ANKARA SCIENCE UNIVERSITY, RESEARCHER
Vol. 5, No. 1, July 2025
e-ISSN: 2717-9494
Research Article/Araştırma Makalesi

A New Multiobjective Harris Hawk Optimization Algorithm for the Diagnosis of Breast Cancer

Alara SERMUTLU[1], Tansel DÖKEROĞLU[2*]
[1]TED University, Computer Engineering Department, https://orcid.org/0009-0005-3479-4640
[2]TED University, Computer Engineering Department, https://orcid.org/0000-0003-1665-5928
* tansel.dokeroglu@tedu.edu.tr

Received: 14 January 2025; Accepted: 29 July 2025

Reference/Atıf: A. Sermutlu ve T. Dökeroğlu, "A new multiobjective Harris Hawk Optimization algorithm for the diagnosis of breast cancer", Researcher, c. 05, sy. 01, ss. 17–25.

Abstract

Breast cancer, a highly prevalent and life-threatening disease, affects millions of individuals worldwide, particularly women. Feature-based methods are widely employed for early diagnosis of breast cancer, and selecting the optimal feature set remains a significant and challenging problem. In this study, we introduce a novel Multi-objective Harris Hawk Optimization algorithm, which integrates an adaptive K-Nearest Neighbor classifier. Comprehensive experiments were conducted on two well-known datasets. The proposed approach achieves 31-45% reductions in the total number of selected features across all datasets, significantly lowering computational costs and improving the accuracy of diagnostics up to 95-97%.

Keywords: Breast cancer, Harris Hawk, metaheuristic, K Nearest Neighbor

1. Introduction

Multi-objective metaheuristic algorithms are effective in feature selection by optimizing both accuracy and the number of selected features simultaneously [1]. They aim to balance the trade-off between maximizing classification performance and minimizing feature subsets. By exploring Pareto-optimal solutions, these algorithms enhance model efficiency, reduce computational complexity, and avoid overfitting, making them suitable for high-dimensional datasets in various applications. The Harris Hawk Optimization (HHO) algorithm is a recent nature-inspired metaheuristic optimization technique introduced in 2019 [2]. It mimics the cooperative hunting behaviour of Harris hawks in the wild, where they target prey in groups using both surprise and persistence strategies. HHO is designed to solve complex optimization problems by balancing exploration and exploitation in the search space.

In a recent study, Dokeroglu et al. proposed a multiobjective HHO for binary classification, aiming to reduce selected features while maximizing prediction accuracy. The algorithm demonstrates superior performance on benchmark datasets and a COVID-19 dataset [3]. Piri, J., & Mohapatra proposed a multi-objective optimization problem by proposing a Multi-Objective Quadratic Binary HHO (MOQBHHO) technique, integrating KNN as a wrapper classifier and crowding distance for solution selection [4]. Experimental results on twelve medical datasets demonstrate that MOQBHHO outperforms deep-based FS methods and other multi-objective algorithms in achieving an optimal trade-off between feature selection and classification accuracy. Selim et al. introduced an improved HHO algorithm for optimal Distributed Generation placement in radial distribution systems, aiming to minimize power loss, reduce voltage deviation, and improve voltage stability [5]. By enhancing HHO with a rabbit location mechanism and employing grey relation analysis for Pareto solutions, the proposed methods demonstrate superior performance on IEEE 33-bus and 69-bus systems compared to other optimization techniques.

Thawkar proposed a hybrid CSAHHO algorithm, combining the Crow Search Algorithm (CSA) and HHO, for feature selection and classification of masses in mammograms [6]. Using ANN and SVM classifiers, CSAHHO achieves superior performance with 97.85% accuracy, outperforming original CSA, HHO, and other state-of-the-art algorithms while using fewer features to enhance breast cancer diagnosis. Bandyopadhyay et al. proposed a two-stage pipeline for COVID-19 detection in CT scans, combining feature extraction with DenseNet and feature selection using an HHO algorithm enhanced

with Simulated Annealing (SA) and Chaotic initialization [7]. Evaluated on the SARS-COV-2 CT-Scan dataset, the method achieves an accuracy of 98.85% and reduces selected features by 75%, outperforming many state-of-the-art and hybrid meta-heuristic algorithms in both accuracy and feature reduction.

