Novel feature selection method for accurate breast cancer classification using Correlation coefficient and Modified GWO Algorithm

Ali Mezaghrani, Mohammed Debakla, Khalifa Djemal

Abstract

Breast cancer is perceived as the most common cause of mortality among women globally. Early detection of this disease is critical to reduce significantly the possibility of death. Machine learning techniques have been proved to be efficient and very successful for an accurate breast cancer diagnosis. In this paper, an efficient hybrid Feature Selection (FS) method named a Correlation technique-Modified Grey Wolf Optimizer (CMGWO) was proposed for accurate breast cancer classification based on dimensionality reduction. The suggested technique is based on two stages: the feature selection step and the classification step. Feature selection is the process of picking the most significant characteristics from a dataset. This stage is crucial in machine learning. Firstly, we focus on the filter method by using a Correlation technique for dimensionality reduction. This technique is intended to eliminate and reduce the number of features by selecting one feature from the other correlated features. Secondly, we use the Modified Grey Wolf Optimization algorithm (MGWO) to locate and determine the most significant features from uncorrelated features. After that, we use multiple classifiers to classify breast cancer disease based on the selected features. The Wisconsin Diagnostic Breast Cancer (WDBC) database was used to prove the performance of our proposed work. The experimental results show that the combination of the correlation method and MGWO for feature selection increases the accuracy rate of classification with a minimum number of features. The performances of different machine learning algorithms were evaluated, including Random Forest classifier (RF), Support Vector

doi:10.56415/csjm.v32.10

Machine (SVM) Classifier, and Naïve Bayes (NB) Classifier for the classification step. The suggested technique proves to be the best approach and reliable one among all studied approaches since it increases classification accuracy to 99.12% obtained by CMGWO using Random Forest classifier and demonstrates its significance in detecting breast cancer.

Keywords: Breast cancer, Feature selection, Correlation coefficient, Grey wolf optimizer, Classification.

MSC 2020: 68T09, 68T10, 68T20, 62H30.

1 Introduction

Breast cancer is perceived as the most deadly disease globally [1]. Breast cancer develops in the breast cells. The latter tend to become worse and increase quicker over time, ultimately leading to death. Breast cancer can be treated and avoided in its early stages. However, many women receive cancer symptoms when it is too late. Breast cancer tumors are classified into two classes, benign and malignant. A benign tumor is not harmful to the human body and seldom causes death. A malignant tumor is fatal. This type of tumor evolves quickly because of uncontrolled cell development. Early prevention of breast cancer is essential to heighten the survival chances. For this reason, diagnosis of breast cancer involves accurate identification of breast cancer tumor [2]. Otherwise, to enhance the capability of breast cancer diagnosis classification, researchers over the globe proposed many techniques using machine learning algorithms to obtain the best results and to complete the weaknesses of each other. However, there is a huge opportunity to adopt more efficient breast cancer detection systems. This paper proposes a new method for breast tumor classification using machine learning algorithms. We develop a novel approach that effectively classifies breast cancer. The standard benchmark Wisconsin Diagnostic Breast Cancer (WDBC) database [3] has been used to test and compare the classification performance of the proposed approach with a number of existing studies. The proposition is based on the advantages of the FS process. FS is a preprocessing technique that could largely influence data mining methods [4]. The latter was done to enhance classification accuracy through eliminating unnecessary and insignificant data from original datasets. Thus, FS has become a crucial component in developing machine learning models. In many cases, FS can enhance the performance of a machine learning model as well. In literature, there are three variable selection methods [4]- [7]. Filter methods like Chi-square Test technique, Correlation Coefficient, and Variance Threshold. The second one is the Wrapper methods, for example, Forward Feature Selection technique, Backward Feature Elimination, and Exhaustive Feature Selection. Finally, the embedded methods are like LASSO Regularization (L1) and Random Forest Importance. Among the challenges we encounter when applying a single FS technique is the low accuracy that was obtained for feature subsets selection and that requires artificial analysis of different datasets [5]. On the other hand, high computational cost is another disadvantage of these approaches [5]. In recent years, to solve optimization issues, hybrid algorithms have attracted attention. Hybrid algorithms are the ones that mix many algorithms to create a more effective one in order to address more challenging optimization issues. To enhance the possibility of rapidly and efficiently discovering the best solution and selecting the most important features from a dataset, we propose in this paper a robust hybrid feature selection method. In this method, we use a correlation technique for variable selection. Then, we apply a modified GWO algorithm on chosen attributes to get the most important features from uncorrelated features in order to classify breast cancer disease accurately. The performance of the proposed approach is tested on the WDBC dataset [3] to determine the optimum combination of features that would maximize the classification accuracy. Different classifiers will be trained on the optimized subset of features identified by correlation and modified GWO algorithm. As a result, the main objective of the proposed work is to determine the most significant features from the large dataset that help to make an effective and efficient classification of breast cancer.

This paper is divided into five sections. Section 1 is an introduction to this study. Section 2 discusses relevant studies and several cutting-edge techniques for breast cancer diagnostics. The proposed strategy is presented in Section 3. Section 4 is concerned with experimentation and discussion. Section 5 is the conclusion of the paper.

