the algorithm and in guiding the selection of appropriate parameter settings.

Stability analysis by eq. (1) the proposed WOA_m algorithm and all the state-of-the-art metaheuristic algorithms are shown in Table 5. Stability analysis aims to determine how consistent the algorithm

is in finding near-optimal solutions across different runs. According to the findings analyzed from Table 5 observations, the proposed WOA_m has superior stability compared to other algorithms for the CNS, colon, Leukemia, and OSCC datasets. The highest stability score range [0.48, 0.94] achieved by WOA_m for the cancer microarray datasets, shows that it is the most robust technique for biomarker identification. Additionally, the results show that WOA_m is robust to the curse of dimensionality and can effectively search for informative features among the large number of features in microarray datasets. Furthermore, the WOA_m algorithm is not sensitive to the initial solution and can converge to the optimal solution in a reasonable number of iterations.

Table 5. Comparing the Stability of Proposed-WOA with other meta-heuristic contemporary algorithms

Datasets	**Grey Wolf Optimization (GWO)**	**Marine Predator Optimization (MPO)**	**Whale Optimization Algorithm (WOA)**	**Improved Whale Optimization Algorithm (iWOA)**	**Proposed Whale Optimization Algorithm (WOAm)**
Breast	0.46	0.18	0.36	0.54	0.52
CNS	0.54	0.25	0.61	0.75	0.89
Colon	0.57	0.24	0.65	0.77	0.94
Leukemia	0.51	0.23	0.66	0.69	0.81
Ovarian	0.49	0.22	0.35	0.49	0.48
OSCC	0.53	0.17	0.41	0.58	0.58

Figure 2 is the convergence graph plot that shows how the value of an objective function changes as an optimization algorithm progresses. The x-axis represents the number of iterations, while the y-axis represents the value of the objective function. For the proposed WOA_m the value of the objective function is decreasing over time, this indicates that the optimization algorithm is making progress towards finding a good solution. It's worth noting that a convergence graph is just one way to visualize the progress of an optimization algorithm. Another parameter stability of the selected feature set by the proposed WOA_m is compared in Figure 3 with the stability score of the other state-of-the-art algorithms. The stability score of a selected feature subset is a measure of how robust the feature selection process is to small perturbations in the data. A feature subset generated by the proposed WOA_m achieved a high stability score which is likely to be composed of features that are consistently informative across different samples of the same data. The highest stability index achieved by proposed WOAm 0.89, 0.94, 0.84, and 0.58 for CNS, Colon, Leukemia, and OSCC datasets respectively recommends that the feature selection process is relatively stable and is likely to produce consistent results across different samples of the data.

The proposed modified Whale Optimization Algorithm (WOAm) has demonstrated superior performance in feature selection for microarray datasets compared to other algorithms. Experimental results on various microarray datasets, including Breast, CNS, Colon, Leukemia, Ovarian, and OSCC,

