Robust Method for Breast Cancer Classification Based on Feature Selection Using RGWO Algorithm

Ali Mezaghrani[1](✉), Mohamed Debakla[1], and Khalifa Djemal[2]

[1] Faculty of Science Exact, University of Mustapha Stambouli Mascara, Mascara, Algeria
{Ali.mezaghrani,Debakla_med}@univ-mascara.dz
[2] IBISC Laboratory, Evry Val d'Essone University, Evry, France

Abstract. Breast cancer is a leading cause of mortality in women all over the world. According to the worldwide cancer statistics, early detection and treatment are keys components for improving the recovery rate of breast cancer and lowering the death rate. Machine learning solutions have been proved to be particularly very successful in exploring the origins of such severe diseases, which requires processing vast amounts of data.

In the present study, robust grey wolf optimisation-Random Forest (RGWO-RF) approach was proposed. Our proposed approach based on two steps feature selection process and classification. Modified Grey Wolf Optimizer is used to locate and determine the most significant features. Then, utilizing the prior optimum selections of features, by using Random Forest (RF) classifier to classify breast cancer disease. The reason for using RF it's robustness and highest accuracy.

We apply the proposed approach on Wisconsin Diagnostic Breast Cancer (WDBC) database. The experimental result improve that the hybridation between RGWO for feature selection and RF classifier increase the accuracy rate of classification and demonstrating it's robustness in identifying the breast cancer.

Keywords: Breast cancer · Grey wolf optimizer · Feature selection · Random Forest

1 Introduction

Breast cancer disease has been considered as one of the deadly disease in the world [1]. To increase the odds of survival and save more women's lives, early detection of breast cancer is critical and very important factor in the diagnostic. Breast cancer identification requires precise categorization of the tumor as benign or malignant [2]. Several methods proposed by researchers to enhance the classification capability of their system of breast cancer diagnosis. But still, there is a huge opportunity to create a breast cancer categorization system that is more efficient.

In this study we try to develop an approach which effectively classifies the breast cancer tumor using RGWO for feature selection and RF for disease classification. The suggested strategy aims to extract the most significant and optimum subset of features

M. Salem et al. (Eds.): ICAITA 2022, CCIS 1769, pp. 18–27, 2023.
https://doi.org/10.1007/978-3-031-28540-0_2

from the dataset which helps to make an efficient and effective classification of breast cancer. In this study, the RGWO-RF method is suggested to identify the ideal set of features that would improve the RF classifier classification performance. The RF classifier will be trained on the optimized subset of features identified by RGWO.

The WDBC breast cancer dataset, which has a total of 569 instances and 33 characteristics. The WDBC dataset accessible through the UCI Machine Learning Repository [3], was used to test the suggested methodology.

The experimental result shown that the proposed system increases the accuracy rate when we use RGWO for feature selection. The suggested method outperformed recent studies, obtaining an accuracy of 98.60%, Precision (98.1%), F1-Score (98.1%), Sensitivity (98.1%) and a Specificity value of 98.9%.

The remainder of the essay is structured as follows: Sect. 2 discusses relevant research and several cutting-edge methods for diagnosing breast cancer. The description of the GWO algorithm and the Mathematical model of GWO. In Sect. 4, We provide a thorough explanation of the strategy we suggest. Experimentation and debate are covered in Sect. 5. Finally, in Sect. 6, we conclude our paper with a summary and outlooks for future work.

2 Related Works

The current section provides a summary of the techniques and algorithms used in the suggested study. Based on feature selection and machine learning methodologies, several strategies have been established to identify breast cancer.

Based on tumor traits, the study in [4] sought to make a diagnosis of breast cancer. K-means and K-SVM were combined in order to extract meaningful information from WDBC dataset. Hidden patterns of benign and malignant tumors are found using the K-means method. The outcomes showed that the suggested approach was effective in diagnosing breast cancer while also reducing training time.