2 Related works

Many researchers have used machine learning techniques for breast cancer diagnosis to improve the accuracy of classification based on FS techniques. In this section, we have reviewed some relevant studies dealing with different FS methods and machine learning techniques for image classification.

In [6], the authors present a novel approach based on Recursive Feature Elimination (RFE) and SVM classifier. The new Enhanced RFE (EnRFE) method is an improved recursive feature elimination approach. They evaluate RFE with EnRFE on a variety of real datasets and find that EnRFE increases classification accuracy considerably, especially for a small number of features.

An effective heterogeneous image recognition system was introduced in [7]. The multi-model classification technique (MM-CM) and adaptive relevant feature selection are the key components of the proposed system. They use Fisher Linear Discriminant (FLD) to choose the most pertinent features based on the SVM evaluation that was performed. Two image databases including the COREL and CALTECH-256 with a significant number of features are used to evaluate the performance of the suggested image recognition system. The findings obtained show that adaptive feature selection based on MM-CM enhances recognition accuracy in different image databases.

Darzi et al. [8] studied feature selection for breast cancer detection. The suggested techniques involve a wrapper approach based on Genetic Algorithm (GA) and case-based reasoning (CBR). The GA algorithm was used to locate all feasible subsets of attributes, and case-based reasoning was utilized to estimate the evaluation outcome of each subset. The results show that the proposed model performed comparably to the other models on the WBCD dataset. They achieved an accuracy of 97.37% after the feature selection process.

In [9], the authors present an improved and novel strategy combining reliefF and SVM-RFE algorithm to select optimum subset of features performing well in image classification. According to the experimental findings, the suggested relief-SVM-RFE approach greatly improves feature selection in image classification.

Shen et al. [10] studied the SVM algorithm with the Fruit-fly Optimization Algorithm (FOA) in various medical datasets obtained from the UCI repository. The ML SVM technique is combined with Particle Swarm Optimization Algorithm-based SVM (PSO-SVM), Genetic Algorithm-based SVM (GASVM), Bacterial Forging Optimization-based SVM (BFOSVM), and Grid Search Technique-based SVM (Grid-SVM). The SVM-FOA gives the highest accuracy at 96.9% in the Wisconsin dataset.

To overcome the problem of overfitting, Bharat et al. [11] used k-fold cross-validation methods. The different algorithms used are SVM, Decision Tree (CART), NB, and k-Nearest Neighbours (k-NN). They find that SVM with a Gaussian kernel is the best approach for predicting breast cancer accurately.

Jamal et al. [12] worked on two machine learning algorithms, a support vector machine (SVM) and extreme gradient boosting, and compared their performances. For classification, they reduced the number of data attributes by extracting the features with the help of principal component analysis (PCA) and clustering with k-means. In their case, feature extraction is accomplished by changing data from one dimension to another depending on the Euclidian distance between cluster centroids. The metric assessment shows that the dimensionality reduction using the K-means cluster is nearly as excellent as PCA. The experimental findings revealed that sensitivity was the most essential parameter for early diagnosis of breast cancer.

A novel feature selection technique presented in [13], that is based on bee colonies, and a gradient boosting decision tree was proposed. Experiments are carried out using two breast cancer datasets and six public data repository datasets. The experimental findings demonstrate that the suggested strategy efficiently minimizes the dimensions of the dataset and achieves improved classification accuracy.

In [14], the authors suggested a novel breast cancer intelligent diagnosis approach that employed information gain-directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection. The efficacy of the proposed approach is tested on Wisconsin Original Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) breast cancer data sets. The results demonstrate that the proposed method outperforms other works.

A. Rahmani et al. [15] suggested a novel method to enhance the process of breast cancer diagnosis by using the Grasshopper optimization algorithm (GOA) to select the optimal features that are classified by SVM. The experiments for this study were performed on the WBC, WDBC, and WPBC datasets, with accuracy values of 99.51%, 98.83%, and 91.38%, respectively.

In [16], M. Abdel-Basset et al. developed a novel Grey Wolf Optimizer technique coupled with a Two-phase Mutation to tackle feature selection for classification issues using wrapper approaches using a KNN classifier. 35 datasets are used in the studies including the WDBC dataset. Statistical analyses are performed to demonstrate the efficacy and outperformance of the suggested method.

Kumar and Singh in [17], in order to select the optimum subset of features for accurate identification of benign and malignant breast cancer tumors, have proposed an enhanced GWO in combination with SVM applied to the WDBC dataset. Experimental results show that the proposed method improves accuracy to 98,24%.

In [18], a comparison to the performance of support vector machine (SVM)), artificial neural network (ANN), SVM with reduced features, and hybrid SVM-ANN model in the breast cancer diagnosis was carried out. It is found that the hybrid SVM-ANN model gives the best accuracy of 98%.

In [19], feature selection and dimensionality reduction were implemented using principal component analysis and evaluating the correlations between distinct sets of features and their variation. The performances of different machine learning algorithms, including logistic regression, support vector machine, naïve Bayes, k-nearest neighbor, random forest, decision tree, and stochastic gradient descent learning were assessed. The proposed method was tested on the WDBC dataset, and findings show that the proposed approach performed better than other existing works.

