A Novel Multi-objective Model for Feature Selection in Cancer Data

Himanshu Sekhar Panigrahi
M. Sc. Data Science
Silicon University
Bhubaneswar, India
datasceince.22mdsa76@silicon.ac.in

Itishree Mishra
M. Sc. Data Science
Silicon University
Bhubaneswar, India
datasceince.22mdsa67@silicon.ac.in

Pritish Mishra
M. Sc. Data Science
Silicon University
Bhubaneswar, India
datasceince.22mdsa62@silicon.ac.in

Jayashree Piri
Dept. of CSE
Silicon University
Bhubaneswar, India
jayashree.piri@silicon.ac.in

Abstract*—Cancer is a vital disease taking several lives every year. Early detection can prevent lots of terrible cases. Traditional methods frequently don't achieve the best possible accuracy of diagnosis in the detection of cancer. This paper presents a complete framework to improve the feature selection process for cancer detection by integrating a multi-objective optimization approach. Developing a strong multi-objective feature selection method that optimizes computing efficiency, sensitivity, and specificity is one of the main goals. The proposed methodology leverages multi-objective feature selection and the Crayfish Optimization Algorithm (COA), using clues from crayfish behavior to improve exploration and exploitation at different phases. Thorough research and analysis have shown that the hybrid technique works well for optimizing feature subsets. When multi-objective optimization and COA are used, interpretability is enhanced as well as the precision of classification. The study's findings indicate positive advancements in the field of cancer, and the suggested design holds great potential for streamlining feature selection procedures and lowering the complexity of cancer data. MO-COA excels over MOGA and MOPSO, achieving 96.7% average accuracy in breast cancer with 3 average features, surpassing MOGA's 95.3% (2 features) and MOPSO's 95.7% (4 features). In lung and cervical cancer, MO-COA achieves 60.5% (10 features) and 95.5% (5 features) accuracies respectively, outperforming MOGA (52.2% lung, 95.2% cervical) and MOPSO (52.4% lung, 94.2% cervical). The average is taken for total non-dominated solutions for each algorithm over different datasets. These findings underscore MO-COA's efficiency over other algorithms in enhancing cancer diagnosis accuracy with minimal features, benefiting both research and clinical practice.

***Index Terms*—Cancer, MO-Optimization, Feature Selection, COA, Crowding distance**

I. Introduction

In modern medical research, cancer diagnosis is a major difficulty because of the increasing complexity of data, the vital need for early detection, and the requirement for accurate diagnostic procedures. Due to the complex and multidimensional nature of cancer data, innovative methods are required to improve the accuracy of diagnosis and direct efficient treatment strategies [1]. Conventional diagnostic techniques struggle to handle the complexity of cancer data, which leads to less-than-ideal computational efficiency, sensitivity, and specificity.

To address these issues, this work offers a novel framework that revolutionizes cancer classification [2] [3] [4] by combining multi-objective optimization techniques with the Crayfish Optimization Algorithm (COA). Developing a strong multi-objective feature selection method specifically for cancer data is one of the goals, as is maximizing vital variables like sensitivity, specificity, and the effectiveness of computing. In addition, the research aims to smoothly integrate sophisticated machine learning models for categorization, making use of the features that are retrieved to improve the predicted accuracy of cancer detection. Even with the great advancements in optimization algorithms for feature selection, handling interdependent attributes and traversing vast search spaces continue to be difficult, particularly when dealing with cancer data. To discover the best feature subsets, many studies used evolutionary techniques including DE [5], GA [6], GWO [7], ACO [8], WOA [9] and BHHO [10], BPSO [11], among others. To address competing goals in cancer research, this paper presents MOFS (multi-objective feature selection), in recognition of the necessity for efficient global search tools. The analysis emphasizes how crucial natively inspired algorithms like COA are in the ever-changing field of cancer data. The integration of multi-objective optimization techniques guarantees an exhaustive search of the solution space by concurrently maximizing competing objectives. By utilizing a modified COA for cancer data, the study seeks to substantially advance the improvement of classification accuracy and interpretability in cancer research by converting it into a multi-objective feature selection technique. Thorough assessments utilizing three widely-used medical datasets contrast the suggested MO-COA with renowned multi-objective methods, such as MOGA [12] and MOPSO [11]. The potential of the suggested framework to minimize feature subsets through careful analysis and testing, promises advancements in the field of cancer research. By doing through rigorous comparisons with

established methods, the proposed framework demonstrates its potential to revolutionize prominent feature selection.

