BREAST CANCER CLASSIFICATION USING A NOVEL HYBRID FEATURE SELECTION APPROACH

E. Akkur[*], *F. Türk*[†], *O. Eroğul*[*]

Abstract: Many women around the world die due to breast cancer. If breast cancer is treated in the early phase, mortality rates may significantly be reduced. Quite a number of approaches have been proposed to help in the early detection of breast cancer. A novel hybrid feature selection model is suggested in this study. This novel hybrid model aims to build an efficient feature selection method and successfully classify breast lesions. A combination of relief and binary Harris hawk optimization (BHHO) hybrid model is used for feature selection. Then, k-nearest neighbor (k-NN), support vector machine (SVM), logistic regression (LR) and naive Bayes (NB) methods are preferred for the classification task. The suggested hybrid model is tested by three different breast cancer datasets which are Wisconsin diagnostic breast cancer dataset (WDBC), Wisconsin breast cancer dataset (WBCD) and mammographic breast cancer dataset (MBCD). According to the experimental results, the relief and BHHO hybrid model improves the performance of all classification algorithms in all three datasets. For WDBC, relief-BHO-SVM model shows the highest classification rates with an of accuracy of 98.77%, precision of 97.17%, recall of 99.52%, F1-score of 98.33%, specificity of 99.72% and balanced accuracy of 99.62%. For WBCD, relief-BHO-SVM model achieves of accuracy of 99.28%, precision of 98.76%, recall of 99.17%, F1-score of 98.96%, specificity of 99.56% and balanced accuracy of 99.36%. Relief-BHO-SVM model performs the best with an accuracy of 97.44%, precision of 97.41%, recall of 98.26%, F1-score of 97.84%, specificity of 97.47% and balanced accuracy of 97.86% for MBCD. Furthermore, the relief-BHO-SVM model has achieved better results than other known approaches. Compared with recent studies on breast cancer classification, the suggested hybrid method has achieved quite good results.

Key words: *breast cancer, hybrid feature selection, relief, binary Harris hawk optimization, machine learning*

Received: August 11, 2022 **DOI:** 10.14311/NNW.2023.33.005
Revised and accepted: April 30, 2023

[*]Erkan Akkur; Osman Eroğul; Department of Biomedical Engineering, TOBB ETU University, Ankara, Turkey

[†]Fuat Türk – Corresponding author; Department of Computer Engineering, Cankiri Karatekin University, Cankiri, Turkey, E-mail: fuatturk@karatekin.edu.tr

1. Introduction

Breast cancer (BC) is affecting a lot of women around the world. According to the global statistics, BC affected approximately two million women and caused more than half a million deaths in 2020 [1, 2]. Early diagnosis of BC is crucial to minimize the mortality rate. The use of machine learning (ML) algorithms can help to detect BC at early stage [3–7]. When diagnosing breast cancer using a ML algorithm, the high dimensional dataset needs to be analyzed and processed. High dimensional dataset may cause overfitting, increase the training time and affect the performance of ML algorithm. Therefore, it is extremely beneficial to use feature selection approaches to increase the classification rates of ML algorithms. Feature election (FS) approaches are used to reduce the number of input variables by eliminating irrelevant features and narrowing the feature set to those the most relevant to the ML algorithms. FS approaches help build simpler models, make the training process faster and increase the performance of ML algorithms. Thus, prior to classification, FS are used to reduce the size of the feature space and select the most discriminating features [8].

In this study, a novel hybrid FS model is suggested for the diagnosis of BC. The main goal of the suggested hybrid model is to find the most discriminating features. The features determined by the hybrid FS model are used for the classification process of breast masses. Four ML (LR, NB, SVM and k-NN) algorithms are used for classification process, respectively. The effect of presented the hybrid FS model on the classification performance of ML algorithms is also investigated. The summary of this study and its contribution to science is given below:

1. No one has before used a hybrid method based on relief and binary Harris hawk optimization for FS and BC prediction. Therefore, this hybrid FS method can be used for future studies on BC.

2. Four different ML algorithms (SVM, k-NN, NB, and LR) are used. Among four different algorithms, the most optimal ML approach is suggested for the classification of BC.

3. The aforementioned hybrid FS method and ML algorithms are tested on two different BC data sets, which are frequently used in the literature. The suggested hybrid model is also tested a new BC dataset which was confirmed by the Ankara Training and Research Hospital institutional Ethics Committee (319/E-20).

The study is organized as follows: Section 2 summarizes the literature survey on the subject. The methodology of the study is given in Section 3. The search results are shown in Section 4 and Section 5 summarizes the conclusion and future works of the study.

2. Literature survey

Many FS algorithms have been used for the prediction of BC in recent years. Generally, FS algorithms can be examined in two main groups which are filter and

wrapper approaches. Filter approaches use statistical functions to choose and rank the feature subsets. Due to simplicity and efficiency, chi-square (CS), relief, mutual information (MI), maximum relevance minimum redundancy (MRMR) and correlation-based techniques are some of the most used filter FS approaches in the literature [8, 9]. Unlike filter-based approaches, the FS process is based on a specific learning algorithm in wrapper approaches. Nature-inspired metaheuristics (NIM) have been used over the last two decades for wrapper approaches due to their suitability for global search and their ability to escape from the local optima subset. Genetic algorithm (GA), particle swarm optimization (PSO) and grey wolf optimizer (GWO) are some popular ones used as NIM algorithms for BC prediction [10, 11] Sakri *et al.* [12] used PSO with three different ML models for the prediction of BC. NB showed the best accuracy of 81.3%. Chauhan *et al.* [13] used GA with an ensemble method for the diagnosis of BC. Their methods achieved 99.14% of accuracy. Kumar *et al.* [14] used GWO with SVM for the diagnosis of BC. The suggested model of their study achieved 98.24% of accuracy. Harris hawk optimization (HHO) is a new type of NIM used in FS. When compared to other NIM models, HHO shows superior performance for several benchmark datasets [15, 16]. Jiang *et al.* [17] used HHO and extreme learning machine (ELM) on Wisconsin breast cancer dataset. The model achieved 98.76% of accuracy. Alshayeji *et al.* [18] used artificial neural network (ANN) algorithm for BC prediction. The experiments were tested on Wisconsin breast cancer dataset (WBCD) and the Wisconsin diagnostic breast cancer (WDBC) dataset. ANN showed 99.85% of accuracy for WBCD and 99.47% of accuracy for WDBC.

