Applying Various Machine Learning Techniques for Early Diagnosis of Breast Cancer

Mohamed Shaban Abden
Faculty of Computer Science,
Nahda University in Beni Suef, Egypt
Faculty of Computers and Artificial Intelligence,
Fayoum University, Egypt

Mostafa Ali Elmasry
Faculty of Computers and Artificial Intelligence,
Fayoum University, Egypt

Kamel Hussein Rahouma
Faculty of Computer Science,
Nahda University in Beni Suef, Egypt
Faculty of Engineering,
Minia University, Minia, Egypt

ABSTRACT
Cancer disease is a category of diseases distinguished as an uncontrolled increase and extension of unnatural cells within the body, often caused by genetic mutations and various risk factors. Breast cancer (BC) stands as a common cancer forms. The early detection through timely examination and treatment greatly improves the chances of a successful outcome. To enhance early detection and improve treatment outcomes, a gene expression data set was used, but the curse of dimensionality appears when trying to analyze such data. We aim to create an accurate model. So, it is important to filter this noise and lower the dimensions in the microarray data, which is considered a mandatory step. In this study, we conducted experiments for the early identification of breast cancer. For this task, we used breast cancer microarray data to classify patients. First, the dataset was normalized using the min-max scalar technique, and then its features were obtained using Binary Harris Hawks Optimization (BHHO). The application of machine learning models like k-nearest neighbor (KNN), support vector machine (SVM), logistic regression (LR), decision tree (DT), and neural network (NN) are investigated. Our experiments show that DT outperformed the other models producing the highest performance across Van't Veer dataset.

Keywords
Breast cancer classification. Microarray data. Binary Harris Hawks Optimization (BHHO)

1. INTRODUCTION
Breast cancer is considered the second-most dangerous health problem in the world, as it is defined as the uncontrolled increase of unnatural cells. These cells are called malignant, tumor, or cancerous cells, as they affect normal body tissues [1]. Carcinoma arises from a sequence of genetic mutations that disrupt the normal cellular growth, causing the affected cells to expand increase, divide, and ultimately form carcinoma [2, 3]. Typically, it originates within the lobules or the inner lining of the milk ducts, which are responsible for supplying milk [4, 5]. The symptoms include skin swelling, surface peeling, skin irritation, a change in skin color resembling orange, discharge from the nipple, and the nipple retracting upon touch. Recently, it has been noted that the death rate of women with breast cancer has increased, with about 10%–15% of breast carcinoma patients suffering from breast pain and also it has been found that breast cancer is spreading between men [6]. Early diagnosis and detection have shown to greatly increase patient survival rates [7], while simultaneously reducing the cost and intricacy associated with cancer treatment. The microarray method has emerged as a standard approach for early breast carcinoma detection [8], However, microarray data often consists of numerous genes with limited samples, a considerable part of which are either noisy, redundant, or irrelevant [9].

Although machine learning (ML) has displayed tremendous potential in the field of bioinformatics. However, there is a challenge known as the curse of dimensionality when handling this specific data type. To create an accurate model, it is crucial to reduce noise and lower the dimensions in microarray data. To do this, feature selection (FS) approaches are essential. They help reduce the dimensionality of high-dimensional datasets like microarray data and enhance the accuracy and generalizability of machine learning models. FS techniques can improve model performance, prevent overfitting, and speed up training by selecting relevant features [10]. There are different types of FS techniques, including filter, hybrid, wrapper, and embedding approaches, each with its strengths and weaknesses. Filter methods use statistical or information-theoretic measures to select features independently from the learning algorithm. Wrapper methods, on the other hand, evaluate subsets of features using the learning algorithm to maximize performance. Hybrid methods combine the strengths of filter and wrapper techniques, while embedded methods incorporate FS into the learning algorithm itself. The choice of an FS technique depends on the specific characteristics of the data and the machine learning model being used. Sometimes, a simple filter method is sufficient, but in other cases, a more advanced wrapper or embedded approach may be necessary. [11].