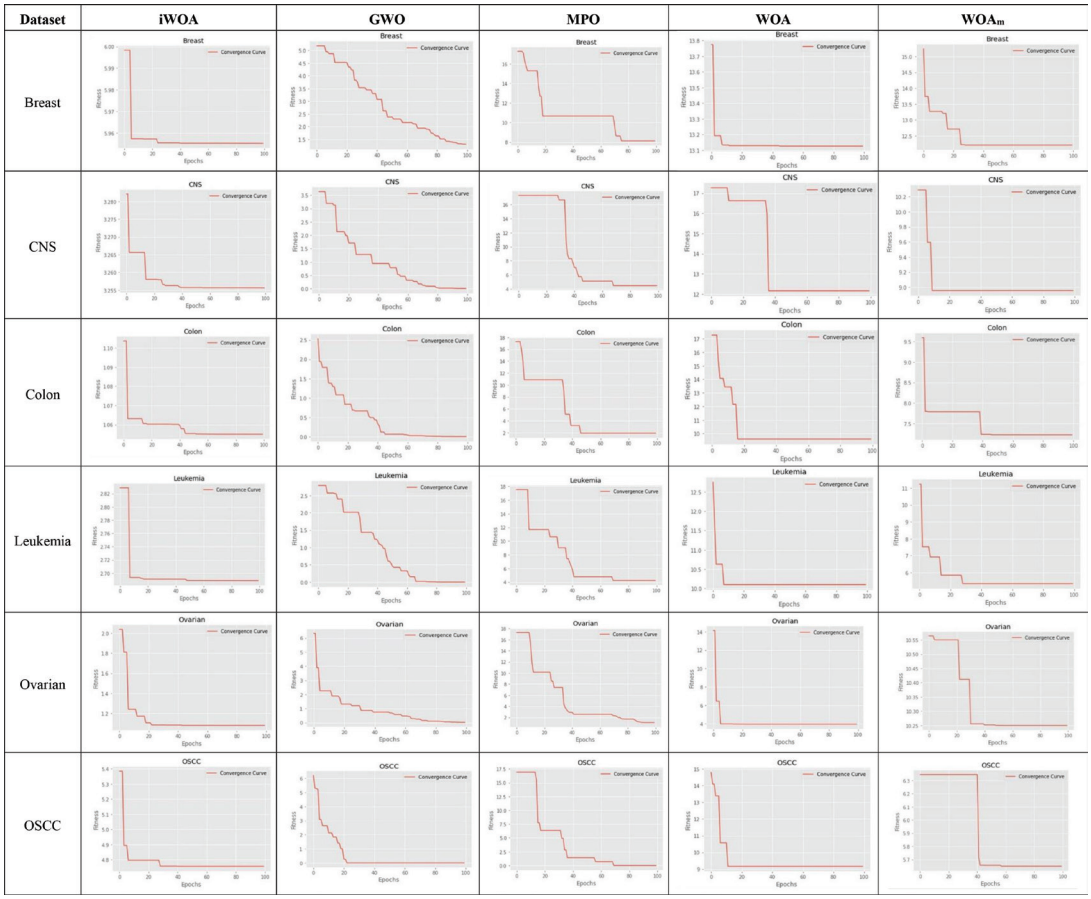

Figure 2. Comparing the WOAm convergence curves with other metaheuristic algorithms for six microarray datasets

showed that WOAm achieved higher accuracy, precision, recall, and F1-score compared to algorithms such as iWOA, GWO, MPO, and WOA. For example, in the Breast dataset, WOAm achieved an accuracy of 0.78, outperforming the other algorithms. Furthermore, stability analysis using the Kuncheva index revealed that WOAm exhibited the highest stability scores across different datasets, ranging from 0.48 to 0.94. This indicates the consistency of WOAm in identifying informative features and its robustness to small perturbations in the data. The convergence analysis also showed that the value of the objective function decreased over time, indicating the algorithm's progress towards finding optimal solutions. These results highlight the effectiveness and reliability of WOAm for feature selection in microarray datasets, making it a valuable tool for biomarker identification and optimization tasks in the field.

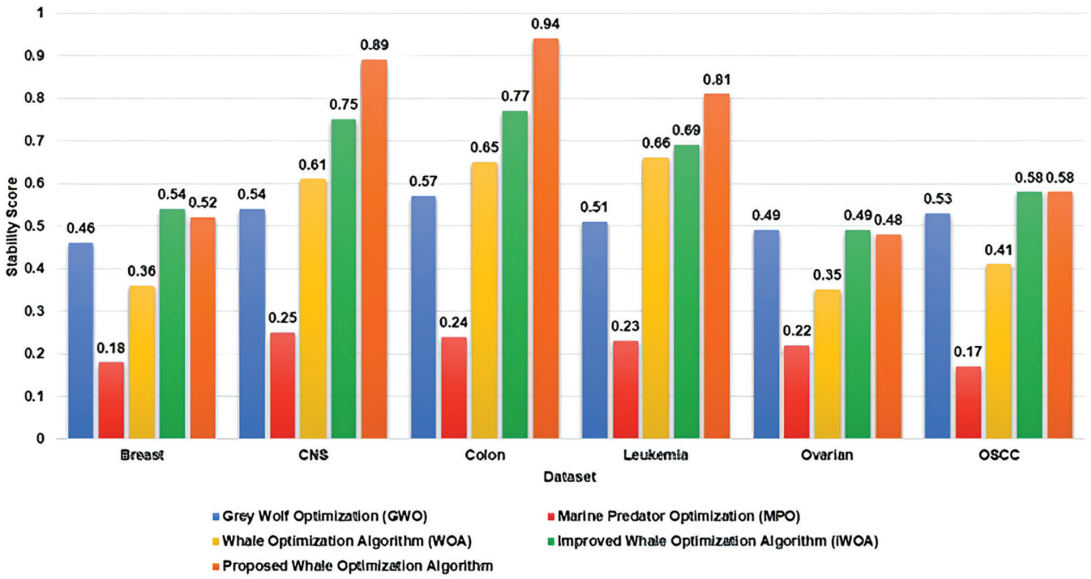

Figure 3. Stability Comparison of Proposed WOAm with other Metaheuristics Optimization Techniques