II. Background and Related work

The COA, proposed by Hemming in 2023, draws inspiration from the Crayfish [13]. Within the program, its behaviors are logically split into distinct phases to efficiently reconcile exploration and exploitation. Mimicking the adaptability of crayfish, which thrive in diverse freshwater habitats and exhibit resilience to varying temperature conditions, COA proves to be a robust optimization technique. With a focus on diversification strategies to enhance universal search potential, COA outperforms well-known optimization techniques like PSO [11], and GA [6]. The algorithm incorporates a dynamically randomized escape energy criterion to ensure smooth switching between searches conducted locally and globally. Hemming's COA has demonstrated superior performance by transforming it into a rational and worthwhile choice for optimization tasks, surpassing the capabilities of several established techniques. The steps of COA are:

A. Initialize Population

All the crayfish in the multiple dimension-based optimization problem are a $1 \times \text{dim}$ matrix. A problem's solution is represented by each column matrix. Every variable (X_i) is a collection of variables $(X_{i,1}, X_{i,2}, \ldots, X_{i,\text{dim}})$ and must be positioned between the top and bottom boundary. Initially, a collection of feasible fixes X inside the area is generated at random by the COA, and N number of samples which is given by equation 1.

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,dim} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i,1} & \cdots & X_{i,j} & \cdots & X_{i,dim} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & X_{N,j} & \cdots & X_{N,dim} \end{bmatrix} \tag{1}$$

where,

$$X_{i,j} = \text{lb}_j + (\text{ub}_j - \text{lb}_j) \times \text{rand} \tag{2}$$

Each sample in the matrix is randomly assigned binary values (1, 0) denoting the presence or absence of particular features. Since the original COA was created for continuous optimization [13] rather than binary-encoded matrices, it is modified.

B. Specify the Temperature and Crayfish intake.

Crayfish's action is affected by temperature changes.

$$\text{temp} = \text{rand} \times 15 + 20 \tag{3}$$

$$p = a \times \left(\frac{1}{\sqrt{2\pi\sigma}} \times \exp\left(-\frac{(\text{temp} - \mu)^2}{2\sigma^2} \right) \right) \tag{4}$$

(μ=25) is the most suitable temp and (σ=3, C_1=0.2) used to control intake at different temp.

C. Summer Resort Stage (Exploration)

$$X_{\text{shade}} = \frac{F_{\text{g}} + F_{\text{l}}}{2} \tag{5}$$

The occupancy of caves is random. If rand < 0.5 and $T > 30$, a crayfish will enter the cave straight at iteration t by using the following equation:

$$X_{i,j}^{t+1} = X_{i,j}^{t} + b \times \text{rand} \times (X_{\text{shade}} - X_{i,j}^{t}) \tag{6}$$

where, T= maximum number of iterations

$$b = 2 - \frac{t}{T} \tag{7}$$

This stage promotes quicker convergence.

D. Competition Stage (Exploitation)

When rand ≥ 0.5 and $T > 30$, indicating the curiosity in the cave shown by other crayfish. So Crayfish engage in a contest for the cave. Here,

$$X_{i,j}^{t+1} = X_{i,j}^{t} - X_{z,j}^{t} + X_{\text{shade}} \tag{8}$$

$$where, z = \text{round}(\text{rand} \times (N-1)) + 1 \tag{9}$$

Crayfish adjusts their positions based on another crayfish's location (X_z) which expands the search range of COA, enhancing exploration capabilities.

E. Foraging Stage (Exploitation)

When temp ≤ 30, i.e. suitable for feeding.

$$food\ location, (X_{food}) = F_{\text{g}} \tag{10}$$

$$\text{food size, Q} = c \times \text{rand} \times \frac{\text{fitness}_i}{\text{fitness}_{\text{food}}} \tag{11}$$

where c is the largest food (set to 3). When $Q > \frac{C_3+1}{2}$, crayfish rip the meal using first claw foot by

$$X_{\text{food}} = \exp\left(-\frac{1}{Q}\right) \times X_{\text{food}} \tag{12}$$

For shredded food, crayfish uses cosine and sine blending functions as given below to provoke the alternating procedure.

$$X_{i,j}^{t+1} = X_{i,j}^{t} + X_{\text{food}} \times p \times (\cos(2 \times \pi \times \text{rand}) - \sin(2 \times \pi \times \text{rand})) \tag{13}$$

When $Q \leq \frac{C_3+1}{2}$, crayfish advances straight in the direction of the food (where C_3 = 3), as per the formula given below.

$$X_{i,j}^{t+1} = (X_{i,j}^{t} - X_{\text{food}}) \times p + p \times \text{rand} \times X_{i,j}^{t} \tag{14}$$

Crayfish adjust their feeding strategies throughout the foraging stage in response to changes in food size Q, with X_{food} serving as the ideal outcome. The foraging step facilitates convergence toward the best option and improves COA's capacity for exploitation.