This paper has five sections. Section (1) gives an introduction and section (2) presents a review of some of the used literature. Section (3) explains the proposed methodology and section (4) highlights as overview of the dataset used and presents the experimental results, which are discussed and compared with the previous research results. Section (5) illustrates some conclusions and introduces some points for the future research work. At the end of the paper, some of the used references are listed.

2. A LITERATURE REVIEW
In [12], the authors introduced a novel method aimed at addressing the challenges posed by high-dimensional microarray datasets. Their approach involves a combination of filter feature selection techniques, specifically information gain, gain ratio, and Chi-squared, in conjunction with a genetic algorithm (GA). This strategy entails two distinct phases: firstly, employing filter feature selection methods to identify and retain the most informative 5% of features while eliminating those that are redundant or irrelevant, thus reducing

the dataset's dimensionality; secondly, utilizing a genetic algorithm (GA) to further refine the reduced datasets, resulting in enhanced cancer classification outcomes. The outcomes of this hybrid approach demonstrated a notable reduction of approximately 50% in irrelevant features and a significant enhancement in cancer classification performance compared to using standalone classifiers or filter algorithms in isolation. When applied to the breast cancer dataset, the accuracy rates achieved were as follows: SVM 82.47%, NB 79.38%, KNN 84.54%, DT 90.72%, and RF 93.81%.

In [13], the authors proposed a hybrid strategy combining a genetic algorithm (GA), LASSO, and mutual information (MI) for early cancer diagnosis. LR, SVM, KNN, and RF were used to evaluate this approach. This approach can accurately diagnose breast carcinoma with fewer genes than state-of-the-art models. With just 23 features, it was able to achieve a classification accuracy of 96.04% for the benchmarked Van't Veer dataset.

In [14], the authors introduced a novel hybrid model designed to distinguish between benign and malignant breast cancer samples by integrating two optimization strategies with machine learning. Initially, K-Means was applied to address the complexities arising from nonlinear and imbalanced data distribution by assigning feature weights. Subsequently, a novel method known as PHHO, which combines Harris Hawks Optimization with PSO, was utilized to optimize the extreme learning machine. To evaluate the efficacy of their proposed model, the authors conducted tests on the Wisconsin diagnosis breast carcinoma dataset, yielding impressive results. The model demonstrated exceptional performance with high accuracy, sensitivity, and specificity, achieving values of 98.76%, 97.37%, and 99.46%, respectively.

In [15], the authors created a hybrid strategy consisting of enhanced GA and SLI. The suggested SLI filter strategy was made to rank and order features based on how effective they are. The best feature subset was then found by using GA to identify the features at the upper of the feature classification. Specifically, the KNN and ANN classifiers were used to determine the fitness value. To assess the performance of this strategy, eleven Wisconsin datasets and high-dimensional microarray datasets have used. In their research, KNN and ANN classifiers were used. Utilizing 10-fold cross-validation with and without FS techniques, the suggested strategy was evaluated 100 times. The suggested technique showed outstanding classification accuracy using hybrid FS with a KNN classifier compared to other current techniques with a limited number of chosen genes.

In [16], a hybrid feature selection (FS) approach named pyramid (PGSA) that combines an enhanced binary gravitational search algorithm (BGSA) with gene ranking. The primary objective of this approach was to reduce the number of genes in the microarray data. Initially, gene ranking was employed as a filtering strategy to choose the most relevant genes. Subsequently, an improved binary GSA (IBGSA) was utilized as the wrapper approach to select the optimal gene subset. The PGSA method collaborates with the classifier in each gene selection cycle, thereby enhancing the overall accuracy. To assess the effectiveness of the suggested technique, a 10-fold cross-validation was performed, and a support vector machine was employed to obtain the fitness value. The results demonstrated that the suggested method outperformed existing wrapper methods, successfully eliminating more than 70% of the features from the original feature set.

In [17], A combination of multi-objective particle swarm optimization (PSO) and signal-to-noise ratio (SNR) was utilized for microarray data identification. The primary objective of the proposed technique, which incorporated the SNR filter method, was to eliminate insignificant attributes and effectively rank the top 100 features. The results produced by the filter method were subsequently integrated into the multi-objective PSO approach, which sought to maximize accuracy while minimizing the number of selected features. An adaptive K-nearest neighbors (KNN) classifier was then applied to the chosen genes, and its performance was assessed across 10 iterations. The outcomes revealed that the suggested approach exhibited outstanding classification accuracy, achieving a perfect accuracy rate of 100% on two datasets.