5. Conclusion

Metaheuristic algorithms give promising results for feature selection on high-dimensional datasets. A study of selected benchmark metaheuristic algorithms for biomarker identification from the microarray dataset is presented with comparative results. Our proposed algorithm (WOAm) has improved the classification performance evaluated using parameters like accuracy, precision, and recall. The proposed modification in the whale optimization algorithm is implemented and compared with the traditional whale optimization algorithm, improved whale optimization algorithm, grey wolf optimization, and marine predator optimization algorithm. These comparative results are presented for six microarray datasets related to various types of cancers. From the results obtained, it is observed that the proposed modification for the whale optimization algorithm has significantly improved the results of feature selection across all the datasets. It is found that our proposed whale optimization algorithm has outperformed all the benchmarking algorithms discussed in this study. Proposed WOAm have shown a good convergence rate over the other metaheuristic algorithms.

Also performed stability analysis using the Kuncheva index for all implemented algorithms. Stability indicates the confidence and efficiency of biomarker identification in microarray datasets. Our proposed whale optimization algorithm has shown good stability for CNS, Colon, Leukemia, and OSCC datasets. However, improved whale optimization and grey wolf optimization have shown better stability for the ovarian dataset. As proposed whale optimization achieved better stability for most of the algorithms, this has proved it an effective algorithm for feature selection. The proposed approach can be further investigated using datasets other than microarray datasets where feature reduction is highly required. Future directions for feature selection include exploring hybrid metaheuristic approaches, enhancing scalability and efficiency, investigating cross-dataset generalization, extending beyond

microarray datasets, and enhancing interpretability. By exploring these future directions, researchers can further advance the field of feature selection for microarray datasets, improving the accuracy, efficiency, and applicability of metaheuristic algorithms like the proposed WOAm.

6. References

Abu Khurma, R., Aljarah, I., Sharieh, A., Abd Elaziz, M., Damaševičius, R., & Krilavičius, T., 2022. A review of the modification strategies of the nature inspired algorithms for feature selection problem. *Mathematics, 10*(3), 464. https://doi.org/10.3390/math10030464

Agrawal, Prachi, et al. "Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019)." IEEE Access 9 (2021): 26766-26791. https://doi.org/10.1109/ACCESS.2021.3056407

Alrefai, N., & Ibrahim, O., 2022. Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. *Neural Computing and Applications*, 1-16. https://doi.org/10.1007/s00521-022-07147-y

Alzaqebah, M., Briki, K., Alrefai, N., Brini, S., Jawarneh, S., Alsmadi, M. K., & Alqahtani, A., 2021. Memory based cuckoo search algorithm for feature selection of gene expression dataset. *Informatics in Medicine Unlocked*, 24, 100572. https://doi.org/10.1016/j.imu.2021.100572

Aziz, R. M., 2022. Cuckoo Search-Based Optimization for Cancer Classification: A New Hybrid Approach. *Journal of Computational Biology, 29*(6), 565-584. https://doi.org/10.1089/cmb.2021.0410

Fahrudin, T. M., Syarif, I., & Barakbah, A. R., 2016, September. Ant colony algorithm for feature selection on microarray datasets. In *2016 International Electronics Symposium (IES)* (pp. 351-356). IEEE. https://doi.org/10.1109/ELECSYM.2016.7861030

Faramarzi, A., Heidarinejad, M., Mirjalili, S., & Gandomi, A. H., 2020. Marine Predators Algorithm: A nature-inspired metaheuristic. *Expert systems with applications, 152*, 113377. https://doi.org/10.1016/j.eswa.2020.113377

Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst. 12*, 95-116 (2007). https://doi.org/10.1007/s10115-006-0040-8

Khaire, U. M., & Dhanalakshmi, R., 2022. Stability investigation of improved whale optimization algorithm in the process of feature selection. *IETE Technical Review, 39*(2), 286-300. https://doi.org/10.1080/02564602.2020.1843554