In this section, a related work about FS methods was presented. FS is regarded as one of the most important and difficult challenges in machine learning. As we have already seen, it is frequently employed to resolve the issue of dataset dimension reduction in a variety of sectors, particularly the medical one.

According to this state-of-the-art, it is noticed that whatever the FS method is used alone, it is challenging to offer a satisfying answer due to the peculiarities of high-dimensional data space. In the most of previous studies, the authors proposed many approaches to reduce the dimensions of the datasets in order to minimize the computational cost and improve the classification accuracy. However, there is no satisfactory solution, and an optimal solution can't be obtained. Focusing on improving the classification accuracy, a new hybrid FS method has been proposed. A hybrid FS method combines the advantages of the different approaches. Therefore, the search space of the subset of relevant features can be significantly reduced. Indeed, the hybrid algorithm is able to avoid early convergence and more efficiently explore the entire data space when a large amount of noisy data is eliminated. In this context, the proposed hybrid method based on the technique of correlation coefficients and the MGWO algorithm reduces the possibility of falling into the local optimal solution.

3 Proposed method

The process of FS becomes crucial for creating efficient machine learning models. In many situations, FS may help a machine learning model perform better. In this part, we present a new method CMGWO for the classification of breast cancer by applying the FS technique on the WDBC dataset. A hybrid FS technique using the Correlation and Modified Grey Wolf optimizer technique was implemented. The organization of this section is as follows. Firstly, a definition of the Pearson correlation technique was presented. Then, this section demonstrated a description of the base GWO algorithm including the mathematical model and Modified GWO. After that, it provided an explanation of the architecture of the proposed model with a description of the WDBC dataset. Finally, this section addressed in detail the implementation of the correlation technique on WDBC to select the correlated features and the MGWO applied to the selected features.

3.1 Pearson Correlation technique

Pearson Correlation method is a measure of the linear relationship between any two quantitative and categorical variables [20], [21]. We can predict one variable based on the other(s) through correlation. The target should be associated with the variables, but they shouldn't have any relationship with each other. In cases where two characteristics are associated, the model only actually needs one of them since the other one does not provide any new data. The range of the correlation coefficient is between -1 and +1. The correlation's strength rises from 0 to 1. Zero (0) means there is no correlation, while one (1) means there is a perfect correlation.

3.2 Modified Grey Wolf Optimizer

Grey Wolf Optimization algorithm was a new optimization method proposed by Mirjalili et al. [22] in 2014. The main idea of GWO is inspired from the nature and emulated the behavior of hunting of wolves (agents). There are four levels in the social hierarchy in the pack of wolves, alpha (α), beta (β), delta (Δ), and omega (ω), depending on the wolf's participation in the hunting process. Alpha (α) wolf is the dominant wolf in the pack, and his decisions should be respected and followed by the pack members. The second best solution is beta (β) wolf. The third level is occupied by delta (δ) wolf. The last level is omega (ω) wolf. They are the least important individuals in the pack. Grey wolves collaborate in intelligent manner. They follow the prey in a team and try to encircle it and increase the chance of hunt. This process is known as the encircling process. The mathematical model of encircling is defined as follows (Eq. (1) and Eq. (2)):

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A}.\vec{D}, \tag{1}$$

$$\vec{D} = \left|\vec{C}.\vec{X}_p(t) - \vec{X}(t)\right|, \tag{2}$$

where t indicates the current iteration, A and C are coefficient vectors, X_P is the position vector of the prey, and X indicates the position vector of a grey wolf. The vectors A and C are calculated as follows:

$$\vec{A} = 2\vec{a}.\vec{r}_1 - \vec{a}, \tag{3}$$

$$\vec{C} = 2.\vec{r}_2, \tag{4}$$

where $r1$ and $r2$ are random vectors in range $[0, 1]$, a is vector which linearly decreased from 2 to 0 over the course of iterations, see [22]. The second process was the hunting process. It is usually guided by alpha (α), beta (β), and delta (δ) agents. These three wolves are considered as the best solution in the pack and they have better knowledge about the potential location of the prey. Therefore, the three leading search agents are responsible to guide every search agent in the direction of optimal prey. Mathematically, the hunting process is formulated as follows [22]:

$$\vec{D}_\alpha = \left|\vec{C}_1.\vec{X}_\alpha - \vec{X}\right|, \vec{D}_\beta = \left|\vec{C}_2.\vec{X}_\beta - \vec{X}\right|, \vec{D}_\delta = \left|\vec{C}_3.\vec{X}_\delta - \vec{X}\right|; \tag{5}$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1.\vec{D}_\alpha, \vec{X}_2 = \vec{X}_\beta - \vec{A}_2.\vec{D}_\beta, \vec{X}_3 = \vec{X}_\delta - \vec{A}_3.\vec{D}_\delta; \tag{6}$$

$$\vec{X}(t+1) = (\vec{X}_1 + \vec{X}_2 + \vec{X}_3)/3. \tag{7}$$

At the beginning of the method, the GWO starts with a random population of wolves and a searching process guided by alpha (α), beta (β), and delta (δ) wolves. When ($A > 1$), they diverge from each other to search for the best prey (Exploration). In addition, if ($A < 1$), they diverge from each other and force the wolves to attack the prey (Exploitation).