In [18], the authors presented a hybrid approach, combining the information gain (IG) filter method with a genetic algorithm (GA) wrapper technique, along with the fuzzy logic neural network (FLNN) classifier, for breast cancer (BC) data prediction. The GA served as a wrapper approach, while the IG was employed as a filter method to evaluate gene importance. During the classification phase, FLNN was utilized to assess the gene selection obtained from the wrapper technique. The suggested technique was evaluated on the Microarray BC dataset. In the classification stage, two different learning rate (LR) parameter values, namely 0.01 and 0.6, were employed in different orders to explore their impact. The results revealed that BC achieved better outcomes with an LR of 0.01, exhibiting an accuracy of 85.63%. Furthermore, the authors extended their proposed technique to classify four other cancer types (lung, colon, prostate, and ovarian), achieving an accuracy above 90% for disease classification.

In [19], the authors directed several studies utilizing machine-learning algorithms to get better the datasets classification of breast tumor. It was shown that logistic regression produces accurate results when applied to the training set. The confusion matrix was depicted, and the accuracy was evaluated using seaborn and sklearn metrics. The accuracy of this model is 97.63%. However, the data set may contain incremented data, which can increase accuracy.

In [20], the authors aimed to enhance the classification results of the JELM for BC classification. They utilized Jaya optimization to select the optimal hidden biases and input weights for the ELM. Additionally, the authors employed the WRST to identify relevant genes for the classification task. Comparing the performance of JELM with other classifiers such as SVM, KNN, NB, and c4.5, JELM achieved a higher accuracy rate of approximately 90.91%. However, despite the promising accuracy, the suggested model selected a large subset of around 505 genes, indicating the need for further reduction of the gene subset.

2.1 Discussion

Although the use of microarray data has proven to be effective in diagnosing breast cancer, there is a challenge known as the 'curse of dimensions' due to the numerous features and limited sample size. For instance, the Van't Veer dataset[13] contains 24,481 features with only 97 samples. To address this challenge, hyper- and filter-selection techniques have been utilized. The Van't Veer dataset[13] has been employed in studies [12],[14],[20],and[21]. In [20], the lowest results were obtained using the IG-GA selection approach, which resulted in a subset of 49 genes. Conversely, [21] achieved a high score of 90.91% using the Wilcoxon rank-sum test (WRST) but with a larger subset of genes (505). In [14],the highest results were obtained using the MI-LASSO-GA selection approach, which

resulted in the lowest number of genes (23). In our comparison with [14], we used the same dataset and classifiers but employed the feature selection method BHHO, resulting in higher results in most of the classifiers.

3. THE PROPOSED METHODOLOGY

This section provides a description of the proposed Binary Harris-Hawks optimization algorithm-based feature selection approach used for breast cancer classification in high-dimensional microarray gene datasets, as shown in figure (1). In this figure, the dataset is prepared and normalized. Then, the BHHO technique is applied for the best gene subset selection. Training and testing of the model are applied and then the model performance is evaluated. In the following subsections, these steps are presented.

3.1 Normalization

Normalization of the dataset is essential to ensure that each numeric value is placed on a uniform scale, preventing certain features from exerting undue influence due to their higher values [22]. To achieve this, the min-max normalization technique is employed. Initially, the minimum value within the dataset is subtracted from each feature, and subsequently, the outcome is divided by the range between the maximum and minimum values. This normalization method ensures that all features are scaled within the 0 to 1 scale, as expressed by the following equation (1) [23] where X stands for the feature's value

3.2 Features Selection

In the processing of microarray data, it is crucial to remove duplicate data and choose only relative characteristics. One of the methods for selecting features based on classifiers is wrapper. It works by selecting a subset of genes that yield the best results for a specific learning model. There are two categories of wrapper methods: stochastic and greedy search. Because they employ a single-track search, greedy search techniques like forward selection and sequential backward selection can become stuck in a local optimum. On the other hand, stochastic search techniques make use of randomness to investigate the solution space and can make use of meta-heuristic algorithms to enhance the selection procedure[24, 25].