In recent years, several researchers have used hybrid feature selection approaches to exploit the advantages and disadvantages of both approaches, such as filter and wrapper, trying to find a good compromise between efficiency and effectiveness [19–26] The filter approach is mostly used as preprocessing step to decrease the sizes of features and then wrapper FS is utilized to select the optimal feature subset. Some of the hybrid models used in breast cancer prediction are summarized below.

Sangiah *et al.* [20] presented a hybrid model for BC prediction. The hybrid model incorporates relief and entropy-based GA. SVM outperformed with 85.89% of accuracy. Loey *et al.* [21] presented a hybrid model using information gain and GWO for BC and colon cancer prediction. The model achieved 94.87% of accuracy for BC and 95.935% for colon cancer. Alomari *et al* [22] suggested a hybrid FS method using MIR and flower pollination algorithm (FPA) for cancer classification. The model achieved 85.88% of accuracy. Mufassirin *et al.* [23] suggested a hybrid FS approach using IG and wrapper subset evaluator on five different cancer datasets. The best performer accuracy was 89.69%. Alzubaidi *et al.* [24] suggested a hybrid FS technique using GA and MI on WBCD. SVM algorithm showed 0.9702 of the area under the curve (AUC). Noori *et al.* [25] proposed a hybrid FS model using MI and BGWO for cancer classification. NB showed 88.39% of accuracy. Jain *et al.* [26] utilized a hybrid FS model using ReliefF and principal component analysis (PCA) for diagnosis of diabetes and BC. The model achieved 82.16 % of accuracy for diabetes and 81.73% of accuracy for BC.

3. Methodology

The suggested architecture is presented in Fig. 1. The model starts with the acquisition of breast cancer datasets. Secondly, features are determined. Then, data-preprocessing process is used to fix the errors and improve the quality of the feature dataset. After that, using the hybrid FS model, the most discriminating feature are determined. The hybrid feature selection method consists of two stages. Using relief approach is used to select the top-ranking features. In the second step, BHHO is used to find the optimal feature subset from the filtered sorted data. Finally, using machine learning models, the malign and benign lesions are classified. Codes related to the proposed model have been uploaded to github https://github.com/turkfuat/erkanoptcodes.

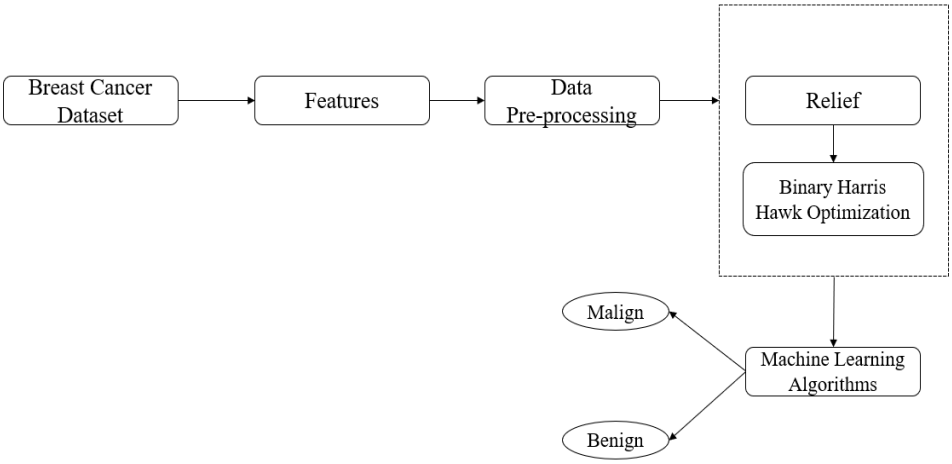

Fig. 1 *Suggested framework for breast cancer classification.*

3.1 The descriptions of breast cancer datasets

Three different breast cancer datasets are used in this study. The first dataset is the Wisconsin diagnostic breast cancer dataset (WDBC) which consists of 569 instances with 31 features. The first feature represents the patient ID number and the remaining feature are 30 input features. The features are extracted from images of cell nuclei. Each instance is labeled as benign and malign. There are 357 benign and 212 malign instances. The second dataset is the Wisconsin breast cancer dataset (WBCD) which consists of 699 instances with 10 features. The first feature represents the patient ID number again. 241 cases are malignant and 458 cases are benign [26–28]. A clinical dataset containing mammography images of a total of 101 patients is used as the third dataset (40 patients are benign, 61 patients are malign). The patients were confirmed with benign and malign breast lesions by hispathologic examinations or the ones who were confirmed with benign lesions as a result of two years radiological periodic follow-up. All patients who underwent digital mammography between April 2015 and April 2020 were

reached through the picture archiving and communication system (PACS). Using IMS Giotto digital mammography (Bologna-Italy), all patients underwent. This retrospective study was approved by local institutional review board and informed consent was waived due to the retrospective nature of the study.

3.2 Data pre-processing

Data normalization is a preprocessing technique that aims to identify numeric values in the datasets within certain range. Z-score normalization technique is used in this study. The data are standardized by calculating the new value for each attribute according to the distance from the mean value and the standard deviation in the attribute values [29].

3.3 Features of breast cancer datasets

Full knowledge of defined attributes related to WDBC and WBCD can be referred to [18,30]. However, the mammogram images dataset has not defined any features and feature extraction which is needed. Before the feature extraction, the breast lesions need to be identified. Therefore, a segmentation process has been applied to all images to determine the region of interest (ROI). The process of extraction of ROI is shown in Fig. 2. First, the mammogram images are retrieved (Fig. 2(a)). Then, two radiologists manually defined the boundaries of the ROIs with a green contour using RadiAnt DICOM Viewer software (Fig. 2(b)). Disagreements between the expert radiologists were resolved by consensus. As a result, a total ROI of 195 were determined (116 images are malign, 79 images are benign). Subsequently, using gray level thresholding and morphological operations techniques, each of ROI were segmented on "MATLAB 2020a" program (Fig. 2(c)).