The updating of each search agent's position in the base GWO included the average of the three best grey wolves, alpha, beta, and delta wolf's location (Eq. (7)). This technique produces low-quality solutions. To boost the efficiency of base GWO in [17], the authors proposed a weighted position update technique and changed the Eq. (7) by Eq. (8) which is calculated by Eq. (9):

$$\vec{X}(t+1) = (\vec{X}_1.\vec{W}1 + \vec{X}_2.\vec{W}2 + \vec{X}_3.\vec{W}3)/(\vec{W}1 + \vec{W}2 + \vec{W}3), \tag{8}$$

where $W1$, $W2$, and $W3$ were calculated as follows:

$$\vec{W}1 = \vec{C}_1.\vec{A}_1, \vec{W}2 = \vec{C}_2.\vec{A}_2, \vec{W}3 = \vec{C}_3.\vec{A}_3. \tag{9}$$

3.3 Architecture of the proposed method

In the present research, to increase the accuracy of the classification of breast cancer using a minimum number of features, we divided our work into two stages, including Feature Selection and Classification (see Fig. 1).

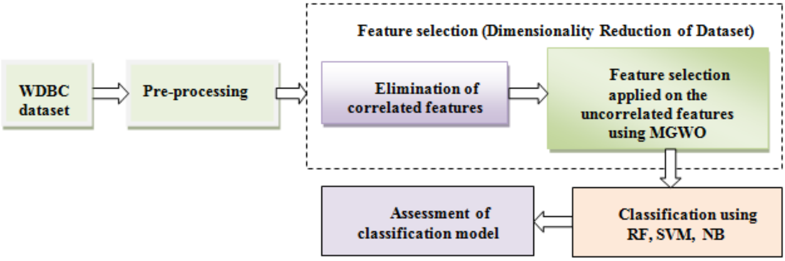
Feature selection (Dimensionality Reduction of Dataset)
WDBC dataset
Pre-processing
Elimination of correlated features
Feature selection applied on the uncorrelated features using MGWO
Assessment of classification model
Classification using RF, SVM, NB

Figure 1. Flowchart of the proposed approach. In our model, the WDBC dataset was used; preprocessing step to cleaning dataset by removing unused features. FS strategy consists of two algorithms: correlation technique and MGWO algorithm. For classification purpose, we use multiple machine learning algorithms including SVM, RF, and NB

The first one combined two techniques for dimensionality reduction (filter method and wrapper method). We selected the non-correlated features from the WDBC dataset by removing correlated features among them and with the target (cancer tumor or not) by applying the correlation technique and reducing the number of features from 30 to 16 (see Section 3.3.2). Then, we passed the non-correlated features to the Modified GWO algorithm to get a minimum number of variables (see Section 3.3.3). Those features represent the most relevant and significant ones for efficient identification in order to reach the best accuracy of classification of breast cancer. Furthermore, we should point out that applying MGWO algorithm to 16 attributes would be better than to 32 attributes. The second stage is the classification the breast cancer by using the SVM Classifier, RF classifier, and NB classifier. Algorithm 1 represents a pseudo code of the proposed method. There are two principal stages. The implementation of the first one consists of three steps to select the best combination of features and get satisfactory results. Then, the classification stage uses multiple machine learning classifiers to classify breast cancer disease.

Algorithm 1.

```
Phase 1:
Input: upload WDBC Dataset
Step1 : Preprocessing and removing unused features from Dataset.
Step2 : Feature selection with correlation technique and removing correlated features from original dataset (see Section 3.3.2).
Step3 : Feature selection applied on uncorrelated features using MGWO algorithm (see Section 3.3.3).
Output : Selected features
Phase 2:
Classification of breast cancer using selected features based on the output of Phase1 and assessment the accuracy of classification.
```

3.3.1 Description of WDBC dataset

We tested the proposed model using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset that is accessible in the UCI Machine Learning Repository [3], [23]. There are 569 cases totally in the dataset, split into two classes. There are 357 instances in the malignant class and 212 cases in the benign class, respectively. 32 attributes are used to represent each record [17], [23]: patient ID, diagnosis, and 30 real-valued attributes. These parameters define the features of cell nuclei as obtained by the digitized FNA images of the breast mass. Ten distinct characteristics of each cell nucleus are represented by the 30 attributes: radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness ($perimeter^2/area - 1.0$), concavity (severity of concave portions of the contour), concave point (number of concave portions of the contour), symmetry, and fractional dimension ("coastline approximation"- 1). There are actually three values listed for each feature: mean value, standard error, and maximum value.

3.3.2 FS using Correlation

As demonstrated in Fig.2, to analyze the correlations between features of the dataset, a heat map was used. A high correlation was observed among “radius-mean”, “parametric-mean”, and “area-mean” features as all these features contain information about the size of breast cancer cells. Therefore, only the “area-mean” feature was selected to further represent the information about the size of breast cancer cell.