In this study, we implemented a multi-objective HHO algorithm for feature selection on the Wisconsin breast cancer dataset (WBCD) [8] and Wisconsin diagnostic breast cancer (WDBC) dataset, achieving notable results. The algorithm successfully balanced the trade-off between classification accuracy and feature reduction, improving accuracy significantly while reducing the number of selected features by 31-45%. This reduction enhances model efficiency and decreases computational complexity without compromising performance. By exploring Pareto-optimal solutions, the proposed approach demonstrates its ability to identify relevant features effectively. These results highlight the algorithm's potential for handling high-dimensional datasets, providing a robust method for feature selection in various classification tasks. Its performance surpasses many existing techniques.

2. Poroposed Multiobjective HNO Algorithm

This section briefly explains the proposed multiobjective HHO algorithm for the diagnosis of breast cancer. The algorithm begins by initializing a population of hawks, each representing a candidate solution. During the optimization process, the hawks adopt different strategies to simulate their natural hunting behaviours. In the exploration phase, hawks search for prey by randomly moving across the search space, promoting diversity and avoiding premature convergence. In the exploitation phase, hawks converge towards the prey using dynamic strategies, such as surprise pounce and soft or hard besiege tactics, to refine solutions and exploit the best regions.

One of HHO’s strengths is its simplicity and adaptability, making it suitable for various optimization problems in engineering, feature selection, scheduling, and machine learning. It is computationally efficient and requires fewer parameters compared to many other metaheuristics. Researchers have further enhanced HHO with hybrid models and variants, improving its performance in specific applications. Overall, HHO has proven to be a robust and versatile algorithm for solving real-world optimization challenges.

The HHO algorithm incorporates key parameters to mimic the hawks' energy and hunting behaviors, enhancing its ability to balance exploration and exploitation effectively. Among these parameters, energy (E) and r play crucial roles in determining the hawks' strategies:

The energy level of the prey (E) is modeled as a dynamic parameter to simulate the prey's attempt to escape. It decreases linearly over iterations and is represented as: $E=2E0\ (1-t/T)$, where E0 is the initial energy, t is the current iteration, and T is the maximum number of iterations.

When $|E|>1$, the Hawks focus on exploration, searching widely across the search space. When $|E|\leq 1$, the hawks shift to exploitation, zeroing in on the prey with more refined strategies like "soft besiege" or "hard besiege."

The parameter random factor (r) is a random value in the range [0, 1], used to probabilistically decide the Hawks’ movement strategy: $r < 0.5$: Indicates that hawks move randomly, mimicking the unpredictability of nature. r \geq 0.5: The hawks target the prey directly, focusing on convergence.

These parameters enable HHO to dynamically adjust its behaviour, making it versatile for a wide range of optimization problems. By modulating energy and incorporating randomness, the algorithm balances exploring the search space and exploiting promising regions to find optimal solutions efficiently. Figure 1 shows the energy level (E), q and r values of the HHO metaheuristic according to its diversification and intensification efforts. Algorithm 1 presents the details of the proposed multiobjective HHO algorithm for the diagnosis of breast cancer.