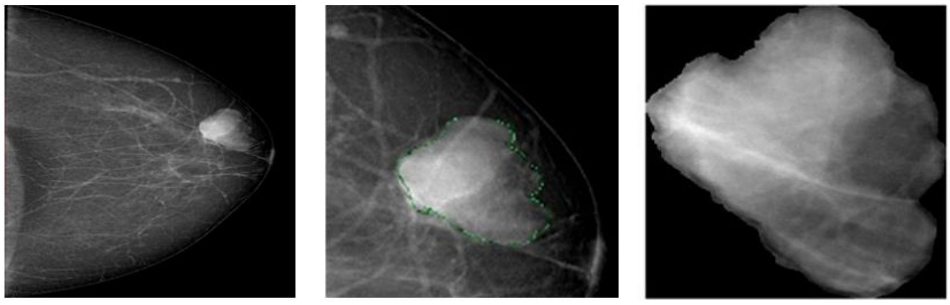

Fig. 2 *a) Original mammogram image, b) Marking of ROI with green contours, c) Extracted ROI.*

A total of 127 features are calculated for each ROI, including 16 shape-related, 15 histogram-based, 52 gray level co-occurance matrix and 44 gray level run matrix, respectively [31–33].

3.4 The proposed hybrid feature selection approach

The suggested hybrid FS model is depended on relief and BHHO algorithm. Fig. 3 illustrates the hybrid FS model process consisting of two stages. The first-stage model uses relief ranking for data size reduction by assessing of weights of each feature that distinguishes different categories. The main idea of this model is to calculate the importance of each feature according to each class. The "relieff" function in the MATLAB program can calculate the importance of each feature by ranking the features according to their weights. Higher weights illustrate higher discrimination of this feature from the other categories. This means that the feature with high weight can be used to calculate classification results effectively. A threshold value is applied to select distinctive features after all weights have been calculated. Here we selected 0 as a threshold value. If the weights of the features are higher than 0, the feature is selected, otherwise, it is ignored. Using filter-based FS model, the number of features can be significantly diminished. A wrapper-based FS model is used to further reduce the number of features in the second stage. The most distinctive features are determined using the BHHO FS method. The BHHO method is wrapper-based FS method that mimics the cooperative behavior and chase style of Harris hawks called surprise attack. The maximum number of iterations and the fitness function at a certain value were determined as the stopping criterion of the model. When the maximum number of iterations and the fitness function with a certain value are reached, the most selective features are determined. Finally, using the ML algorithm, the breast lesion are classified.

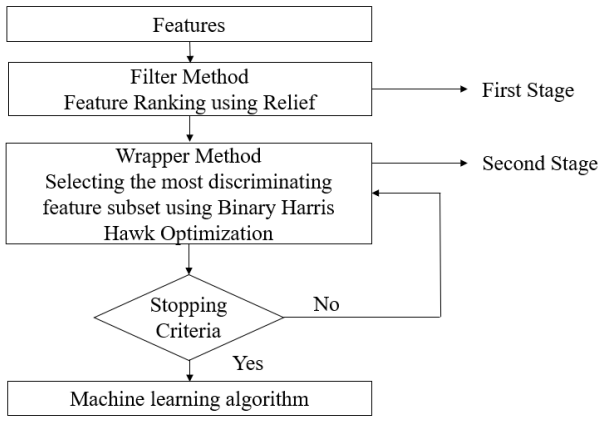

Fig. 3 *The proposed hybrid feature selection.*

3.4.1 Relief

Relief is an effective filter FS algorithm method that was first introduced by Komonenko. This method weighs features according to the relationship between them. The most important features get high weights while the remaining features get small weights. All features are ranked according to this measure [34, 35].

3.4.2 Harris hawk's optimization

In nature, Harris hawks aim to catch their prey by following different strategies. Harris hawk optimization (HHO) is NIM based approach that mimics this hunting strategy of hawks. The algorithm is presented by Heidari et al. [15, 16] in 2019. There are two main phases in these algorithms which are exploration and exploitation. The mathematics process used in the exploration phase is given in Eq. (1) and Eq. (2). X_m refers to the average values of population, ub and lb refer to the lower and upper limits, r_1, r_2, r_3, and r_4 refer to the random values, X_{prey} refers to the current position of the target, t refers to the current position in Eq. (1) and Eq. (2), respectively.

$$X(t+1) = \begin{cases} X_{\text{rand}}(t) - r_1 \left| X_{\text{rand}}(t) - 2r_2 X(t) \right| & \text{if } p \geq 0.5 \\ X_{\text{prey}}(t) - X_m(t) - r_3(lb + r_4(ub - lb)) & \text{if } p < 0.5 \end{cases} \tag{1}$$

$$X_m = \frac{1}{N} \sum_{i=1}^{N} X_i(t) \tag{2}$$

After exploitation phases, exploitation phases is starting. The energy rate of prey is calculated which is shown in Eq. (3). E_0 is the initial value of prey defined in the range $[-1, 1]$. T shows the number of maximum step and t is the current step.

$$E = 2E_0 \left(1 - \frac{t}{T}\right) \tag{3}$$

$$E_0 = (2r - 1) \tag{4}$$

The hawks start the exploitation process with many different attack methods. Firstly, a random number (r) is assigned between $[0, 1]$. According to this r value and E energy, different strategies are exploited. The exploitation phases consist four different stages which are soft besiege, hard besiege, soft besiege with progressive rapid dives (SBPRD) and hard besiege with progressive rapid dives (HBPRD). The process of soft besiege is shown in Eq. (5) and Eq. (6). $\Delta X(t)$ represents the different between the position of the prey and the present hawk. J is the random jump and it takes value between $[0, 2]$. The calculation of J is shown in Eq. (7).

$$x(t+1) = \Delta x(t) - E|JX_{\text{prey}} - X_i(t)| \tag{5}$$

$$\Delta X_i(t) = X_{\text{prey}} - X_i(t) \tag{6}$$

$$J = 2(1 - r) \tag{7}$$

The mathematical modeling of hard besiege is shown in Eq. (8).

$$X_i(t) = X_{\text{prey}} - E\left|\Delta X_i(t)\right| \tag{8}$$

If $E \geq 0.5$ and $r < 0.5$, the hawks perform SBPRD. The mathematical model is shown in Eq. (9). S refers to a random vector and *Levy* shows the Levy flight function and f is the fitness value for the given optimized problem.