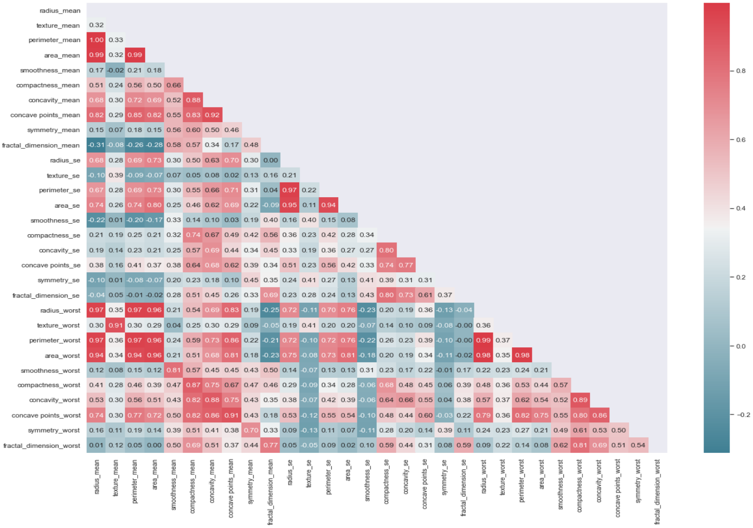

Figure 2. Heat map plot showing the correlations among all features of WDBC

We dropped a total of 14 features: ’perimeter-mean’, ’radius-mean’, ’compactness-mean’, ’concave points-mean’, ’radius-se’, ’perimeter-se’, ’radius-worst’, ’perimeter-worst’, ’compactness-worst’, ’concave-points worst’, ’compactness-se’, ’concave points-se’, ’texture-worst’, ’area-worst’. This way, we had 16 features remaining for further processing. Fig.3 displays the relationships between the chosen features.

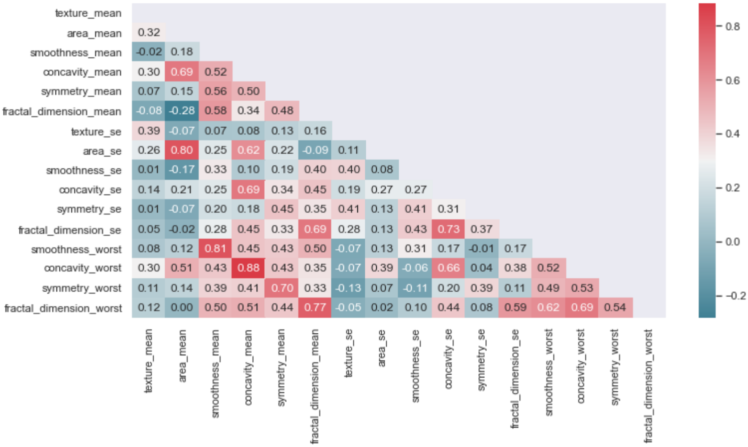

Figure 3. Heat map plot showing the correlations among selected features of WDBC

3.3.3 FS using MGWO

In order to successfully detect the breast cancer tumor in our study, we made use of the strengths of the MGWO algorithm to choose the most pertinent subset of attributes. Algorithm 2 represents a pseudo code of the MGWO algorithm. As we mentioned before, in MGWO, Eq. (8) is utilized in place of Eq. (7) to generate new findings. The Different classifiers were trained using the subset of characteristics that MGWO determined. An example position vector for an alpha search agent of the MGWO algorithm that is utilized for feature selection, is shown in Fig.4. There are two possible values for the solution location, “1” and “0”. To solve a problem with n-dimensions, the position vector would consist of n bits. The property is not picked if the amount is equal to 0. On the other hand, the feature is picked if the value is 1. As a result, the number of 1s in the position vector is exactly the same as the number of features that were chosen (Optimal subset of features). Algorithm 3 explains how MGWO chooses the optimal subset of features. The most important features chosen by applying the MGWO method on uncorrelated features were nine features including texture-mean, area-mean, concavity-mean, symmetry-mean, fractal-dimension-mean, area-se, concavity-se, smoothness-worst, and fractal-dimension-worst.

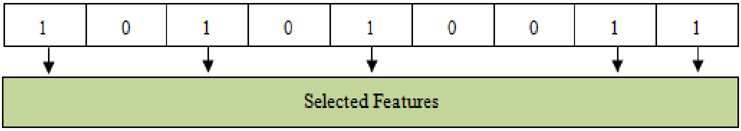

Figure 4. Representation of feature selection technique with MGWO

Algorithm 2.

```
Input:
- Dataset
- Number of features (Dim)
- Population of GWO (Searchagent no)
- Number of Iteration (Max Iteration)

Output:
Minimum number of selected features by MGWO
initialize alpha, beta, and delta positions
Initialize alpha pos, beta pos, and delta pos
Initialize the positions of search agents
For each Iteration
    For each Searchagent no
        - Calculate objective function for each search agent
        - Update Alpha pos, Beta pos, and Delta pos
    end For
    For each Searchagent no
        For each features
            Update the Position of search agents
            including omegas using Equations (1)-(6)
            and Equation (8)
        end For
    end For
end For
return Alpha pos
```

Algorithm 3.

```
For each feature in alpha pos[i] (i=1,2,...,Dim)
if (alpha pos[i] $> 0,5$)
alpha pos[i] =1
Else if (alpha pos[i] $< 0,5$)
alpha pos[i] $= 0$
End if
End for
```

4 Experimental Results

In this research, FS was performed using Correlation technique in conjunction with modified GWO algorithm. The performances of different machine learning algorithms, including SVM, RF, and NB, were evaluated. The suggested method was tested on WDBC dataset. Experimental results were obtained using Python and by fixing the number of iterations for Modified GWO algorithm at 20 iterations with 10 search agents. The algorithm was implemented in PYTHON and run on an Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz with 8GB of RAM.