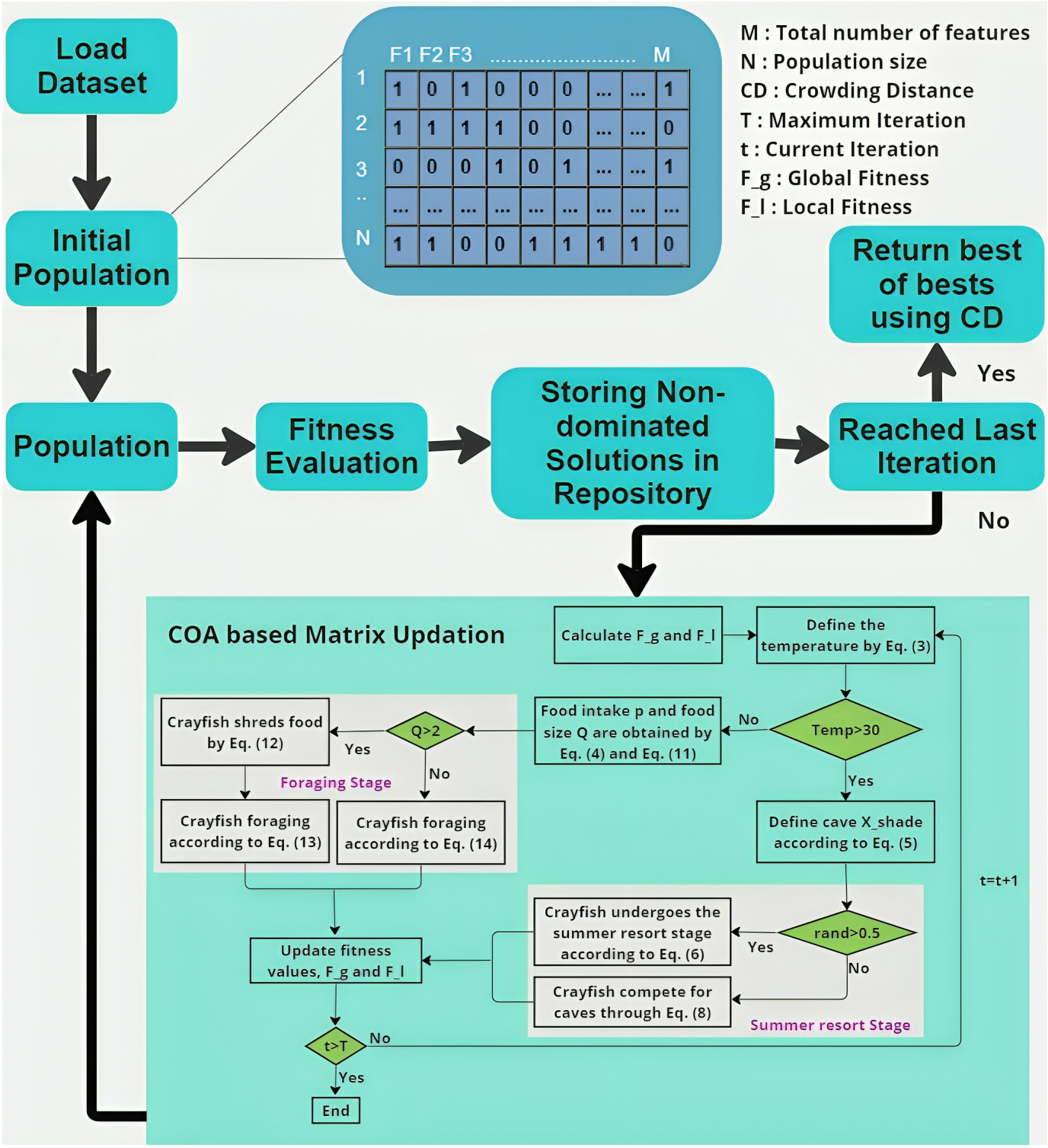

Fig. 1. System architecture of MO-COA based feature selection

F. Multi-objective Optimization (MOO)

MOO [11] [10] entails simultaneously optimizing several objective functions. Optimal solutions are found by applying the Pareto dominance principle, which states that a solution S_1 dominates S_2 if S_1 is purely better in a minimum of one goal and is not bad in any of the objectives considered for evaluation. The answers that are not suppressed by another solution among the search spaces are the optimal Pareto solutions, which make up the Pareto front.