Binary Harris Hawks Optimization (BHHO) is a meta-heuristic optimization algorithm employed for genes chosen in order to attain excellent classification accuracy while utilizing a limited number of genes. BHHO incorporates two fitness functions to strike a balance between classification performance and the number of chosen genes. The first fitness function employs a weight that progressively increases during the optimization process to effectively manage the two objectives. The second fitness function operates in two stages, with the initial stage solely focusing on optimizing classification performance, while the second stage takes into account the number of chosen genes. Ultimately, BHHO aims to identify a reduced set of genes while simultaneously achieving high classification accuracy. [26].

3.3 Evaluation

DT, SVM, LR, KNN, and NN classifiers have all been used to evaluate the suggested method. The supervised classification algorithm known as K-nearest neighbors (KNN). It finds a label for a new point by looking at the closest labeled (K) points [27]. Similar measures, including Euclidean distance, Manhattan distance, Hamming distance, and Makowski distance, are used to determine the KNN rankings [28].

Decision tree (DT) can be represented microarray data as trees. To do this, instances are organized as nodes within the tree structure. In this representation, the decision nodes have two branches, while the leaf nodes contain a single decision. [29].

Logistic regression (LR) is a classification method employed to categorize observations into distinct classes. It accomplishes this by utilizing the logistic sigmoid function to convert its output into a probability value. LR is known for its simplicity and effectiveness in classification tasks.[30].

Neural network (NN) is a type of technique that uses a simulation of the human brain to find patterns in data. Without having to change the output criteria, it can adjust to changing input and produce an optimal result [31].

4. EXPERIMENTAL RESULTS

The suggested BHHO is used with five BC datasets for gene expression and KNN, SVM, LR, and NN classifiers. The BC datasets description is summarized in Table (2). The classifiers results, according to these data sets, are described in Table (3).

Multiple BC datasets are used to evaluate the suggested approach utilizing BHHO. Comparing our results with the previous research, we can find:

1) In [14], the authors used Van't Veer [13] with the classifier MI-GA, and they could reach an accuracy level of 81.32 %with a precision level of 0.80 and a number of features 366 using the logistic regression algorithm (LR). A following step was using a hybrid classifier of MI-LASSO-GA. They could reach an accuracy level of 96.04% with a precision level of.989 and a number of features 23 using (LR). In our study, we have reached either same results or better using the same classifiers. For instance, we obtained an accuracy of 96.67% in LR, 96.67% in KNN, 93.33% in SVM, 100 %in DT and 93.33% in NN where the number of features is 31. Figure (3) shows the performance in each of them.

2) In [14], the authors used Chowdary [32] with the classifier MI-GA, and they could reach an accuracy level of 98.09 % with a precision level of 0.98 and a number of features 338 using (LR). A following step was using a hybrid classifier of MI-LASSO-GA. They could reach an accuracy level of 99.27% with a precision level of 0.98 and a number of features 12 using (LR). In our study, we have reached either same results or better using same classifiers. For instance, we obtained an accuracy of 95% in LR, 100% in KNN, 95.23% in SVM, 80.95 % in DT and 85.71% in NN where the number of features is 15. Figure (4) shows the performance in each of them.

3) In [14], the authors used Chin [33] with the classifier MI-GA, and they could reach an accuracy level of 90.69 % with a precision level of 0.896 and a number of features 337 using (LR). A following step was using a hybrid classifier of MI-LASSO-GA. They could reach an accuracy level of 95.50% with a precision level of 0.94 and a number of features 18 using (LR). In our study, we have reached either same results or better using the same classifiers. For instance, we obtained an accuracy of 100% in LR, 100% in KNN, 91.67% in SVM, 83.3% in DT and 95.83% in NN where the number of features is 23. Figure (5) shows the performance in each of them.