4.1 Comparison of different performance metrics between different classifiers

In medical research, most existing studies have evaluated performance based only on accuracy evaluation measures. Therefore, we focused not only on accuracy but also evaluated performance based on sensitivity, specificity, precision, and F1-score. Table 1 shows that SVM performed better than other machine learning techniques by sensitivity 93%, specificity 100%, precision 100%, and F1-score 96,4%. On the other hand, we found that RF performed better than SVM and NB by achieving accuracy equal to 99,12%.

Table 1. Comparison of different performance measurements between different classifiers for breast cancer classification using the data of Confusion Matrix.

Evaluation-Mesurement	SVM (%)	RF(%)	NB (%)
Precision	100	97,6	100
F1-score	96,4	95,2	92,5
Sensitivity	93	93	86
Specificity	100	98,6	100
Accuracy	97,4	99,12	96,5

4.2 Comparison of the classification accuracy between CBGWO (Correlation + Base GWO) and CMGWO (Correlation + Modified GWO)

In this part, comparisons are made for classification of breast cancer. As can be seen in Table 2, we can easily observe how the FS step increases the accuracy of the classification of breast cancer. We implement the feature selection process with the proposed approach using the Correlation technique with the Base Grey Wolf optimization algorithm in the first scenario; in the second scenario, we use a Correlation technique with the Modified GWO.

Table 2. Comparison of classification accuracy using proposed approach between CBGWO and CMGWO

Classifiers	Without Feature selection (%)	Correlation + Base GWO (%)	Correlation + Modified GWO (%)
RF	97,07	98,83	99,12
SVM	92,1	92,98	97,36
NB	94,4	93,85	96,5

After feature selection stage, we compared the performances of different machine learning classification techniques for breast tumor classification. Table 2 shows that using Correlation in addition with Modified

GWO in the proposed work gives best result comparing with our novel method based on the original GWO. As shown in the Table, RF outperformed the other classifiers by obtaining accuracy equal to 99,12%.

4.3 Comparison of the classification accuracy between different classifiers using ROC curve (receiver operating characteristic curve)

ROC curve helps to better understand the power of a machine learning algorithm. We can easily observe in Figure 5 that RF is the perfect classifier. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes, and it is used as a summary of the ROC curve. The higher the AUC, the better the performance among classifiers. From Fig.5, we see that RF gives good results compared with SVM and NB classifier in terms of ROC-AUC metric by achieving an AUC criterion equal to 99,3%.

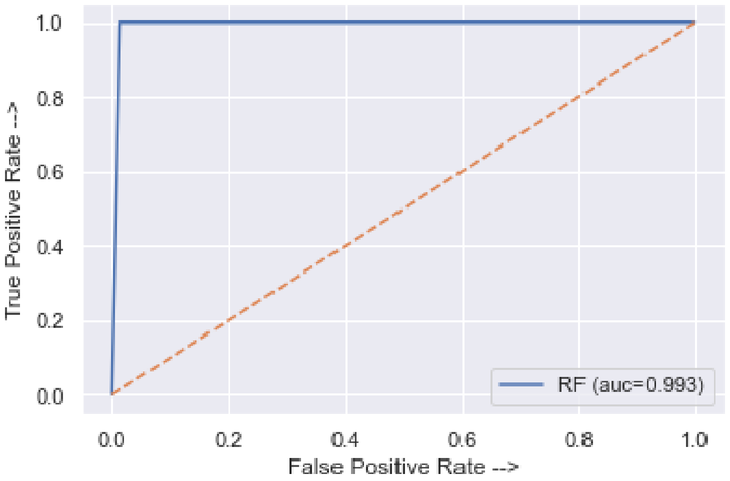

Figure 5. ROC curve metric of RF classifier

The second best classifier was SVM by obtaining 97% as shown in Fig.6. Fig.7 represents the ROC-AUC metric obtaining by NB classifier and achieving 94,6%.