Using various data mining approaches, the study in [5] aimed to evaluate the likelihood of developing breast cancer as well as the likelihood of the disease returning. The Wisconsin dataset of UCI machine learning was used to collect cancer patient data. The results show that the Naive Bayes and the decision tree algorithm are more accurate and deliver superior outcomes.

Breast cancer detection was investigated using the SVM approach in [6]. The accuracy, ROC, measurement, and computational time of training were employed as benchmarks in this study. The results supported the SVM algorithm's better performance over other classification methods.

Dora et al. [7] suggested a new technique for calculating the ideal weight coefficients in order to train samples termed GNRBA (Gauss–Newton representation-based approach). The purpose of this method is to reduce computing complexity while also reducing reaction time. GNRBA beats the previous methods in both the UCI cancer datasets.

Shahnaz et al. [8] performed and examined many statistical and deep learning data studies on a dataset of breast cancer cases in order to improve the classification accuracy through feature selection.

Li et al. [9] introduced a medical diagnostic system that used the grey wolf optimizer with the kernel extreme learning machine to determine the ideal feature subset for medical data. With the use of the wrapper approach and a novel fitness function,

Liu et al. [10], presented a novel breast cancer intelligent detection method, Implementing a feature selection process using Information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW). In this procedure, they rank the features using the IG method, and then they use the cost-sensitive support vector machine (CSSVM) learning algorithm to extract the top m optimum features. The efficacy of the suggested approach is tested on Wisconsin Original Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) breast cancer data sets, and the outcomes show that the suggested hybrid algorithm works better than existing techniques.

A chaotic crow search algorithm technique, which was a meta-heuristic optimizer, was recommended by Sayed et al. [11] to address the issue of poor convergence rate.

A technique for determining the ideal qualities for a decision tree's input using a bee colony algorithm was described by Rao et al. [12], along with a way for generating decisions using an artificial bee colony algorithm.

In [13], To handle feature selection for classification issues based on wrapper approaches employing KNN classifier, M. Abdel-Basset and D. El-Shahat developed a novel Grey Wolf Optimizer algorithm coupled with a Two-phase Mutation. To demonstrate the effectiveness and performance of the suggested method, statistical studies was performed.

From the mini MIAS dataset, 80 digital mammograms of normal breasts, 40 benign cases, and 40 malignant cases were selected in [14]. In this study, comparison of classification process performance of support vector machines (SVM), artificial neural networks (ANN), then a hybrid SVM-ANN model was used to create a computer-aided detection (CAD) system, and the later model demonstrated a respectable accuracy of 98%.

In [15], In order to accurately identify benign and malignant tumors, the best group of traits must be chosen. an enhanced GWO has been suggested with SVM applying on WDBC dataset, Experimental result show that the new method improve accuracy by 98.24%.

3 Grey Wolf Optimizer

The Grey Wolf Optimizer, developed by Seyedali Mirjalili et al. in 2014 [16], is a population-based meta-heuristics algorithm that mimics the natural leadership structure and hunting behavior of grey wolves. A social hierarchy is observed within the grey wolf group. According to their functions within the pack, the wolves in the pack are assigned positions in the hierarchy. The wolf pack is typically divided into four different groups: alpha (α), beta (β), delta (δ), and omega (ω), depending on how each wolf participates to the hunting process. Figure 1 shows the social hierarchy of the grey wolves, Alpha (α) wolf is the pack commander and should be obeyed by the other wolves in the pack. The second place of hierarchy is occupied by beta wolves (β), Beta wolf assist the alpha in making decisions and are seen to be the best candidate to be the alpha wolf. Delta (δ) wolf represent the third rank in the pack, must subordinate to the alpha and beta, but they rule the omega (ω). Omega (ω) wolves are the least significant members of the pack and are only permitted to eat last, occupy the lowest rung of the hierarchy. They are viewed as the scapegoats in the pack.