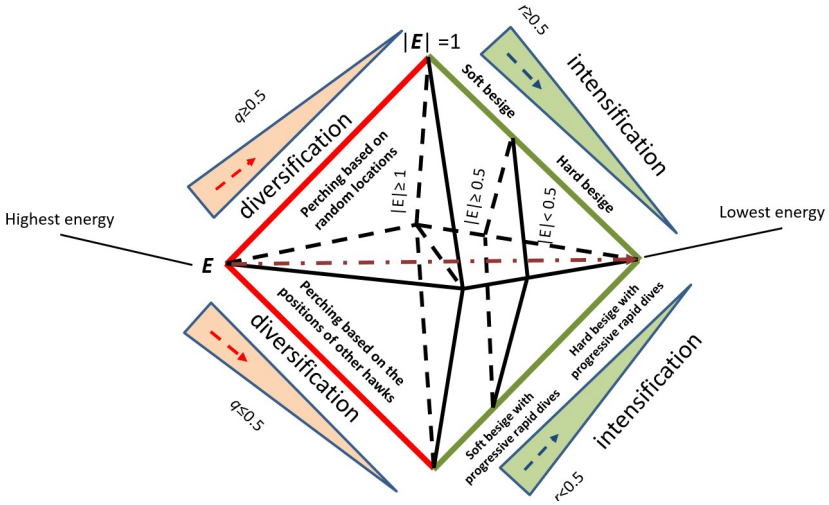

Figure 1. The steps of the HHO metaheuristic according to the energy level (E), q and r values.

Algorithm I. Multiobjective Harris Hawk Optimization Algorithm for feature selection

1	Initialize the population of hawks randomly (binary representation for feature subsets)	15	if r ≥ 0.5 and |E| ≥ 0.5 then Perform soft besiege
2	Set algorithm parameters (max iterations, population size, etc.)	16	else if r ≥ 0.5 and |E| < 0.5 then Perform hard besiege
3	Evaluate each hawk using two objectives:	17	else if r < 0.5 and |E| ≥ 0.5 then Perform soft besiege with rapid dives
4	1. Classification accuracy of the selected features (maximize)	18	else Perform hard besiege with rapid dives end if
5	2. Number of selected features (minimize)	19	end if Apply a repair mechanism if a solution violates constraints (e.g., empty feature subset)
6	Use a dominance-based mechanism (e.g., Pareto dominance) to identify the best solutions (prey)	20	end for Evaluate new solutions and update Pareto front
7	for each iteration do:	21	Update prey (best non-dominated solutions) if better solutions are found
8	Update to escaope energy of the prey(E)	22	end for Return the Pareto front of solutions (trade-off between accuracy and number of features)
9	for each hawk do:		
10	if |E| ≥ 1 then		
11	Perform exploration:		
12	Update hawk's position using a random search in binary space		
13	Else		
14	Perform exploitation:		

3. Experimental setup and evaluation of the results

The experiments in this study were conducted on a Huawei MateBook 14 laptop equipped with an AMD Ryzen 7 4800H processor, 16 GB of RAM, and 512 GB SSD, running the Windows 10 operating system.

Two datasets are used during our experiments, Wisconsin breast cancer dataset (WBCD) [8] and Wisconsin diagnostic breast cancer dataset (WDBC) [9]. The WBCD dataset consists of 699 samples characterized by nine numerical features. These features, derived from Fine Needle Aspiration (FNA) samples, capture various cellular and structural characteristics like cell size, shape and mitoses. The

dataset contains 16 incomplete records, which were excluded during preprocessing to ensure data homogeneity. This resulted in 683 complete samples, distributed as 444 benign and 239 malignant cases. The WDBC dataset includes 569 instances, each with 30 features extracted from a digitized image of an FNA of a breast mass. The features in the image represent the characteristics of the cell nuclei [9]. The details of the datasets are presented in Table 1.

Table 1. The details of the breast cancer datasets used in the experiments.

Dataset	Features	Instances	Malignant	Benign
WBCD	9	683	239 (35%)	444 (65%)
WDBC	30	569	212 (37.2%)	357 (62.8%)

In this part of our study, we conducted experiments on our datasets to evaluate the algorithm's performance with different numbers of generations and determine the optimal value for this parameter. A population size of ten individuals was chosen for this study, as it was previously shown to be sufficient for achieving reliable optimization. We tested both WDBC and WBCD datasets with 1, 2, 5, 10, 20, 30, and 50 generations. For the WBCD dataset, the optimal number of generations was 30, with accuracy declining beyond this point. Whereas, for the WDBC dataset, the highest accuracy was achieved at 20 generations, after which performance dropped. These results indicate that the ideal number of generations varies by dataset, with 30 being optimal for WBCD and 20 for WDBC.