$$X(t+1) = \begin{cases} Y = X_{\text{prey}} - E\left|JX_{\text{prey}} - X_i(t)\right| & \text{if} \quad f(Y) < f(X_i(t)) \\ Z = Y + S \times Levy(d) & \text{if} \quad f(Z) < f(X_i(t)) \end{cases} \tag{9}$$

If $E < 0.5$ and $r < 0.5$, the Harris hawks use the HBPRD. The mathematical model is shown in Eq. (10).

$$X(t+1) = \begin{cases} Y = X_{\text{prey}} - E\left|JX_{\text{prey}} - X_m(t)\right| & \text{if} \quad f(Y) < f(X_i(t)) \\ Z = Y + S \times Levy(d) & \text{if} \quad f(Z) < f(X_i(t)) \end{cases} \tag{10}$$

The algorithm of HHO selects the most discriminating feature set by eliminating unnecessary attributes. This task is performed by using the fitness function which measures the quality of the search agent. The search agent quality is assessed by its ability to get the highest accuracy results [36]. The fitness function is defined Eq. (11).

$$fitness = \alpha\, error + (1-\alpha)\frac{|\mathbf{S}|}{|F|} \tag{11}$$

Error denotes the error rate (ER) which is calculated by a learning algorithm, $|\mathbf{S}|$ denotes the length of the feature matrix, $|F|$ denotes the total number of features and α refers parameter which is utilized to control the effects of the learning algorithm and feature size. The first term of the Eq. (11) represents the classification performance and the second term represents the feature reduction. K-NN algorithm is utilized to calculate the ER. K-NN algorithm is selected due to its simplicity and ease of implementation. In order to calculate the ER in the fitness function, a 10-fold cross-validation is used. In the final FS step, the global best solution (best feature subset) consisting of optimal features is generated. The feature subset is then fed into the k-NN in the next step [37]. Also, the population size is selected 10, the maximum iterations are selected 10, 50, 100, respectively and $\alpha = 0.99$, $lb = 0$, and $ub = 1$ are selected.

3.4.3 Binary Harris hawk optimization

HHO originally was used for continuous processes. Therefore, HHO should be converted into a binary version for FS. The sigmoid function (SF) is used to create the binary version (Eq. (12)). For converting to the binary version, continuous-valued $x(t)$ input is given to the sigmoid function. A transfer function refers to the opportunity of changing a position vector' element from 0 to 1. A new position is then computed for each agent in the binary search space in Eq. (13).

$$T\left(x_i^j(t)\right) = \frac{1}{1+\exp\left(-x_i^j(t)\right)} \tag{12}$$

$$x_i^j(t) = \begin{cases} 1 & \text{if} \quad r < T\left(x_i^j(t)\right) \\ 0 & \quad\;\; r \geq T\left(x_i^j(t)\right) \end{cases} \tag{13}$$

where $T(x)$ denotes the SF, r denotes a random value in $[0,1]$, X denotes the location of the hawk, i denotes the order of the hawk, j denotes the dimension, and t denotes the present iteration [37].

3.5 Classification

Four different machine learning methods, SVM, LR, k-NN, and NB were applied to classify malign and benign lesions. LR aims to build a linear model which defines the relationship between dependent and independent variables [38]. SVM aims to determine the hyperplane that will allow the separation between two classes to be optimal [39]. K-NN aims to classify a new sample according to the similarity between the sample s in the training set [40]. NB algorithm uses a series of probability principles to determine the class of the data [41].

4. Results and discussion

The suggested hybrid relief-BHHO model is evaluated using three different breast cancer datasets; (i) Wisconsin diagnostic breast cancer dataset (WDBC) – Dataset 1 (ii) Wisconsin breast cancer dataset (WBCD) – Dataset 2 (iii) mammographic breast cancer dataset (MBCD) – Dataset 3. The suggested model is implemented in MATLAB 2020a program. 10-fold cross validation is utilized for the classification process [42]. This model separates the whole dataset into ten blocks of equal size. 90% of the data is employed to train the suggested model and the remaining 10% for testing. The results of classification are evaluated in terms of accuracy (Acc.), precision (Prec.), recall, F1-score, specificity (Spe.) and balanced accuracy (BA) [43].

The classification process is performed with and without the proposed hybrid FS model (relief-BHHO). In the first stage of the proposed hybrid FS model, the features in the data sets are weighted and ranked according to their importance with the relief algorithm. Then, the value 0 is chosen as the threshold value and the features higher than this threshold value are determined as distinctive features. In next step, BHHO algorithm is used to in order to more efficiently eliminate the unwanted features and the size of the feature set has been significantly reduced. In the BHHO method, the most distinctive features can be selected when the algorithm reaches a fixed fitness value and the determined maximum iteration value. As the maximum number of iterations, 10, 50 and 100 values are selected, respectively, and the algorithm is performed separately for each ML and each dataset. The most discriminating features selected by relief-BHHO algorithm is presented in Tab. I. As a result of the experiments performed on the three datasets, the relief-BHHO method has the lowest fitness value and the lowest number of features using 10th iteration. This result implies that the algorithm provides a good balance between exploration and exploitation phases in 10th iteration. Therefore, the features obtained as a result of the 10th iteration are given as input data to the machine learning algorithms and the classification processes of the machine algorithms are calculated separately in three data sets.