Mahendru, S., & Agarwal, S., 2019. Feature selection using metaheuristic algorithms on medical datasets. In *Harmony Search and Nature Inspired Optimization Algorithms* (pp. 923-937). Singapore: Springer. https://doi.org/10.1007/978-981-13-0761-4_87

Mirjalili, S., & Lewis, A., 2016. The whale optimization algorithm. *Advances in engineering software, 95*, 51-67. https://doi.org/10.1016/j.advengsoft.2016.01.008

Mirjalili, S., Mirjalili, S. M., & Lewis, A., 2014. Grey wolf optimizer. *Advances in engineering software, 69*, 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007

Pashaei, E., & Pashaei, E., 2022. Hybrid binary COOT algorithm with simulated annealing for feature selection in high-dimensional microarray data. *Neural Computing and Applications*, 1-22. https://doi.org/10.1007/s00521-022-07780-7

Perez, M., & Marwala, T., 2012, November. Microarray data feature selection using hybrid genetic algorithm simulated annealing. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel* (pp. 1-5). IEEE. https://doi.org/10.1109/EEEI.2012.6377146

Septian, F., & Utami, E., 2022. Whale Optimization Algorithm for Medical Diagnostic: A Systematic Literature Review. *Jurnal Sistem Informasi Komputer dan Teknologi Informasi (SISKOMTI), 5*(2), 33-44.

Shojaee, Z., Shahzadeh Fazeli, S. A., Abbasi, E., Adibnia, F., Masuli, F., & Rovetta, S., 2022. A Mutual Information Based on Ant Colony Optimization Method to Feature Selection for Categorical Data Clustering. Iranian Journal of Science and Technology, Transactions A: Science, 1-12. https://doi.org/10.1007/s40995-022-01395-2

Shukla, A. K., Tripathi, D., Reddy, B. R., & Chandramohan, D., 2020. A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges. *Evolutionary Intelligence, 13*(3), 309-329. https://doi.org/10.1007/s12065-019-00306-6

Silva, S. R. D., & Gertrudes, J. C., 2022, July. Chaotic genetic bee colony: combining chaos theory and genetic bee algorithm for feature selection in microarray cancer classification. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 296-299). https://doi.org/10.1145/3520304.3528901

Simumba, Naomi, et al., 2022. Multiple objective metaheuristics for feature selection based on stakeholder requirements in credit scoring. *Decision Support Systems, 155*, 113714. https://doi.org/10.1016/j.dss.2021.113714

Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G., 2006, July. Improving feature subset selection using a genetic algorithm for microarray gene expression data. In *2006 IEEE International Conference on Evolutionary Computation* (pp. 2529-2534). IEEE.

Tawhid, M. A., & Ibrahim, A. M., 2020. Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm. *International journal of machine learning and cybernetics, 11*(3), 573-602. https://doi.org/10.1007/s13042-019-00996-5

Theng, D., & Bhoyar, K. K., 2022, July. Stability of Feature Selection Algorithms. In *Artificial Intelligence on Medical Data: Proceedings of International Symposium, ISCMM 2021* (pp. 299-316). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0151-5_26

Theng, D., & Bhoyar, K. K., 2022, October. Feature Selection Techniques for Bioinformatics Data Analysis. In *2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)* (pp. 46-50). IEEE. https://doi.org/10.1109/GECOST55694.2022.10010541

Wang, X., & Jia, W., 2022, December. A Feature Weighting Particle Swarm Optimization Method to Identify Biomarker Genes. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 830-834). IEEE. https://doi.org/10.1109/BIBM55620.2022.9995376

Xie, W., Fang, Y., Yu, K., Min, X., & Li, W., 2022. MFRAG: Multi-Fitness RankAggreg Genetic Algorithm for biomarker selection from microarray data. *Chemometrics and Intelligent Laboratory Systems, 226*, 104573. https://doi.org/10.1016/j.chemolab.2022.104573

Zhong, C., Li, G., Meng, Z., Li, H., & He, W., 2023. A self-adaptive quantum equilibrium optimizer with artificial bee colony for feature selection. *Computers in Biology and Medicine*, 106520. https://doi.org/10.1016/j.compbiomed.2022.106520

7. Conflict of Interest

Authors declare no conflicts of interest