4) In [14], the authors used Gravier [34] with the classifier MI-GA, and they could reach an accuracy level of 81.24 % with a precision level of 0.85 and number of features 46 using (LR). A following step was using a hybrid classifier of MI-LASSO-GA, and they could reach an accuracy level of 86.73% with a precision level of.95 and number of features 13 using (LR). In our study, we have reached either same results or better using

the same classifiers. For instance, we obtained an accuracy of 73.53% in LR, 88.24% in KNN, 79.41% in SVM, 67.65% in DT and 76.47% in NN where the number of features is 13. Figure (6) shows the performance in each of them.

5) In [14], the authors used West [35] with the classifier MI-GA, and they could reach an accuracy level of 94.09% with a precision level of 0.95 and a number of features 108 using (LR). A following step was using a hybrid classifier of MI-LASSO-GA, and they could reach an accuracy level of 100% with a precision level of 1.00 and a number of features 17 using (LR). In our study, we have reached either same results or better using the same classifiers. For instance, we obtained an accuracy of 80% in LR, 90% in KNN, 70% in SVM, 70 % in DT and 90% in NN while the number of features is 5. Figure (7) shows the performance in each of them.

5. CONCLUSIONS

Early identification of breast cancer is essential to preventing suffering for patients. Microarray technology can be used to achieve this goal. However, the classification process is complicated by the large number of its features. Hence, the Binary Harris Hawks Optimization (BHHO) approach is used to select a small number of genes and achieve a high level of classification accuracy. The proposed approach was used on five breast carcinoma datasets and assessed by the following classifiers: KNN, SVM, LR, DT, and NN. It obtained high accuracy for all datasets, with the West dataset having an accuracy of 90%, the Van't Veer dataset having an accuracy of 100%, and the Gravier dataset having an accuracy of 88%. and the Chin dataset has an accuracy of 100%. and the Chowdary dataset having an accuracy of 100%, the suggested model will be assessed for other types of cancer.

6. REFERENCES

[1] C.-H. Lee, W.-H. Kuo, C.-C. Lin, Y.-J. Oyang, H.-C. Huang, and H.-F. Juan, "MicroRNA-regulated protein-protein interaction networks and their functions in breast cancer," *International journal of molecular sciences*, vol. 14, no. 6, pp. 11560-11606, 2013.

[2] E. v. d. Akker *et al.*, "Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis," *Journal of Integrative Bioinformatics*, vol. 8, no. 2, pp. 222-238, 2011.

[3] A. Chakraborty *et al.*, "Determining protein–protein interaction using support vector machine: A review," *IEEE Access*, vol. 9, pp. 12473-12490, 2021.

[4] Sarkar, J. P., Saha, I., Rakshit, S., Pal, M., Wlasnowolski, M., Sarkar, A., ... & Plewczynski, D. (2019, July). A new evolutionary rough fuzzy integrated machine learning technique for microRNA selection using next-generation sequencing data of breast cancer. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 1846-1854).

[5] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, and Q. Ma, "Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique," *Bioinformatics*, vol. 35, no. 14, pp. 2395-2402, 2019.

[6] Y.-B. Wang *et al.*, "Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network," *Molecular BioSystems*, vol. 13, no. 7, pp. 1336-1344, 2017.

[7] R. Sheikhpour, M. A. Sarram, and R. Sheikhpour, "Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer," *Applied Soft Computing*, vol. 40, pp. 113-131, 2016.

[8] C. L. Chowdhary, N. Khare, H. Patel, S. Koppu, R. Kaluri, and D. S. Rajput, "Past, present and future of gene feature selection for breast cancer classification–a survey," *International Journal of Engineering Systems Modelling and Simulation*, vol. 13, no. 2, pp. 140-153, 2022.

[9] J. Pirgazi, M. Alimoradi, T. Esmaeili Abharian, and M. H. Olyaee, "An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets," *Scientific reports*, vol. 9, no. 1, p. 18580, 2019.

[10] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Information sciences*, vol. 282, pp. 111-135, 2014.

[11] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection--a comparative study," *Lecture notes in computer science*, vol. 4881, pp. 178-187, 2007.

[12] W. Ali and F. Saeed, "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data," *Processes*, vol. 11, no. 2, p. 562, 2023.