III. Proposed Feature Selection Method

The step-by-step procedure for the suggested feature selection technique is shown in Fig. 1.

The detailed explanation of each step is given below:

- **Population Initialization:**
  Each crayfish is encoded as a vector of 0s and 1s showing the absence and presence of a particular feature in the feature vector. For the pictorial representation of the crayfish location string, refer to Fig. 1.
- **Fitness Assessment**
  Each sample is taken and two objectives are evaluated as no. of features (F1) and corresponding classification accuracy of selected columns in the original dataset (F2) [10]. A repository will be created to store the F1 and F2 values for each sample. The highest rank sample of the matrix will be local fitness(F_l).
- **Repository Management**
  After evaluating all samples in the population, the external archive [12] is updated in the multi-objective COA. At the end of each cycle, a collection of Pareto optimum results is generated. The newly generated Pareto solutions are again compared with already existing solutions in the repository one by one. If it wins and space is there in the repository then it will be added. After completing all the iterations, the best from the best is chosen by using crowding distance. The most effective non-dominated solution from the repository will be global fitness(F_g).
- **Updating Population**
  1) While calculating Objective 2 the index of rows containing ones are matched with those rows of the original dataset and selected rows are taken into consideration for classification so for each sample in the matrix a different classification accuracy comes out by using KNN.
  2) The local and global fitness [13] values remain similar for the initial state but after the first iteration the F_G and F_L are calculated and are updated after each iteration as the original matrix changes.
  3) When it comes to population update the (X_shade) is calculated for each value in the corresponding row with three possible values as 0, 0.5, or 1.
  4) The process for temperature intake and Summer resort stage remain similar except for one thing for eq.(6) sample value update we have to calculate it for m x n elements of the population.
  5) If crayfish undergoes Competition stage (exploitation) the z value will be calculated as per eq. (9) and put in the eq. (8).
  6) For the foraging stage few modifications will be done while calculating Q using eq. (11), the $fitness_{\text{food}}$ will be F_g i.e. global fitness value and $fitness_i$ will be calculated using the below formula $fitness_i$ = (N-$rank_i$)+1 where if $dominate_{\text{count}}$=0, rank =1. Use the sigmoid function as per eq. (15) to make it binary.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{15}$$

IV. Experimental Evaluation

The MOCOA-based approach for feature selection was first evaluated by implementing a set of 3 clinical datasets i.e. Lung, Breast, and Cervical cancer from the UCI repository. The details of all datasets are given in Table I. Our Experiment began with repository maintenance and fitness selection by crowing distance-based algorithm.

The sigmoid function transforms the updated matrix value into a discrete binary format after each iteration.

All of the answers that are not dominated by one another are stored in the external archive after the predetermined number of repetitions. The crowding distance selects the most important features based on objectives 1 and 2 as discussed earlier. Table II shows a statistical comparison between content of the non-dominated repository after applying different algorithms. For the standard case, we have kept all the processes of

TABLE I
Details of the Dataset

Name	Samples	Features	Classes
Breast Cancer	699	9	2
Lung Cancer	32	56	3
Cervical Cancer	858	35	2

crowding distance-based non-dominated solution selection and objective 1 and 2 selection processes similar for 3 algorithms. Only the matrix updation logic after each iteration is different for different algorithms.

The primary goal of this method involves successfully selecting a set of important features from a large no. of existing features available in data. But to showcase the successful initial implementation and working efficiency of the modified algorithm we have gone through a set of clinical data feature selections by taking the same approach. The set of pairs (F1, F2) represented in Table II are objectives 1 and 2 i.e. no. of selected features and classification accuracy. Important features are considered based on dominance ranking over other solutions. The table provides a distinct idea of objectives selected by different algorithms.

Talking about the results, the MOCOA captured the most important features [14] while maintaining a high classification accuracy over another multi-objective algorithm, as lesser no. of features are selected. It outperforms traditional algorithms in datasets like breast, lung, and cervical cancer datasets.