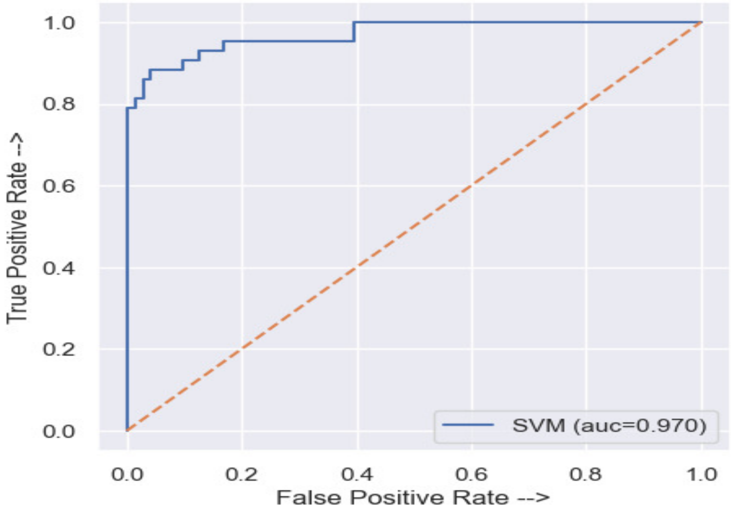

Figure 6. ROC curve metric of SVM classifier

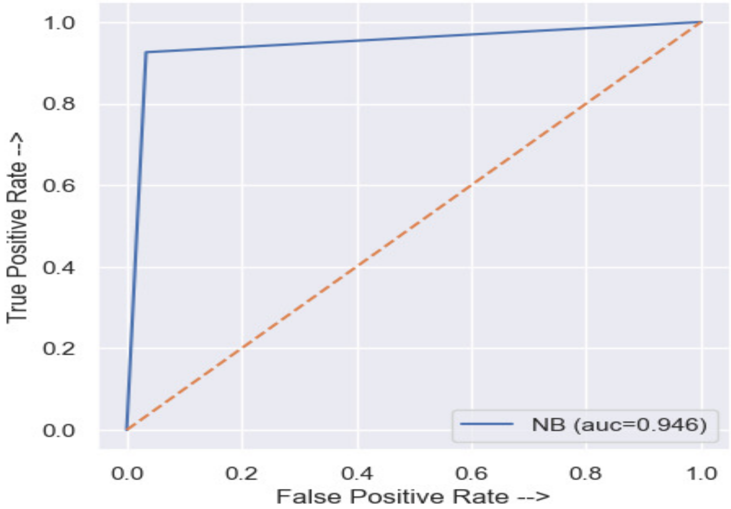

Figure 7. ROC curve metric of NB classifier

4.4 Comparison of the suggested method with existing works

In order to improve the robustness of our proposed method and to predict breast cancer effectively, we compare the performances of different classification models using the WDBC dataset. Table 3 shows a comparison with existing studies for breast cancer identification using ensemble machine learning techniques. From the table, we can see that our method performed better than other works. We have evaluated the performances of three different classification algorithms, i.e., a support vector machine, a naïve Bayes, and a Random forest. As a result, the RF performed better than SVM and NB in terms of classification accuracy.

Table 3. Evaluation of the proposed method by comparing results with existing feature selection methods.

Authors	Feature Selection Technique	Classifier	Accuracy (%)
Darzi et al. [8]	Genetic Algorithm	case-based reasoning (CBR)	97,37
A. Rahmani et al. [15]	Feature Selection with GOA	SVM	98,83
S. Kumar and M. Singh [17]	Feature Selection with Enhanced GWO-SVM	SVM	98,24
Ibrahim et Nazir. [19]	Correlation + Principal Component Analysis	Ensemble machine learning	98,24
Proposed	Proposed-CMGWO	SVM NB RF	97,36 96,5 99,12

5 Conclusion

An extensive research is outgoing on in order to reduce mortality rate due to breast cancer. In this respect, a quick and accurate detection is a critical step in the diagnostic of breast cancer disease. In the present work, we proposed a new method for breast cancer classification; the suggested technique is based on two principal stages. Firstly, we combined the correlation coefficient technique with the Modified GWO algorithm for the dimensionality reduction step. Then, the performance of several machine learning techniques for classification purposes, such as Random Forest, support vector machine, and naive Bayes, was assessed. We reported on the performance of many classifiers using various performance criteria, including accuracy, precision, sensitivity, F1-score, specificity, and ROC-AUC curve. The suggested approach outperformed other efforts according to experimental results by achieving an accuracy of 99,12%, and the value of the AUC criterion was 99,3%. In terms of future projects, to generalize our proposition, we will apply our algorithm to other related datasets such as Wisconsin breast cancer databases including WBC (Wisconsin breast cancer) and WPBC (Wisconsin prognosis breast cancer). This work could be enhanced through the use of parallel methods to increase the accuracy of the classification of breast cancer and reduce computation time.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018, DOI: 10.3322/caac.21492.

[2] F. Ades, D. Zardavas, I. Bozovic-Spasojevic, L. Pugliano, D. Fumagalli, E. de Azambuja, G. Viale, C. Sotiriou, and M. Piccart, "Luminal B Breast Cancer: Molecular Characterization, Clinical Management, and Future Perspectives," *Journal*

of Clinical Oncology, vol. 32, no. 25, pp. 2794–2803, 2014. DOI: https://doi.org/10.1200/JCO.2013.54.1870.

[3] "UCI Machine Learning Repository," University of California, School of Information and Computer Science, Irvine, CA, 2019. [Online]. Available: http://archive.ics.uci.edu/ml.

[4] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (Opatija, Croatia), 2015, pp. 1200–1205, DOI: 10.1109/MIPRO.2015.7160458.

[5] Y. Zhu, T. Li, and W. Li, "An Efficient Hybrid Feature Selection Method Using the Artificial Immune Algorithm for High-Dimensional Data," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID: 1452301, 21 pages, 2022. [Online]. Available: https://doi.org/10.1155/2022/1452301.