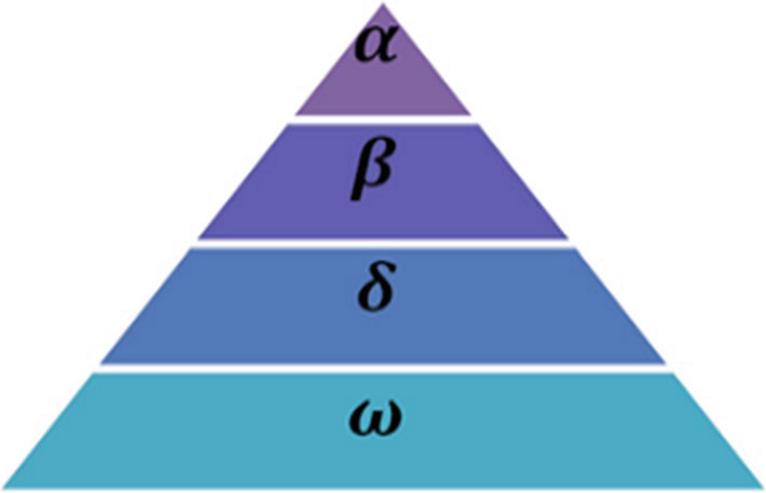

Fig. 1. Grey wolf social structure.

3.1 Mathematical Model of the GWO

In a hunting formation, the pack leaders serve as the spearheads. Once a broad location is established for the prey, they send the omega wolves to encircle it, getting closer as the precise position is sought. The wolves strike after completely encircling their victim.

The procedure might be broken down into three separate parts to represent this as a mathematical model.

3.1.1 Encircling

In the GWO algorithm, hunting is guided by alpha, beta, and delta wolves, and Omega wolves follow them.

The mathematical simulation of grey wolves encircling is shown in Eqs. (1) and (2):

$$\vec{X}(t+1) = \vec{X}_P(t) + \vec{A} \cdot \vec{D} \tag{1}$$

$$\vec{D} = \left| \vec{C} \cdot \overrightarrow{X}_P(t) - \vec{X}(t) \right| \tag{2}$$

where t represent the current iteration, A and C are coefficient vectors, $\vec{X}_P$ is the position vector of the prey, and $\vec{X}$ indicates the position vector of a grey wolf. Following are the calculations for the vectors A and C:

$$\vec{A} = 2\vec{a} \cdot \overrightarrow{r_1} - \vec{a} \tag{3}$$

$$\vec{C} = 2 \cdot \overrightarrow{r_2} \tag{4}$$

where r1 and r2 are random vectors in range [0,1] and components of vector a are linearly reduced from 2 to 0 throughout the course of iterations.

3.1.2 Hunting

Because α, β, and δ are more knowledgeable about the probable locations of prey, omega wolves adjust their positions in line with α, β, and δ in each iteration.

The mathematical model to adjust a search agent's location in accordance with the positions of alpha, beta, and delta search agents is represented by the equations below:

$$\overrightarrow{D_\alpha} = \left|\overrightarrow{C_1}.\vec{X}_\alpha - \vec{X}\right|,\ \overrightarrow{D_\beta} = \left|\overrightarrow{C_2}.\vec{X}_\beta - \vec{X}\right|,\ \overrightarrow{D_\delta} = \left|\overrightarrow{C_3}.\vec{X}_\delta - \vec{X}\right| \tag{5}$$

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - A_1 \cdot \left(\overrightarrow{D_\alpha}\right),\ \overrightarrow{X_2} = \overrightarrow{X_\beta} - A_2 \cdot \left(\overrightarrow{D_\beta}\right),\ \overrightarrow{X_3} = \overrightarrow{X_\delta} - A_3 \cdot \left(\overrightarrow{D_\delta}\right) \tag{6}$$

$$\vec{X}(t+1) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3} \tag{7}$$

3.1.3 Attaking

The coefficient vector A plays a crucial role in the GWO Algorithm, is a random value in the range [−a,a] where a decreases from 2 to 0 throughout the duration of iterations. When random values of $\vec{A}$ are in [−1,1], the next position of a search agent can be in any position between its current position and the position of the prey. If $|A| < 1$ compels the wolves to attack towards the prey (**Exploitation**), $|A| > 1$ compels the grey wolves to diverge from the prey to find a best prey (**Exploration**), there is an other component favors the exploration is $\vec{C}$. In contrast to A, the vector C does not decrease linearly and has random values between [0, 2]. However, C may also be thought of as natural impediments that prevent approaching to the prey [16].