Tab. II presents the classification performance of ML in terms of performance measures on the three datasets with and without relief-BBHO. When the results are examined, the SVM algorithm gives the best result in both cases. Using the proposed relief-BBHO method, SVM has achieved an accuracy of 98.77%, a precision of 97.17%, a recall of 99.52%, and a F1-score of 98.33%, a specificity of 99.72% and a balanced accuracy of 99.62% for Dataset 1. Similarly, SVM has showed the

Dataset	Max. iterations	Original features	Features selected	Fitness value
Dataset 1	**10**	**30**	**4**	**0.016**
	50	30	5	0.019
	100	30	4	0.017
Dataset 2	**10**	**9**	**3**	**0.027**
	50	9	4	0.032
	100	9	3	0.03
Dataset 3	**10**	**127**	**7**	**0.026**
	50	127	9	0.031
	100	127	8	0.028

Tab. I *Suggested framework for breast cancer classification.*

Classifier	Metrics	Dataset 1		Dataset 2		Dataset 3	
		Without	With	Without	With	Without	With
LR	Acc.	91.92	97.19	93.71	97.14	86.15	92.82
	Prec.	91.04	95.75	90.87	94.61	94.83	92.24
	Recall	87.73	96.67	90.87	97.02	83.97	95.54
	F1-score	89.35	96.21	90.87	95.80	89.07	93.86
	Spe.	94.56	98.04	95.2	98.47	90.63	93.67
	BA	91.14	97.35	93.03	97.75	87.3	94.6
NB	Acc.	92.62	97.37	93.99	97.71	90.77	94.87
	Prec.	85.38	94.34	91.29	96.68	90.52	95.69
	Recall	94.27	98.52	91.29	96.68	93.75	95.69
	F1-score	89.60	96.39	91.29	96.68	92.11	95.69
	Spe.	91.78	99.16	95.41	98.25	86.75	93.67
	BA	93.02	98.84	93.35	97.46	90.25	94.68
SVM	Acc.	94.73	98.77	94.56	99.28	91.79	97.44
	Prec.	91.04	97.17	94.61	98.76	93.97	97.41
	Recall	94.61	99.52	90.12	99.17	92.37	98.26
	F1-score	92.79	98.33	92.31	98.96	93.16	97.84
	Spe.	94.79	99.72	97.09	99.56	90.91	97.47
	BA	94.70	99.62	93.60	99.36	91.64	97.86
K-NN	Acc.	91.56	96.13	93.85	97.00	91.28	95.38
	Prec.	85.85	93.87	89.63	96.27	93.10	94.83
	Recall	91.00	95.67	92.31	95.08	92.31	97.35
	F1-score	88.35	94.76	90.95	95.67	92.70	96.07
	Spe.	91.87	97.48	94.62	97.38	89.74	96.20
	BA	91.43	96.58	93.47	96.23	91.03	96.77

Tab. II *The classification performance of ML in terms of performance measures on the three datasets with and without relief-BBHO.*

highest classification results compared to other algorithms for Dataset 2 (99.28% of accuracy, 98.76% of precision, 99.17% of recall, 98.96% of F1-score, 99.56% of specificity and 99.36% of balanced accuracy). Comparing the classification results for Data 3, SVM has achieved the best result (97.44% of accuracy, 97.41% of precision, 98.26% of recall, 97.84% of F1-score, 97.47% of specificity and 97.86% of balanced accuracy).

To demonstrate the effect of hybrid feature selection (HFS), Fig. 4 lists the comparison results between relief-BHHO and without feature selection (WFS). The results are shown in terms of accuracy for each ML algorithm. When comparing the relief-BHHO and WFS, the accuracy rates increased with HFS for each ML algorithm. The accuracy rate of LR improved with relief-BHHO by 5.27%, 3.43%, and 6.67% for both datasets, respectively. By using the relief-BHHO, the accuracy values of NB increased by 4.75%, 3.72%, and 4.1% for three datasets, respectively. Similarly, the relief-BHHO method increased the accuracy of SVM by 4.04%, 4.72%, and 5.65%, respectively. The accuracy values of k-NN increased 4.57%, 3.15%, and 4.1%, respectively, when using the relief-BHHO model.

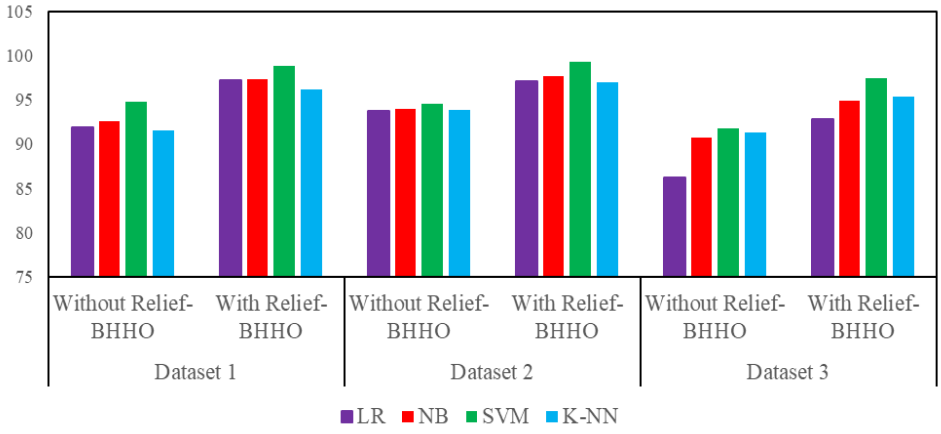

Fig. 4 *Comparison of relief-BHHO with and without feature selection.*

The proposed hybrid relief-BHHO approach is compared in terms of robustness with 3 methods frequently used in the literature: BGWO, GA and BPSO [44–48]. The comparison was conducted same features and the same hybrid feature selection procedure (Fig. 3), only the wrapper part is changed. Instead of binary Harris hawk optimization algorithm, GA, BPSO, and BGWO approaches are applied as wrapper feature selection, respectively. The population size is selected 10, the maximum iterations are selected 10, 50, 100, lb and ub are selected 0 and 1, respectively, crossover and mutation in GA are set 0.8 and 0.01, respectively, inertia weight in and acceleration constant c_1 and c_2 in PSO are set 0.9 and 2, respectively.

Tab. III highlights the comparison results in terms of the number of selected features, fitness value, and accuracy. For Dataset 1, 3 features for relief-BHHO and relief-BPSO and 4 features for relief-GA, and 5 features for relief-BGWO are

Method	Dataset 1			Dataset 2			Dataset 3		
	SF	FV	Acc. (%)	SF	FV	Acc. (%)	SF	FV	Acc. (%)
Relief-BHHO-SVM	3	0.027	98.77	4	0.016	99.28	7	0.026	97.44
Relief-GA-SVM	4	0.033	97.71	3	0.02	97.4	3	0.035	94.36
Relief-BPSO-SVM	3	0.048	96.42	5	0.027	96.5	9	0.038	92.82
Relief-BGWO-SVM	5	0.038	96.85	4	0.019	96.67	7	0.032	95.87

Tab. III *Performances of FS model over BC datasets.*

selected. While relief-BHHO and relief-BPSO select the same number of features, relief-BHHO shows the most oriented fitness value with 0.027. In addition, relief-BHHO-SVM shows the highest performance in terms of accuracy with 98.77%. For Dataset 2, while relief-GA selects the least number of features, their mean fitness value is greater than relief-BHHO. Additionally, the relief-BHHO-SVM method outperforms with 99.28% of accuracy. Similarly, relief-GA selects the least number of features for Dataset 3. However, relief-BHHO selects 7 features with a 0.026 fitness value and this seems to be the most fitness-oriented. When using an SVM classifier with the relief-BHHO approach, it achieves an accuracy of 97.44%.