[6] X. W. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, (Cincinnati, OH, USA), December 2007, pp. 429–435, DOI: 10.1109/ICMLA.2007.35.

[7] R. Kachouri, K. Djemal, and H. Maaref, "Multi-model classification method in heterogeneous image databases," *Pattern Recognition (Elsevier)*, vol. 43, no. 12, pp. 4077–4088, December 2010. DOI: 10.1016/j.patcog.2010.07.001.

[8] M. Darzi, A. AsgharLiaei, M. Hosseini, and H. Asghari, "Feature selection for breast cancer diagnosis: A case-based wrapper approach," *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, vol. 5, pp. 220–223, 2011. https://api.semanticscholar.org/CorpusID:51751976.

[9] X. Zhou and J. Wang, "Feature Selection for Image Classification Based on a New Ranking Criterion," *Journal of*

Computer and Communications, vol. 3, pp. 74–79, 2015. http://dx.doi.org/10.4236/jcc.2015.33013.

[10] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, D. Liu, "Evolving support vector machines using fruit fly optimization for medical data classification," *Knowledge-Based Systems*, vol. 96, pp. 61–75, 2016, DOI: https://doi.org/10.1016/j.knosys.2016.01.002.

[11] A. Bharat, N. Pooja, and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," in *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, (Bengaluru, India), 2018, pp. 1–4, DOI: 10.1109/CIMCA.2018.8739696.

[12] A. Jamal, A. Handayani, A. Septiandri, E. Ripmiatin, and Y. Effendi, "Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 9, pp. 192–201, 2018, DOI: 10.24843/LKJITI.2018.v09.i03.p08.

[13] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, and L. Gu, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing*, vol. 74, pp. 634–642, 2019, DOI: https://doi.org/10.1016/j.asoc.2018.10.036.

[14] N. Liu, E.-S. Qi, M. Xu, B. Gao, and G.-Q. Liu, "A novel intelligent classification model for breast cancer diagnosis," *Information Processing & Management*, vol. 56, no. 3, pp. 609–623, 2019, DOI: https://doi.org/10.1016/j.ipm.2018.10.014.

[15] A.E. Rahmani, M. Katouli, "Breast cancer detection improvement by grasshopper optimization algorithm and classification SVM," *Revue d'Intelligence Artificielle*, vol.34, no.2, pp. 195–202, 2020, DOI: https://doi.org/10.18280/ria.340210.

[16] M. Abdel-Basset, D. El-Shahat, I. El-henawy, V. H. C. de Albuquerque, and S. Mirjalili, "A new fusion of grey wolf optimizer

algorithm with a two-phase mutation for feature selection," *Expert Systems with Applications*, vol. 139, Article ID: 112824, 2020. [Online]. Available: https://doi.org/10.1016/j.eswa.2019.112824.

[17] S. Kumar and M. Singh, "Breast Cancer Detection Based on Feature Selection Using Enhanced Grey Wolf Optimizer and Support Vector Machine Algorithms," *Vietnam Journal of Computer Science*, vol. 8, no. 2, pp. 177–197, 2021, DOI: 10.1142/S219688882150007X.

[18] Tze Sheng Lim, Kim Gaik Tay, Audrey Huong, Xiang Yang Lim, "Breast cancer diagnosis system using hybrid support vector machine-artificial neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3059-3069, August 2021, DOI: 10.11591/ijece.v11i4.pp3059-3069.

[19] Sara Ibrahim, Saima Nazir and Sergio A. Velastin, "Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis," *J.Imaging*, vol. 7, no. 11, Article No. 225, 2021. [Online]. Available: https://doi.org/10.3390/jimaging7110225.

[20] Ali M. Alsaqr, "Remarks on the use of Pearson's and Spearman's correlation coefficients in assessing relationships in ophthalmic data," *African Vision and Eye Health*, vol. 80, no. 1, Article No. 612, 2021, DOI: 10.4102/aveh.v80i1.612.

[21] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018, DOI: https://doi.org/10.1016/j.tjem.2018.08.001.

[22] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[23] A. B. Yusuf, R. M. Dima, and S. K. Aina, "Optimized Breast Cancer Classification using Feature Selection and Outliers Detection," *Journal of the Nigerian Society of Physical Sciences*, vol. 3, no. 4, pp. 298–307, 2021, DOI: 10.46481/jnsps.2021.331.

Mezaghrani Ali[1], Debakla Mohammed[2],
Djemal Khalifa[3]

Received November 23, 2023
Revised March 4, 2024
Accepted May 2, 2024

[1,2] Faculty of Science Exact, University of Mustapha Stambouli, Mascara, Algeria

[1] Mezaghrani Ali
ORCID: https://orcid.org/0009-0003-7186-624X
E–mail: ali.mezaghrani@univ-mascara.dz

[2] Debakla Mohammed
ORCID: https://orcid.org/0000-0003-3057-2408
E–mail: debakla_med@univ-mascara.dz

[3] Djemal Khalifa
ORCID: https://orcid.org/0000-0002-4959-8205
3IBISC Laboratory, Evry Val d'Essone University Evry, France
E–mail: khalifa.djemal@ibisc.univ-evry.fr