4 Proposed Approach

The GWO algorithm has become very popular among other swarm intelligence techniques, Due to its many benefits, including its ease of use, scalability, and most importantly, its capacity to deliver faster convergence by maintaining the proper balance between exploration and exploitation during the search. We are aware that the placements of the alpha, beta, and delta search agents influence where the prey will be found in the best possible way. These top three search agents are in charge of pointing all other search agents in the right direction for the best prey. In order for the other search agents to be effectively led to approach the prey, it is crucial to ensure that these three search agents are the fittest in each iteration. From Eqs. (5) and (6) it can be seen that each search agent position is updated in relation to the locations of the alpha, beta, and delta search agents. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which comprises 569 instances and 32 characteristics, was used in the current study to evaluate the effectiveness of the suggested methodology. The WDBC dataset was retrieved from the UCI Machine Learning Repository. A digital fine needle aspirate (FNA) scan is used to determine the breast mass attribute. The characteristics depict several parameters that might be helpful in determining whether a tumor is benign or malignant. The RGWO-RF approach is suggested to choose the best subset of characteristics that would produce the highest level of classification accuracy to obtain better results the following points was implemented:

- We use a metaheuristic algorithm to eliminate redundant features and irrelevant features to increase classification accuracy and minimise time consuming in classification process, RGWO algorithm was implemented for variable selection on WDBC dataset, then we use RF classifier to classifier breast cancer based on the subset of optimal features gained by RGWO algorithm. The features with the highest accuracy of classification and the fewest number of selected characteristics is the optimal and the best. The fitness function which is also employed in RGWO to assess the selected features and utilized to optimize classification accuracy, is indicated as Eq. (8):

$$\mathrm{Fitness} = \mathrm{w} * \mathrm{accuracy} + (1 - \mathrm{w}) * 1/(\mathrm{len}(\mathrm{features})) \tag{8}$$

- In relation to the locations of alpha, beta, and delta wolves, the basic GWO updates the position of search agents wolves (Omega wolves). The three best locations of grey wolves are averaged for the position update of search agent (Eq. (7)). This approach results in early convergence and poor solutions. The update method of the positions should not be considered the same in Eq. (7). In this work, To enhance base GWO performance, We use weighted position update concept which proposed by S. Kumar, M. Singh [15], Then we implement a RGWO algorithm in combination with RF classifier and modify the update technique of the position. The mathematical model of weighted position update technique is represented in Eqs. (9) and (10).

$$\mathrm{W1} = \mathrm{A1} * \mathrm{C1} \quad \mathrm{W2} = \mathrm{A2} * \mathrm{C2} \quad \mathrm{W3} = \mathrm{A3} * \mathrm{C3} \tag{9}$$

$$\mathrm{X}(\mathrm{t} + 1) = (\mathrm{W1} * \mathrm{X1} + \mathrm{W2} * \mathrm{X2} + \mathrm{W3} * \mathrm{X3})/(\mathrm{W1} + \mathrm{W2} + \mathrm{W3}) \tag{10}$$

4.1 Methodology

As is depicted in Fig. 2, in this work, we implement two scenario, For the diagnosis of breast cancer, the first one with three major phases was proposed. We start with data preprocessing which is a common phase for two scenario, which mean data cleaning and filtering were performed to avoid the establishment of ineffective rules and patterns. The breast cancer dataset was preprocessed in this article, and outliers were removed using the outer line approach, then classification using RF classifier of breast cancer on all features of WDBC dataset, the third phase is the assessment of classification accuracy. **The second scenario** consisting four main phases, data preprocessing. Then Using RGWO for feature selection, feature selection was used to identify the key characteristics of a given outcome. After that, RF was used for classification process and in the end, the evaluation of classification accuracy.