[13] L. J. Van't Veer *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530-536, 2002.

[14] M. Abd-elnaby, M. Alfonse, and M. Roushdy, "A Hybrid Mutual Information-LASSO-Genetic Algorithm Selection Approach for Classifying Breast Cancer," in *Digital Transformation Technology: Proceedings of ITAF 2020*, 2022: Springer, pp. 547-560.

[15] F. Jiang, Q. Zhu, and T. Tian, "Breast Cancer Detection Based on Modified Harris Hawks Optimization and Extreme Learning Machine Embedded with Feature Weighting," *Neural Processing Letters*, pp. 1-24, 2022.

[16] A. Tahmouresi, E. Rashedi, M. M. Yaghoobi, and M. Rezaei, "Gene selection using pyramid gravitational search algorithm," *Plos one*, vol. 17, no. 3, p. e0265351, 2022.

[17] K.-J. Kao, K.-M. Chang, H.-C. Hsu, and A. T. Huang, "Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization," *BMC cancer*, vol. 11, no. 1, pp. 1-15, 2011.

[18] S. Abasabadi, H. Nematzadeh, H. Motameni, and E. Akbari, "Hybrid feature selection based on SLI and genetic algorithm for microarray datasets," *The Journal of Supercomputing*, pp. 1-29, 2022.

[19] Kowsari, Y., Nakhodchi, S., & Gholamiangonabadi, D. (2022). Gene selection from microarray expression data: A Multi-objective PSO with adaptive K-nearest neighborhood. *arXiv preprint arXiv:2205.15020*.

[20] G. G. Afif and W. Astuti, "Cancer Detection based on Microarray Data Classification Using FLNN and Hybrid Feature Selection," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 4, pp. 794-801, 2021.

[21] S. K. Baliarsingh, C. Dora, and S. Vipsita, "Jaya optimized extreme learning machine for breast cancer data classification," in *Intelligent and Cloud Computing*: Springer, 2021, pp. 459-467.

[22] J. M. Moosa, R. Shakur, M. Kaykobad, and M. S. Rahman, "Gene selection for cancer classification with the help of bees," *BMC medical genomics*, vol. 9, pp. 135-165, 2016.

[23] R.-J. Palma-Mendoza, D. Rodriguez, and L. De-Marcos, "Distributed ReliefF-based feature selection in Spark," *Knowledge and Information Systems*, vol. 57, pp. 1-20, 2018.

[24] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[25] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern recognition*, vol. 43, no. 1, pp. 5-13, 2010.

[26] Khurma, R. A., Castillo, P. A., Sharieh, A., & Aljarah, I. (2020). New Fitness Functions in Binary Harris Hawks Optimization for Gene Selection in Microarray Datasets. In *IJCCI* (pp. 139-146).

[27] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, pp. 37-66, 1991.

[28] Mohapatra, P., & Chakravarty, S. (2015, October). Modified PSO based feature selection for Microarray data classification. In *2015 IEEE Power, Communication and Information Technology Conference (PCITC)* (pp. 703-709). IEEE.

[29] Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.

[30] T. Rymarczyk, E. Kozłowski, G. Kłosowski, and K. Niderla, "Logistic regression for machine learning in process tomography," *Sensors*, vol. 19, no. 15, p. 3400, 2019.

[31] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.

[32] D. Chowdary *et al.*, "Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative," *The journal of molecular diagnostics*, vol. 8, no. 1, pp. 31-39, 2006.

[33] K. Chin *et al.*, "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies," *Cancer cell*, vol. 10, no. 6, pp. 529-541, 2006.

[34] E. Gravier *et al.*, "A prognostic DNA signature for T1T2 node-negative breast cancer patients," *Genes, chromosomes and cancer*, vol. 49, no. 12, pp. 1125-1134, 2010.

[35] M. West *et al.*, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11462-11467, 2001.