GWO, PSO and GA are some of the NIM algorithms frequently used in the literature. As a common feature of these algorithms, there are two stages in the search steps, namely exploration and exploitation. Due to these structures, these algorithms may encounter such as trapped in local optima and immature convergence during their exploration and exploitation phases. However, HHO includes a total six stages, two in exploration and four in exploitation, and randomly performs one of these stages to find the optimal solution. In the study of Heidari [15] et al., HHO achieved more successful results compared to other NIM algorithms. Similarly, in our study, HHO achieves better results in terms of fitness value (FV) and classification performance when compared to algorithms such as GWO, PSO and GA. Based on this evidence, it can be said that the relief-BHHO-SVM method is a more convenient method for breast cancer datasets with its low fitness value and high accuracy value.

The result obtained with the proposed hybrid relief-BHHO method are compared with the results of some similar studies on the detection of breast cancer in the literature. Tab. IV shows the comparison of the proposed study with previous studies based on classification accuracy. For fair comparison, previous studies that predicted breast cancer on WDBC and WBCD datasets using feature selection and classification methods are considered. When compared with similar studies, it can be said that the presented method achieves very successful accuracy rates. Compared to Alshayeji et al only, the accuracy seems to be slightly lower. However, we would like to state that our study was tested on 3 different data sets and the total processing complexity is less than artificial neural networks thanks to the feature selection-optimization processes. This situation stands out as an important advantage of our study.

Reference	Methods	Dataset	Accuracy (%)
Alshayeji *et al.* [18]	ANN	WBCD	99.85
		WDBC	99.47
Salama *et al.* [49]	SMO	WDBC	97.71
		WBCD	96.99
Chaurasia *et al.* [50]	SMO	WBCD	96.19
Mafarja *et al.* [51]	BGOA-M	WBCD	97.43
Ibrahim *et al.* [52]	SSA-PSO	WDBC	96.97
		WDBC	98.00
Rao *et al.* [53]	ABCoDT	WDBC	97.18
Suggested model	Relief-BHHO-SVM	WDBC	98.77
		WBCD	99.28
		MBCD	97.44

Tab. IV *Comparison of our proposed method with similar recent studies. ANN: Artificial neural network, SMO: Sequential minimal optimization, ABCoDT: Artificial bee colony-gradient boosting decision tree, SSA: Salp swarm optimization, BGOA-M: Binary grasshopper algorithm-mutation operator.*

5. Conclusion and future works

Recent years, quite a few approaches have been proposed for BC classification. However, building an effective classification model is a challenging issue for researchers. The aim of this study is to present a hybrid FS based on relief-BHHO. For classification process, four different ML algorithms such as LR, NB, SVM and k-NN are used, respectively. The suggested model in the study is tested on three different BC datasets. The relief-BHHO FS approach is utilized to select the most discriminating features and compared to three well-known FS models. The set of features selected by the suggested hybrid FS and three well-known FS models are given as separately inputs to ML algorithms. Among many comparisons, relief-BHHO-SVM has also achieved the best performance when tested on all three datasets. Moreover, the relief-BHHO approach has improved the performance of ML algorithms for the three datasets. When comparing the results of relief-BHHO approach to other well-known FS approaches, relief-BHHO approach has showed the better performance with low fitness value. In the future, other feature selection methods, filter selection approaches, and classification methods such as deep learning and its variants can be considered as potential alternatives to the proposed scheme. We think that this hybrid FS model can be used not only in breast cancer but also in the diagnosis of other cancer types.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Credit authorship contribution statement

Erkan Akkur: Methodology, software, formal analysis, data curation.
Fuat Turk: Conceptualization, supervision, software, writing review & editing.
Osman Eroğul: Supervision, writing review & editing.

Acknowledgement

We would like to express my very great appreciation to MD. Oguz LAFCI, Assoc. Prof. Pelin Seher OZTEKIN, MD. Pınar Celepli and Assoc. Prof. Pınar Nercis KOŞAR for his valuable and constructive suggestions during the development of this research work.

The authors received no specific funding for this study.

References

[1] SUNG H., FERLAY J., SIEGEL R.L., LAVERSANNE M., SOERJOMATRAM I., JEMAL A., BRAY F. *Global cancer statistics 2020: GLOBOCAN Estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA. Cancer J. Clin. 2021, 71(3), pp. 209–249, doi: 10.3322/caac.21660.

[2] LOIBL S., POORTMANS P., MORROW M., DENKERT C., CURIGLIANO G. *Breast cancer.* The Lancet 2021, 397, pp. 1750–1769, doi: 10.1016/s0140-6736(20)32381-3.

[3] KARALE V.A., SINGH T., SADHU A., KHANDELWAL N., MUKHOPADHYAY S. *Reduction of false positives in the screening CAD tool for microcalcification detection.* Sadhana – Acad. Proc. Eng. Sci., 2020, 45(1), pp. 1–11, doi: 10.1007/s12046-019-1260-4.

[4] SINGH B., KUR M. *An approach for classification of malignant and benign microcalcification clusters.* Sadhana – Acad. Proc. Eng. Sci. 2018, 43(3), pp. 1–18, doi: 10.1007/s12046-018-0805-2.

[5] WU J., HICKS C. *Breast cancer type classification using machine learning.* J. Pers. Med. 2021, 11(2), doi: 10.3390/jpm11020061.