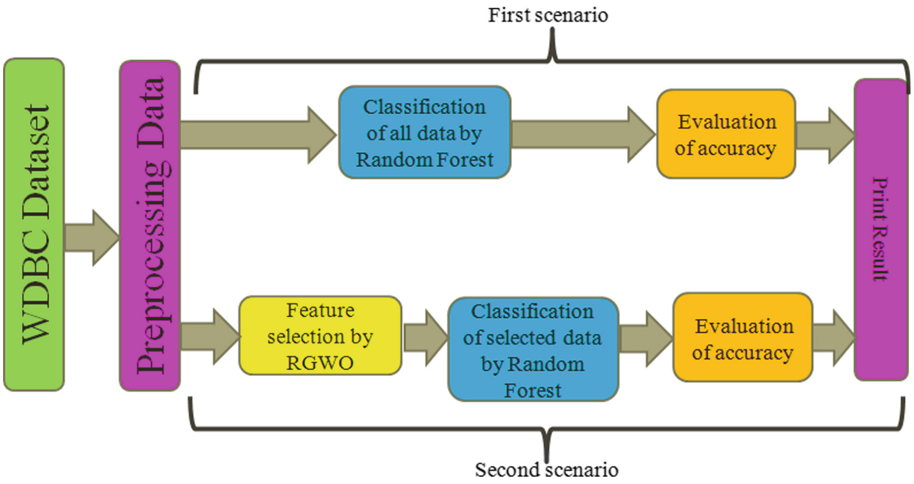
First scenario
WDBC Dataset
Preprocessing Data
Classification of all data by Random Forest
Evaluation of accuracy
Print Result
Feature selection by RGWO
Classification of selected data by Random Forest
Evaluation of accuracy
Second scenario

Fig. 2. Suggested methodology for accurately classifying breast cancer tumors.

5 Experimental Result

By employing the RGWO to reduce the dimensions of features, the primary goal of the current study was to increase diagnostic accuracy while enhancing classification performance. In the experiments, the number of iterations for RGWO algorithm has been fixed at 20 iteration with 10 search agents. We used RF classifier in order to classify the data between malignant and benign tumors, RF is regarded as a very reliable and precise technique. In order to choose the appropriate subset of features, a hybrid strategy using the RF classifier and the RGWO produced the best results. In the proposed approach, the results presented in Table 1 show that when we use dimensionality reduction with RGWO-RF algorithm The Sensitivity, Specificity, Precision, F1-Score, and Accuracy was increased.

5.1 Comparison of the Suggested RGWO-RF Method with the Base GWO-RF

Table 2 compare the RGWO-RF method performance with that of the standard GWO-RF technique. Use of the weighted position update technique in basic GWO is being evaluated through this comparison. The findings have improved in terms of accuracy, F1-score, and sensitivity, as can be seen from the table, which employed the RGWO in the suggested technique.

5.2 Comparing the Suggested Methodology to the Current Feature Selection Methods

The suggested approach was compared with current methods for feature selection-based breast cancer detection approaches in Table 3. It is evident that the recommended RGWO-RF technique outperforms all of the approaches that were considered.

Table 1. Classification results of the suggested RGWO-RF approach using different performance measures.

Performance measures	Classification results (%)	
	Without feature selection	Feature selection with RGWO-RF
Sensitivity	96,3	98,1
Specificity	97,8	98,9
Precision	96,3	98,1
F1-score	96,3	98,1
Accuracy	97,2	98,6

Table 2. Comparison of proposed RGWO-RF approach with the base GWO-RF approaches.

Performance measures	Classification results (%)	
	Proposed RGWO-RF	Base GWO-RF
Sensitivity	98,1	96,3
Specificity	98,9	98,9
Precision	98,1	98,1
F1-score	98,1	97,2
Accuracy	98,6	97,9

Table 3. Evaluation of the proposed RGWO-RF methodology by comparing results with existing feature selection methods.