7. APPENDIX

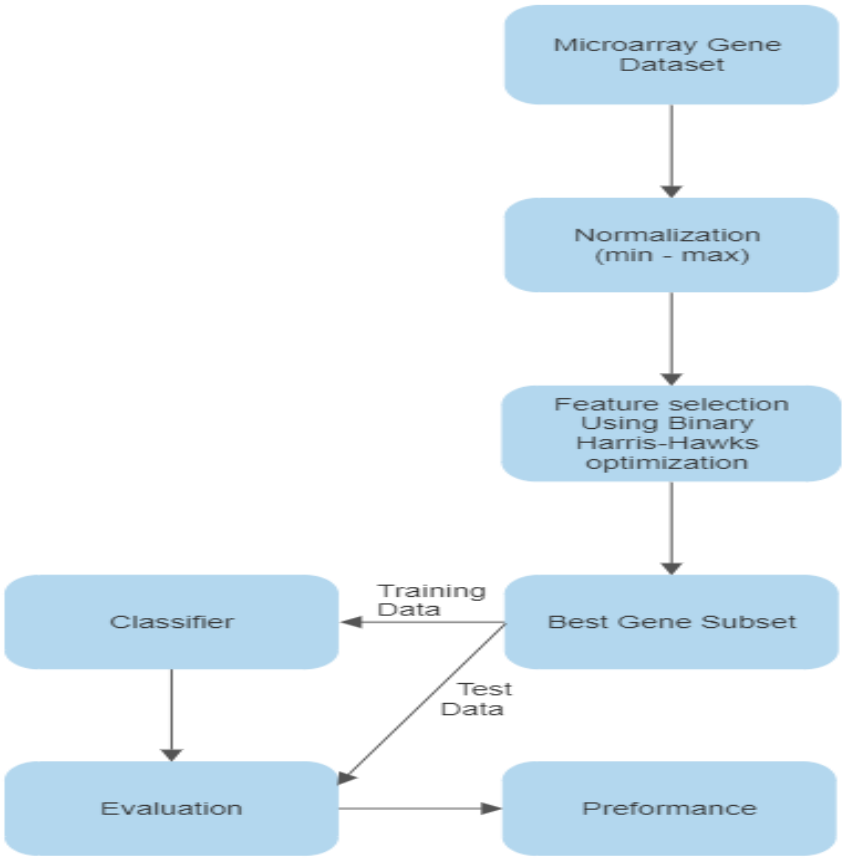

Figure (1): The methodology of the Binary Harris-Hawks optimization algorithm-based feature selection model proposed for breast cancer classification in high-dimensional microarray gene datasets

Table (1): Comparing the previous research works based on feature selection, classifiers, datasets, and number of genes.

Ref	Feature selection	Classifier	Dataset	Classification accuracy	No. of Genes
Ali and Saeed (2023) [12]	Hybrid Filter _ GA	SVM NB KNN DT RF	Breast [13]	84.54 79.38 86.6 90.72 93.81	NAN
Abd-elnaby et al. (2022) [14]	MI-LASSO-GA	LR SVM RF KNN	Breast [13]	96.04 86.02 95.082 87.81	23
Jiang et al. (2022) [15]	Hybrid approach (PSO – HHO)	ELM	Breast	98.76	NAN
Tahmouresi et al (2022) [16]	Gene Rank GSA	SVM	Breast [17]	84.5%	73
Abasabadi et al. (2022) [18]	SLI GA	KNN	DLBCL CNS Colon Leukemia	100% 99.97% 100% 99.99%	22 29 11 29
Kowsari et al.(2022) [19]	SNR PSO	KNN	DLBCL Leukemia Prostate CML Colon	99.47% 94.62% 100% 100% 98.81%	NAN
Afif and Astuti (2021) [20]	IG GA	FLNN	Breast [13]	85.63	49
Baliarsingh et al. (2021) [21]	WRST	JELM	Breast [13]	90.91%	505

Table (2): Description of the BC datasets

Dataset	No. of samples	No. of features	No. of classes
Van't veer [13]	97	24,481	2
Chowdary [32]	104	22,283	2
Chin [33]	118	22,215	2
Gravier [34]	168	2905	2
West [35]	49	7129	2

Table (3): Classifiers results (KNN, SVM, LR, DT, and NN) according to the used datasets