TABLE II
ALGORITHMIC FEATURE SELECTION COMPARISON

DATASET	MOCOA	MOGA	MOPSO
Breast	(2, 0.959) (3, 0.966) (4, 0.971)	(2, 0.95) (3, 0.963) (7, 0.971) (6, 0.97)	(6, 0.97) (4, 0.964) (2, 0.95)
Lung	(4, 0.58) (5, 0.67)	(21, 0.55) (18, 0.45) (17, 0.417) (16, 0.358) (20, 0.467)	(11, 0.658) (9, 0.508) (10, 0.625)
Cervical	(2, 0.952) (8, 0.967) (4, 0.965)	(11, 0.951) (7, 0.959) (11, 0.963) (3, 0.936) (13, 0.969) (10, 0.96)	(6, 0.965) (5, 0.956) (3, 0.94) (2, 0.936) (4, 0.948) (6,0.965)

TABLE III
CD BASED IMPORTANT FEATURES

Dataset	Selected Features
Breast	Uniformity of Cell Size, Marginal Adhesion, Bland Chromatin
Lung	Age, Pollution Exposure Factor, BMI(Body mass index),Respiratory Conditions, Smoking History
Cervical	First sexual intercourse, IUD, STDs, Smokes (years)

Fig. 2, Fig. 3, and Fig. 4 are for the visual performance comparison of MOGA, MOCOA, and MOPSO algorithms over Breast, Lung, and Cervical cancers respectively. The Pareto fronts [12] produced by MOGA and MOPSO, are two well-liked multi-objective evolutionary benchmarking techniques. For all three datasets, our proposed method produces Pareto fronts that are above the other two methods. It shows that the MOCOA-based feature selection method can capture very relevant features by giving the highest classification accuracy.

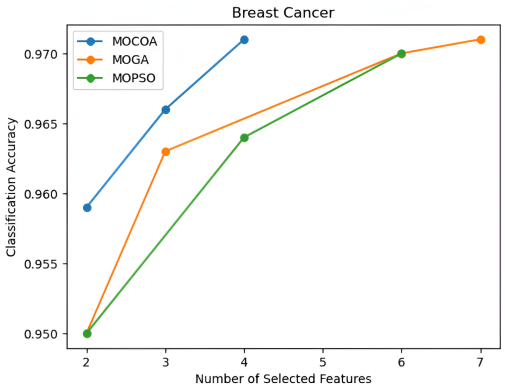

Fig. 2. Performance Comparision of Different Algorithms on Breast Cancer

Table III lists the key features that the MOCOA algorithm identified using the crowding distance measure [10]. The ranking was calculated by taking objectives 1 and 2 into consideration. To choose a specific solution from the less

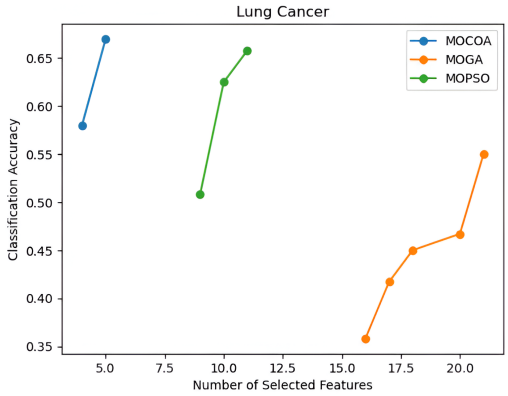

Fig. 3. Performance Comparision of Different Algorithms on Lung Cancer

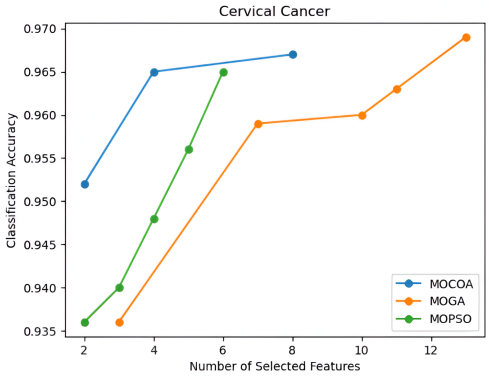

Fig. 4. Performance Comparision of Different Algorithms on Cervical Cancer

congested region of objective space for additional processing, the crowding distance measure is employed as a high-level piece of data.