Figures 2 and 3 present the average accuracy levels for different numbers of generations tested on the WBCD and WDBC datasets respectively.

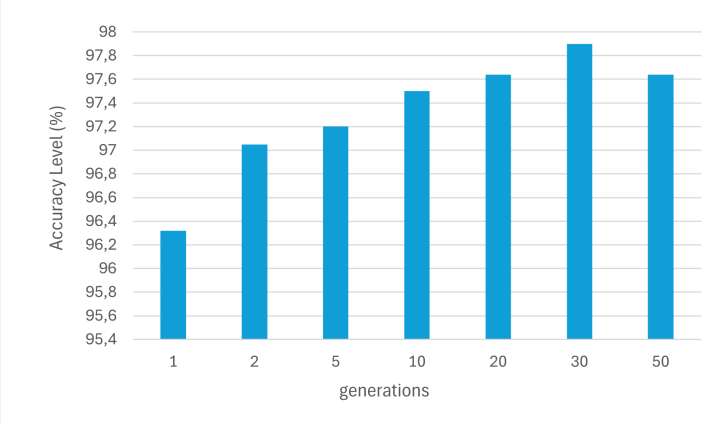

Figure 2. The average accuracy levels for different numbers of generations tested on the WBCD dataset.

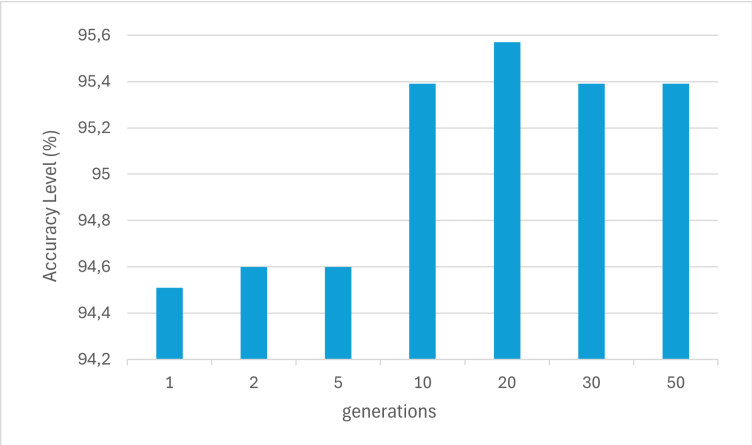

Figure 3. The average accuracy levels for different numbers of generations tested on the WDBC dataset.

The experiments, conducted over 10 iterations, are summarized in Table 2. For the WBCD dataset, the feature count was reduced from 9 to an average of 6.2, representing a 31% reduction, while maintaining an average accuracy of 97.0%. The maximum accuracy observed for the WBCD dataset was 97.5%. Similarly, for the WDBC dataset, the feature count was reduced from 30 to an average of 16.3 features, corresponding to a 45.6% reduction, with an average accuracy of 95.0%. The maximum accuracy for the WDBC dataset was 95.58%. These findings demonstrate the effectiveness of the proposed feature reduction approach in maintaining high classification accuracy.

Table 2. The average number of features in the populations and the accuracy levels for both datasets after executing 10 generations.

generations	1	2	3	4	5	6	7	8	9	10	**avg.**
WBCD # features	4	6	6	5	6	7	7	7	6	8	**6.2**
WBCD accuracy	96.7	97.35	96.76	97.35	96.91	96.91	97.06	97.06	97.35	97.5	**97.1**
WDBC # features	16	17	16	15	15	18	15	18	17	16	**16.3**
WBCD accuracy	95.0	95.22	94.87	94.69	94.04	94.87	95.58	95.04	95.04	95.58	**95.0**

In this section, we compared the accuracy levels and the number of features between the initial and final hawk populations to understand the impact of the evolutionary process on model performance. As shown in Figures 4 and 5, the initial population exhibited a wide range of accuracy levels and feature counts, indicating significant variability in the initial set of models. In contrast, the final population demonstrated a narrower range of accuracy levels with a higher, but more consistent feature count, around 7-8 features. This suggests that the algorithm effectively selected hawks with an optimal performance leading to improved consistency and potentially higher accuracy in the final population.