Ref.	Data-set	Classifiers	Measure	KNN	SVM	LR	DT	NN
In [14]	Van' t veer [13]	MI-GA	Accuracy	76.35	80.02	81.32	-	-
			Precision	0.78	0.81	0.80	-	-
			Recall	0.77	0.77	0.80	-	-
			No. features	**366**				
		MI-LASSO-GA	Accuracy	87.81	86.02	96.04	-	-
			Precision	0.98	0.87	0.989	-	-
			Recall	0.918	0.83	0.93	-	-
			No. features	**23**				
Our study		BHHO	Accuracy	96.67	93.33	96.67	100	93.33
			Precision	0.97	0.94	0.97	1.00	0.93
			Recall	0.96	0.93	0.96	1.00	0.93
			No. features	**31**				
In [14]	Chowdary [32]	MI-GA	Accuracy	96.31	96.18	98.09	-	-
			Precision	0.98	0.98	0.98	-	-
			Recall	0.98	0.93	0.975	-	-
			No. features	**338**				
		MI-LASSO-GA	Accuracy	98.14	96.35	99.27	-	-
			Precision	0.96	0.98	0.98	-	-
			Recall	0.99	0.93	1.00	-	-
			No. features	**12**				
Our study		BHHO	Accuracy	100.0	95.23	95.24	80.95	85.71
			Precision	1.00	.94	0.96	0.80	0.86
			Recall	1.00	.96	0.94	0.80	0.88
			No. features	**15**				
In [14]	Chin [33]	MI-GA	Accuracy	88.67	89.18	90.69	-	-
			Precision	0.88	0.89	0.896	-	-
			Recall	0.94	0.96	0.97	-	-
			No. features	**337**				
		MI-LASSO-GA	Accuracy	93.79	91.78	95.50	-	-
			Precision	0.93	0.91	0.94	-	-

			Recall	0.98	0.96	0.99	-	-
			No. features	**18**				
Our study		BHHO	Accuracy	100.0	91.67	100.0	83.3	95.83
			Precision	1.00	0.94	1.00	0.83	0.97
			Recall	1.00	0.89	1.00	0.84	0.94
			No. features	**23**				
In [14]	Gravier [34]	MI-GA	Accuracy	71.60	77.55	81.24	-	-
			Precision	0.74	0.75	0.85	-	-
			Recall	0.60	0.54	0.56	-	-
			No. features	**46**				
		MI-LASSO-GA	Accuracy	82.33	80.59	86.73	-	-
			Precision	0.95	0.77	0.96	-	-
			Recall	0.72	0.62	0.64	-	-
			No. features	**13**				
Our study		BHHO	Accuracy	88.24	79.41	73.53	67.65	76.47
			Precision	0.92	0.82	0.85	0.64	0.76
			Recall	0.83	0.73	0.62	0.64	0.70
			No. features	**13**				
In [14]	West [35]	MI-GA	Accuracy	71.58	88.47	94.09	-	-
			Precision	0.95	0.91	0.95	-	-
			Recall	0.88	0.85	0.93	-	-
			No. features	**108**				
		MI-LASSO-GA	Accuracy	92.6	94.76	100.0	-	-
			Precision	0.99	0.96	1.00	-	-
			Recall	0.99	0.93	1.00	-	-
			No. features	**17**				
Our study		BHHO	Accuracy	90.0	70.0	80.0	70.0	90.0
			Precision	1.00	.62	0.80	0.67	1.00
			Recall	.80	1.00	0.80	.80	.80
			No. features	**5**				

List of Abbreviations

DT	Decision Tree
RF	Random Forest
KNN	K-nearest neighbor
SVM	Support Vector Machine
LR	Logistic Regression
NN	Neural Network
BHHO	Binary Harris Hawks Optimization
NB	Naive Bayes
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
MI	Mutual Information
LASSO	least Absolute Shrinkage and Selection Operator
ELM	Extreme Learning Machine
GSA	Gravitational Search Algorithm
SLI	Sorted Label Interference
SNR	Signal to Noise Ratio
IG	Information Gain
FLNN	Functional Link Neural Network
JELM	Jaya Optimized Extreme Learning Machine
WRST	Wilcoxon rank sum test
ML	Machine Learning