V. CONCLUSION AND FUTURE WORK

The feature selection problem in cancer data is addressed by the multi-objective crayfish optimization approach presented in this research. To make the feature selection procedure feasible, we converted data in continuous to binary form using the sigmoid function. Every sample in the population is evaluated for fitness using two objective functions. Three clinical data sets were used to compare, evaluate, and apply the methods. Utilizing the Pareto dominance and Crowing distance, the repository is updated and kept after every iteration. MO-COA excels over MOGA and MOPSO, achieving 96.7% average accuracy in breast cancer with 3 average features, surpassing MOGA's 95.3% (2 features) and MOPSO's 95.7% (4 features). In lung and cervical cancer, MO-COA achieves 60.5% (10 features) and 95.5% (5 features) accuracies respectively, outperforming MOGA (52.2% lung, 95.2% cervical) and MOPSO (52.4% lung, 94.2% cervical). The average is taken for total non-dominated solutions for each algorithm over different datasets. After a successful implementation of clinical datasets,

the main target of selecting important characteristics based on numerous objectives was cleared.

Following the implementation's success, the field of microarray data classification gained even more attraction for our research. Future research aims to extend the algorithm to microarray-based genetic data and, if feasible, to multiple other types of genetic disease datasets, without restricting its use to the boundary at the end of genetic cancer. Through slight adjustments to the algorithm parameters, we will analyze and assess the application of MOCOA as a more effective feature selection method.

REFERENCES

[1] H. Fathi, H. AlSalman, A. Gumaei, I. I. Manhrawy, A. G. Hussien, P. El-Kafrawy *et al.*, "An efficient cancer classification model using microarray and high-dimensional data," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[2] M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions," *International Journal of Cognitive Computing in Engineering*, vol. 1, pp. 78–97, 2020.

[3] R. K. Singh and M. Sivabalakrishnan, "Feature selection of gene expression data for cancer classification: a review," *Procedia Computer Science*, vol. 50, pp. 52–57, 2015.

[4] H. Saberkari, M. Shamsi, M. Joroughi, F. Golabi, and M. H. Sedaaghi, "Cancer classification in microarray data using a hybrid selective independent component analysis and υ-support vector machine algorithm," *Journal of medical signals and sensors*, vol. 4, no. 4, p. 291, 2014.

[5] H. Thottathyl and K. K. Pavan, "Differential evolution model for identification of most influenced gene in brest cancer data," *Ingenierie des Systemes d'Information*, vol. 27, no. 3, p. 487, 2022.

[6] M. J. Rani and D. Devaraj, "Microarray data classification using multi objective genetic algorithm and svm," in *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. IEEE, 2019, pp. 1–3.

[7] S. Kumar and M. Singh, "Breast cancer detection based on feature selection using enhanced grey wolf optimizer and support vector machine algorithms," *Vietnam Journal of Computer Science*, vol. 8, no. 02, pp. 177–197, 2021.

[8] A. Gupta, V. K. Jayaraman, and B. D. Kulkarni, "Feature selection for cancer classification using ant colony optimization and support vector machines," in *Analysis of biological data: a soft computing approach*. World Scientific, 2007, pp. 259–280.

[9] S. S. Devi and K. Prithiviraj, "Breast cancer classification with microarray gene expression data based on improved whale optimization algorithm," *International Journal of Swarm Intelligence Research (IJSIR)*, vol. 14, no. 1, pp. 1–21, 2023.

[10] J. Piri and P. Mohapatra, "An analytical study of modified multi-objective harris hawk optimizer towards medical data feature selection," *Computers in Biology and Medicine*, vol. 135, p. 104558, 2021.

[11] C. S. R. Annavarapu, S. Dara, and H. Banka, "Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm," *EXCLI journal*, vol. 15, p. 460, 2016.

[12] J. Piri, P. Mohapatra, and R. Dey, "Fetal health status classification using moga-cd based feature selection approach," in *2020 IEEE international conference on electronics, computing and communication technologies (CONECCT)*. IEEE, 2020, pp. 1–6.

[13] H. Jia, H. Rao, C. Wen, and S. Mirjalili, "Crayfish optimization algorithm," *Artificial Intelligence Review*, vol. 56, no. Suppl 2, pp. 1919–1979, 2023.

[14] A. Anaissi, P. J. Kennedy, and M. Goyal, "Feature selection of imbalanced gene expression microarray data," in *2011 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*. IEEE, 2011, pp. 73–78.