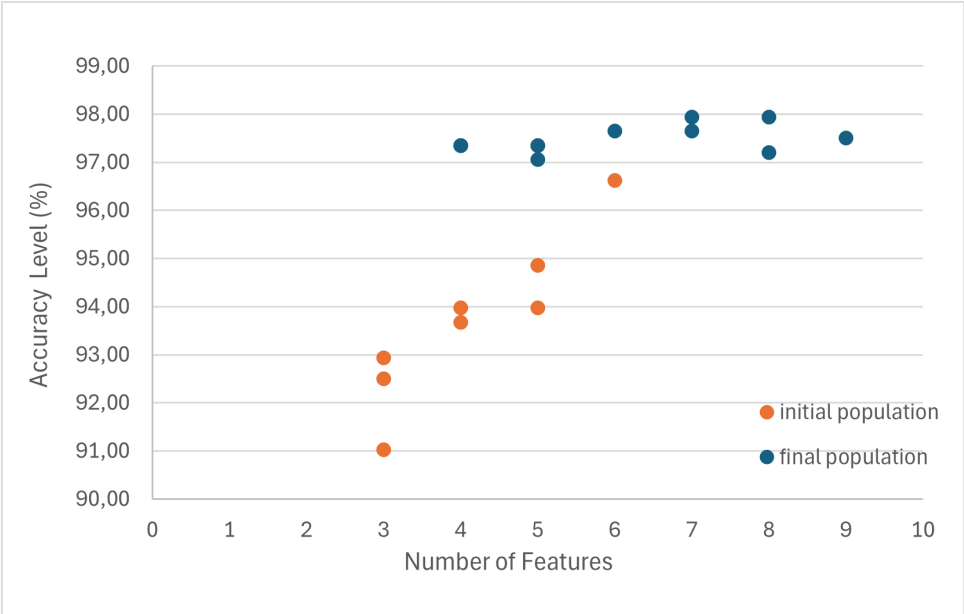

Figure 4. The initial and evolving populations of the proposed algorithm for the WBCD dataset through the generations.

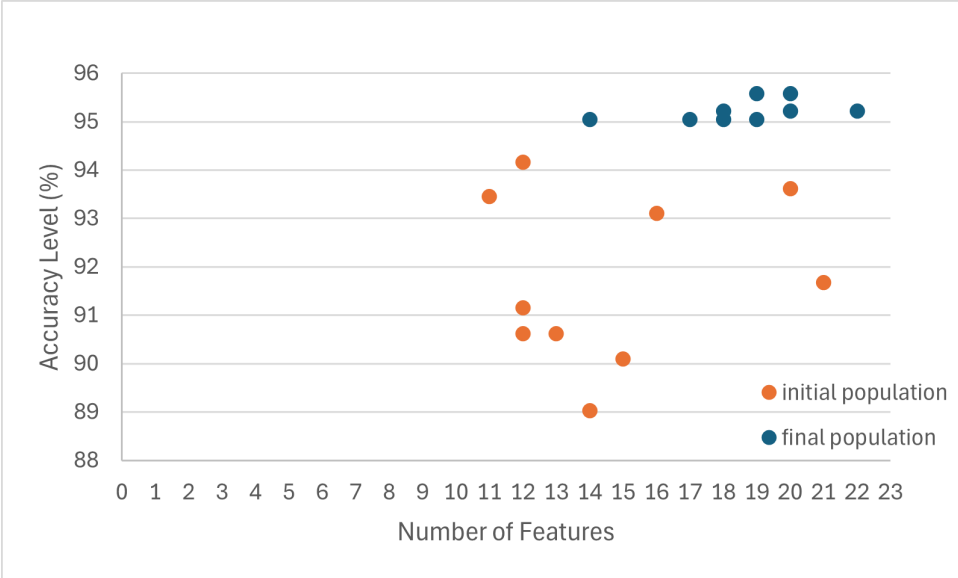

Figure 5. The initial and evolving populations of the proposed algorithm for the WDBC dataset through the generations.