[6] ALZU'BI A., NAJADAT H., DOULAT W., AL-SHARI O., ZHOU L. *Predicting the recurrence of breast cancer using machine learning algorithms.* Multimed. Tools Appl. 2021, 80(9), pp. 13787–13800. doi: 10.1007/s11042-020-10448-w.

[7] DHAHRI H., MAGHAYREH E. AL, MAHMOOD A., ELKILANI W., FAISAL NAGI M. *Automated breast cancer diagnosis based on machine learning algorithms.* J. Healthc. Eng., 2019, 2019, 4253641, doi: 10.1155/2019/4253641.

[8] KHAIRE U.M., DHANALAKSHMI R. *Stability of feature selection algorithm*: A review Journal of King Saud University – Computer and Information Sciences, 2022, 34(4), pp. 1060–1073.

[9] BOMMERT A., SUN X., BISCHL, RAHNENFÜHRER B.J., LANG M. *Benchmark for filter methods for feature selection in high-dimensional classification data*, Comput. Stat. Data Anal., 2020, 143, 106839, doi: 10.1155/2019/4253641.

[10] WONG W.K., MING C.I. *A Review on Metaheuristic Algorithms: Recent trends, benchmarking and applications.* In: *2019 7th International Conference on Smart Computing & Communications (ICSCC). Sawarak,* Malaysia, 2019, pp. 1–5, doi: 10.1109/icscc.2019.8843624.

[11] CARBAS S., TOKTAS A., USTUN D., *Introduction and overview: Nature-inspired metaheuristic algorithms for engineering optimization applications.* In: Nature-Inspired Metaheuristic Algorithms for Engineering Optimization Applications. Springer, Singapore. 2021, pp. 1–9, doi: 10.1007/978-981-33-6773-9_1.

[12] SAKRI S.B., ABDUL RASHID N.B., MUHAMMAD ZAIN Z. *Particle swarm optimization feature selection for breast cancer recurrence prediction*. IEEE Access. 2018, 6, pp. 29637–29647, doi: 10.1109/access.2018.2843443.

[13] CHAUHAN P., SWAMI A. *Breast cancer prediction using genetic algorithm based ensemble approach*. In: 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) Bengaluru, India. 2018, pp. 1–8.

[14] KUMAR S., SINGH M. *Breast cancer detection based on feature selection using enhanced Grey Wolf Optimizer and Support Vector Machine Algorithms*. Vietnam J. Comput. Sci. 2021, 8(2), pp. 177–197, doi: 10.1142/s219688882150007x.

[15] HEIDARI A.A., MIRJALILI S., FARIS H., ALJARAH I., MAFARJA M., CHEN H. *Harris hawks optimization: Algorithm and application*. Futur. Gener. Comput. Syst. 2019, 97, pp. 849–872, doi: 10.1016/j.future.2019.02.028.

[16] HEIDARI A. A., MIRJALILI S., FARIS H., ALJARAH I., MAFARJA M., CHEN H. *Harris hawks optimization: Algorithm and application. Intelligence based optimization method: Harris' Hawk Optimization*. In: International conference on intelligent systems design and applications (ISDA 2018). Vellore, India, Springer, Cham. 2018, 2, pp. 832–842, doi: 10.1016/j.future.2019.02.028.

[17] JIANG F., ZHU Q., TIAN T. *Breast cancer detection based on modified harris hawks optimization and extreme learning machine embedded with feature weighting*. Neural Process. Lett. 2022, pp. 1–24, doi: 10.1007/s11063-021-10700-w.

[18] ALSHAYEJI M.H., ELLETHY H., GUPTA R. *Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach*. Biomedical Signal Processing and Control, 2022, 71, 103141, doi: 10.1016/j.bspc.2021.103141

[19] RANJBARI S., KHATIBI T., VOSOUGH DIZAJI A., SAJAD, H., TOTONCHI M., GHAFFARI F. *CNFE-SE: a novel approach combining complex network-based feature engineering and stacked ensemble to predict the success of intrauterine insemination and ranking the features*. BMC Medical Informatics and Decision Making. 2021, 21(1), pp. 1–29, doi: 10.1186/s12911-020-01362-0.

[20] SANGAIAH I., VINCENT ANTONY KUMAR A. *Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction*. Cluster Comput. 2019, 22(3), pp. 6899–6906, doi: 10.1007/s10586-018-1702-5.

[21] LOEY M., M. JASIM W., EL-BAKRY H.M., TAHA M.H.N., KHALIFA N.E.M. *Breast and colon cancer classification from gene expression profiles using data mining techniques*. Symmetry. 2020, 12(3), 408, doi: 10.3390/sym12030408.

[22] ALOMARI O.A., KHADER A.T., AL-BETAR M.A., ALKAREEM ALYASSERI Z.A. *A hybrid filter-wrapper gene selection method for cancer classification*. In: 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS), 2018, Kuching, Malaysia pp. 113–118, doi: 10.1109/icbaps.2018.8527392.

[23] MUFASSIRIN M.M.M., RAGEL R.G. *A novel filter-wrapper based feature selection approach for cancer data classification*. In: 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), Colombo, Sri Lanka, 2018, pp. 1–6, doi: 10.1109/iciafs.2018.8913362.

[24] ALZUBAIDI A., BROWN D., COSMA G., GRAHAM POCKLEY A. *Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information*. In: 2016 International Conference on Interactive Technologies and Games (ITAG), Nottingham, UK. 2016, pp. 70–76, doi: 10.1109/itag.2016.18.

[25] NOOR MUHAMMED N., OMAR SABER Q. *Improving cancer diseases classification using a hybrid filter and wrapper feature subset selection*. Ann. Proteomics Bioinforma. 2020, 4(1), pp. 6–11, doi: 10.29328/journal.apb.1001010.

[26] JAIN D., SINGH V. *Diagnosis of breast cancer and diabetes using hybrid feature selection method*. In: Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India. 2018, pp. 64–69, doi: 10.1109/pdgc.2018.8745830.

[27] ASUNCION A., NEWMAN D.J. UCI Machine Learning Repository: Data Sets, University of California Irvine School of Information, 2007.

[28] WOLBERG W.H., STREET W.N., MANGASARIAN O.L. *Image analysis and machine learning applied to breast cancer diagnosis and prognosis*. Anal. Quant. Cytol. Histol. 1995, 17(2), pp. 77–87, doi: 10.1016/0304-3835(94)90099-x.