Approaches	Authors	Years	Number of features	Accuracy %
FS-KNN	Sayed et al. [11]	2019	14	90,28
FS-GBDT	Rao et al. [12]	2019	14	92.80
FS-KNN	Abdel-Basset et al. [13]	2020	16	94,82
FS+EGWO-SVM	S. Kumar and M. Singh [15]	2021	6	98,24
Proposed approach	Proposed	2022	12	98,60

6 Conclusion

The presence of a large number of variables is not always correlated with improved classification performance, as some of them may be redundant, irrelevant, or a source of noise. As a result, a Feature Selection phase is frequently applied to high-dimensional datasets.

In this paper, we proposed a Robust GWO in conjunction with RF classifier. The later has been used to get the best parameters for our new approach. We have shown that this step increases the accuracy of the GWO and hence reduces the mortality rate. PYTHON and WDBC datasets were used to get experimental findings. We discovered that the outcomes of our proposed technique outperform other efforts in term of accuracy measurement, Specificity and Precision. In the near future, we plan to adopt this approach in the diagnosis of other disease with other dataset like heart diseases dataset and diabetes dataset.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **68**(6), 394–424 (2018)
2. Ades, F., et al.: Luminal breast cancer: Molecular characterization, clinical management, and future perspectives. J. Clin. Oncol. **32**, 2794–2803 (2014)
3. Dua, D., Gra®, C.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019). http://archive.ics.uci.edu/ml
4. Zheng, B., Yoon, S.W., Lam, S.S.: Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. **41**(4), 1476–1482 (2014)
5. Pritom, A.I., Munshi, M.A.R., Sabab, S.A., Shihab, S.: Predicting breast cancer recurrence using effective classification and feature selection technique. In: 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 310–314. IEEE, New York (2016)
6. Huang, M.W., Chen, C.W., Lin, W.C., Ke, S.W., Tsai, C.F.: SVM and SVM ensembles in breast cancer prediction. PLoS One **12**(1), e0161501 (2017)
7. Dora, L., Agarwal, S., Panda, R., Abraham, A.: Optimal breast cancer classification using Gauss-Newton representation based algorithm. Expert Syst. Appl. **85**, 134–145 (2017)
8. Shahnaz, C., Hossain, J., Fattah, S.A., Ghosh, S.: Efficient approaches for accuracy improvement of breast cancer classification using Wisconsin database. In: IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (2017)
9. Li, Q., et al.: An enhanced grey wolf optimization based feature selection wrapped kernel extreme learning machine for medical diagnosis. Comput. Math. Methods Med. **2017**, 1–15 (2017). https://doi.org/10.1155/2017/9512741
10. Liu, N., Qi, E., Xu, M., Liu, G.: A novel intelligent classification model for breast cancer diagnosis. Inf. Process. Manage. **56**, 609–623 (2019)
11. Sayed, G.I., Hassanien, A.E., Azar, A.T.: Feature selection via a novel chaotic crow search algorithm. Neural Comput. Appl. **31**(1), 171–188 (2017). https://doi.org/10.1007/s00521-017-2988-6
12. Rao, H., Shi, X., Rodrigue, A., Feng, J., Xia, Y.: Feature selection based on artificial bee colony and gradient boosting decision tree. Appl. Soft Comput. **74**, 634–642 (2019)
13. Abdel-Basset, M., El-Shahat, D., El-Henawy, I., Mirjalili, S.: A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Syst. Appl. **139**, 112824 (2020)
14. Lim, T.S., Tay, K.G., Huong, A., Lim, X.Y.: Breast cancer diagnosis system using hybrid support vector machine-artificial neural network. Int. J. Electr. Comput. Eng. **11**(4), 3059 (2021). https://doi.org/10.11591/ijece.v11i4.pp3059-3069

15. 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. **8**(2), 177–197 (2021)
16. Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Adv. Eng. Softw. **69**, 46–61 (2014)