4.Related Works

Several algorithms have been explored for breast cancer classification using the WDBC and WBCD datasets. In 2011, Doddipalli et al. reported high accuracy rates with decision tree classifiers like CART, achieving 96.99% accuracy on the WBCD dataset and 94.72% on the WDBC dataset [10]. Following this, Salama et al. (2012) demonstrated promising results with ensemble methods, achieving 97.28% accuracy on the WBCD dataset by fusing SMO, IBK, NB, and J48 classifiers, while individually, SMO achieved 97.72% on the WDBC dataset [11]. Furthermore, Aalaei et al. (2016) investigated the impact of GA-based feature selection, achieving 97.3% accuracy on the WDBC dataset with ANN and 96.9% on the WBCD dataset with PS-classifier [12]. Optimization techniques have also been explored, such as the adjusted Bat Algorithm (ABA) used by Tuba et al. (2016) to optimize SVM parameters, yielding accuracies of 96.99% on the WBCD dataset and 96.49% on the WDBC dataset [13]. Hybrid models,

like the HECFNN proposed by Alkhasawneh et al. (2018), combining CFNN and ENN, achieved 97.7% accuracy on the WBCD dataset [14].

Notable studies also include Mushtaq et al. (2019), who achieved 99.42% accuracy with KNN on the WBC dataset using Chi-square-based feature selection and the Manhattan distance function [15]. Singh et al. (2020) explored hybrid optimization approaches, combining GWO and WOA for SVM hyperparameter tuning, achieving 97.72% accuracy [16]. Wang et al. (2020) developed the IRFRE method, integrating Random Forest-based rule extraction with a multi-objective evolutionary algorithm, achieving 96.44% ± 3.76% accuracy on the WBCD dataset [17]. More recently, Badr et al. (2022) examined a hybrid GWO-SVM model, achieving 98.60% accuracy on the WDBC dataset with normalization scaling and 99.30% with their proposed scaling techniques [18]. These studies collectively demonstrate the potential of machine learning techniques in accurately classifying breast cancer, providing a foundation for further research and development in this area.
These studies demonstrate the potential of machine learning techniques in accurately classifying breast cancer, providing a foundation for further research and development in this area. The accuracy levels for WBCD and WDBC datasets for each study are summarized in Tables 3 and 4, respectively. It is important to note that, other than CART and the proposed MHHO algorithm, none of the reviewed studies explicitly reported the number of features used in their models.

Table 3. Comparison of the performance of the breast cancer classification methods on the WBCD Dataset

WBCD dataset			
Method	Feature #	Accuracy	Year
CART [10]	9	96.99	2011
SMO, IBK, NB, and J48 [11]	-	97.28	2012
ANN with FS [12]	-	96.90	2016
ABA-SVM [13]	-	96.99	2016
CFNN [14]	-	97.70	2018
KNN [15]	-	99.42	2019
IFRE[17]	-	96.44	2020
MHHO (our model)	**8**	**97.50**	**2025**

Table 4. Comparison of the performance of the breast cancer classification methods on the WDBC Dataset

WBCD dataset			
Method	number of features	Accuracy	Year
CART [10]	8	94.72	2011
SMO [11]	-	97.72	2012
PS [12]	-	97.30	2016
ABA-SVM [13]	-	96.49	2016
GWO-SVM [16]	-	99.30	2020
GWWOA-SVM [18]	-	97.72	2022
MHHO (our model)	**15**	**95.58**	**2025**

5. Conclusion and Future Work

Our study introduces a novel application of the HHO metaheuristic algorithm for multi-objective feature selection, integrated with an adaptive KNN classifier, for breast cancer diagnosis. Extensive experiments demonstrate superior accuracy in datasets and achieve a 31-45% reduction in the number of features significantly lowering computational costs. This feature reduction is accompanied by improved accuracy, confirming the efficiency of our approach. Overall, the proposed HHO algorithm provides a practical and effective solution for feature selection and classification in breast cancer diagnosis, offering promising results for future research. In our future work, we intend to study parallel versions of our proposed algorithm for the multiobjective HHO algorithm on GPU architectures.