[29] MISHRA A.K., CHOUDHARY A., CHOUNDHARY S. *Normalization and Transformation Technique Based Efficient Privacy Preservation In Data Mining*. International Journal of Modern Engineering and Research Technology, 2016, 3(2), pp. 5–10, doi: 10.1007/978-981-15-1480-7_66.

[30] AHMED M.T., IMTIAZ M.N., KARMAKAR A., *Analysis of Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction*. J. Sci. Technol. Environ. Informatics. 2020, 9(2), pp. 665–672, doi: 10.18801/jstei.090220.67.

[31] VADIVEL A., SURENDIRAN B. *A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories*. Comput. Biol. Med. 2013, 43(4), pp. 259–267, doi: 10.1016/j.compbiomed.2013.01.004.

[32] HARALICK R., DINSTEIN M.I., SHANMUGAM K. *Textural features for image classification*. IEEE Trans. Syst. Man Cybern. 1973, SMC-3, pp. 610–621, doi: 10.1109/tsmc.1973.4309314.

[33] GALLOWAY M.M. *Texture analysis using gray level run lengths*, Comput. Graph. Image Process. 1975, 4(2), pp. 172–179, doi: 10.1016/s0146-664x(75)80008-6.

[34] URBANOWICZ R., MEEKER J.M., LA CAVA W., OLSON R.S., MOORE J.H. *Relief-based feature selection: Introduction and review*. Journal of Biomedical Informatics. 2018, 85, pp. 189–203, doi: 10.1016/j.jbi.2018.07.014.

[35] REYES O., MORELL C., VENTURA S. *Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context*. Neurocomputing. 2015, 161, pp. 168–182, doi: 10.1016/j.neucom.2015.02.045.

[36] HANS R., KAUR H., KAUR N. *Opposition-based Harris Hawks optimization algorithm for feature selection in breast mass classification*. J. Interdiscip. Math. 2020, 23(1), pp. 97–106, doi: 10.1080/09720502.2020.1721670.

[37] TOO J., ABDULLAH A.R., SAAD N.M. *A new quadratic binary harris hawk optimization for feature selection*. Electron. 2019, 8(10), 1130, doi: 10.3390/electronics8101130.

[38] HEIBERGER R.M., HOLLAND B. *Logistic regression*. In: Statistical analysis and data display. Springer Texts in Statistics. Springer, New York, NY., 2004, doi: 10.1007/978-1-4757-4284-8_17.

[39] BURGES C.J.C. *A tutorial on support vector machines for pattern recognition*. Data Min. Knowl. Discov.1998, 2(2), pp. 917–937, doi: 10.1049/ic:19990359.

[40] CUNNINGHAM P., DELANY S.J. *K-Nearest Neighbour Classifiers-A Tutorial*, ACM Computing Surveys, 2021, 54(6), pp. 1–25, doi: 10.1145/3459665.

[41] KHARYA S., SONI S. *Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection*, Int. J. Comput. Appl., 2016, 133(9), pp. 32–37, doi: 10.5120/ijca2016908023.

[42] BERRAR D. *Cross-validation*. in Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. 2018, 142, pp. 542–545, doi: 10.1016/b978-0-12-809633-8.20349-x.

[43] MOKARRAMA M.J., KHATUN S., AREFIN M.S. *A content-based recommender system for choosing universities*. Turkish Journal of Electrical Engineering and Computer Sciences., 2020, 28(4), pp. 2128–2142, doi: 10.3906/elk-1911-37.

[44] TOO J., ABDULLAH A.R., SAAD N.M., ALI N.M., TEE W. *A new competitive binary grey wolf optimizer to solve the feature selection problem in EMG signals classification*. Computers., 2018, 7(4), 58, doi: 10.3390/computers7040058.

[45] TOO J., ABDULLAH A.R. *Opposition based competitive grey wolf optimizer for EMG feature selection*. Evol. Intell., 2021, 14(4), pp. 1691–1705, doi: 10.1007/s12065-020-00441-5.

[46] SHOPOVA E.G., VAKLIEVA-BANCHEVA N.G. *BASIC-A genetic algorithm for engineering problems solution.* Comput. Chem. Eng., 2006, 30(8), pp. 1293–1309, doi: 10.1016/j.compchemeng.2006.03.003.

[47] TOO J., ABDULLAH A.R., SAAD N.M. *A new co-evolution binary particle swarm optimization with multiple inertia weight strategy for feature selection.* Informatics, 2019, 6(2), 21, doi: 10.3390/informatics6020021.

[48] TOO J. ABDULLAH A.R., SAAD N.M., TEE W. *EMG feature selection and classification using a Pbest-guide binary particle swarm optimizatio*n. Computation. 2019, 7(1), 12, doi: 10.3390/computation7010012.

[49] SALAMA G.I., ABDELHALIM M., ZEID M.A. *Breast cancer diagnosis on three different datasets using multi-classifiers.* Breast Cancer. 2012, 32(569), 2, doi: 10.1109/icces.2012.6408508.

[50] CHAURASIA V., PAL S. *A novel approach for breast cancer detection using data mining techniques.* International Journal of Innovative Research in Computer and Communication Engineering 2017, 3297(1), pp. 2320–9801, doi: 10.2139/ssrn.3139141.

[51] MAFARJA M., ALJARAH I., FARIS H., HAMMOURI A.I., ALA'M A.Z., MIRJALILI S. *Binary grasshopper optimisation algorithm approaches for feature selection problems.* Expert Syst Appl 2019, 117, pp. 267–86, doi: 10.1016/j.eswa.2018.09.015.

[52] IBRAHIM R.A., EWEES A.A., OLIVA D., ABD ELAZIZ M., LU S. *Improved salp swarm algorithm based on particle swarm optimization for feature selection.* J Ambient Intell Hum Comput, 2019, 10(8), pp. 3155–69, doi: 10.1007/s12652-018-1031-9.

[53] RAO H., SHI X., RODRIGUE A.K., FENG J., XIA Y., ELHOSENY M., YUAN X., GU L. *Feature selection based on artificial bee colony and gradient boosting decision tree.* Appl Soft Comput. 2019, 74, pp. 634–42, doi: 10.1016/j.asoc.2018.10.036.