Contribution of Researchers

All authors contributed equally to the writing of this article.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Ethics committee approval (if needed)

No need to ethics committee approval statement.

References

[1] Liu, Q., Li, X., Liu, H., & Guo, Z. (2020). Multi-objective metaheuristics for discrete optimization problems: A review of the state-of-the-art. Applied Soft Computing, 93, 106382.

[2] Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., & Chen, H. (2019). Harris Hawks optimization: Algorithm and applications. Future generation computer systems, 97, 849-872.

[3] Dokeroglu, T., Deniz, A., & Kiziloz, H. E. (2021). A robust multiobjective Harris' Hawks Optimization algorithm for the binary classification problem. Knowledge-Based Systems, 227, 107219.

[4] Piri, J., & Mohapatra, P. (2021). An analytical study of modified multi-objective Harris Hawk Optimizer towards medical data feature selection. Computers in Biology and Medicine, 135, 104558.

[5] Selim, A., Kamel, S., Alghamdi, A. S., & Jurado, F. (2020). Optimal placement of DGs in distribution system using an improved harris hawks optimizer based on single-and multi-objective approaches. IEEE Access, 8, 52815-52829.

[6] Thawkar, S. (2022). Feature selection and classification in mammography using hybrid crow search algorithm with Harris Hawks optimization. Biocybernetics and Biomedical Engineering, 42(4), 1094-1111.

[7] Bandyopadhyay, R., Basu, A., Cuevas, E., & Sarkar, R. (2021). Harris Hawks optimisation with Simulated Annealing as a deep feature selection method for screening of COVID-19 CT-scans. Applied Soft Computing, 111, 107698.

[8] Wolberg, W. (1990). Breast Cancer Wisconsin (Original) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z.

[9] Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[10] Doddipalli, Lavanya & Rani, K.. (2011). Analysis of feature selection with classification: Breast cancer datasets. Indian Journal of Computer Science and Engineering (IJCSE). 2. 756-763.

[11] Salama, G. I., Abdelhalim, M. B., & Zeid, M. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technology, 1*(1), 1-10.

[12] Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iran J Basic Med Sci. 2016 May;19(5):476-82. PMID: 27403253; PMCID: PMC4923467.

[13] E. Tuba, M. Tuba and D. Simian, "Adjusted bat algorithm for tuning of support vector machine parameters," 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, Canada, 2016, pp. 2225-2232, doi: 10.1109/CEC.2016.7744063.

[14] Alkhasawneh, M. S., & Tay, L. T. (2018). A Hybrid Intelligent System Integrating the Cascade Forward Neural Network with Elman Neural Network. Arabian Journal for Science and Engineering, 43(12), 6737–6749. https://doi.org/10.1007/s13369-017- 2833-3

[15] Mushtaq, Z., Yaqub, A., Sani, S., & Khalid, A. (2019). Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. Journal of the Chinese Institute of Engineers, 43(1), 80–92. https://doi.org/10.1080/02533839.2019.1676658

[16] I. Singh, R. Bansal, A. Gupta and A. Singh, "A Hybrid Grey Wolf-Whale Optimization Algorithm for Optimizing SVM in Breast Cancer Diagnosis," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2020, pp. 286-290, doi: 10.1109/PDGC50313.2020.9315816.

[17] Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., & Jin, Y. (2020). An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing, 86, 105941. https://doi.org/10.1016/j.asoc.2019.105941

[18] Badr, E., Almotairi, S., Abdul Salam, M., & Ahmed, H. (2022). New sequential and parallel support vector machine with grey wolf optimizer for breast cancer diagnosis. Alexandria Engineering Journal, 61(3), 2520-2534. https://doi.org/10.1016/j.aej.2021.07.024