

RESEARCH

Open Access



# Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer

Seyed Reza Kamel<sup>1\*</sup>, Reyhaneh Yaghoubzadeh<sup>1</sup>  and Maryam Kheirabadi<sup>2</sup>

\*Correspondence:  
rezakamel@computer.org  
<sup>1</sup> Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran  
Full list of author information is available at the end of the article

## Abstract

One of the most common diseases among women is breast cancer, the early diagnosis of which is of paramount importance. Given the time-consuming nature of the diagnosis process of the disease, using new methods such as computer science is extremely important for early detection of the condition. Today, the main emphasis is on the science of data mining as one of the computer methods in the field of diagnosis. In the present study, we used data mining as a combination of feature selection method by Gray Wolf Optimization (GWO) and support vector machine (SVM), which is a new technique with high accuracy compared to other methods in this classification, to increase the accuracy of breast cancer diagnosis. The UCI dataset and functional parameters and various statistical criteria were applied to evaluate the proposed method and assess the validity of the results in MATLAB, respectively. Application of the proposed method increased the improvement of the evaluated criteria, which increased the accuracy of diagnosis by 27.68%, compared to former works in the field. As such, it could be concluded that the proposed method had a higher ability to diagnose breast cancer, compared to previous techniques.

**Keywords:** Breast cancer, Feature selection, Support vector machine, Gray Wolf Optimizer

## Introduction

Breast cancer is the second leading cancer among women worldwide, the incidence of which has been reported to increase every year due to factors such as inheritance, lifestyle, and dietary habits [1]. Industrialization and urbanization are among the other contributing factors to the increase incidence of breast cancer. On the other hand, new facilities have been used for the early diagnosis of this chronic disease. The prevalence rate of breast cancer has been reported to be higher in high-income countries, while the incidence rate is on the rise in low-income countries as well, such as the regions in Africa, Asia, and Latin America [2]. In India, the mean age range of the high-risk population for breast cancer has been estimated at 43–46 years, and women aged 53–57 years are reported to be more susceptible to breast cancer [3].

Attempts have been made to reduce the treatment costs of breast cancer and increase the quality of care for these patients by healthcare organizations. Some of the diagnostic and laboratory measures used in the treatment of breast cancer are costly and painful for the patients. Moreover, some of the medications that were initially approved as non-harmful agents to humans have currently shown various side-effects after long-term administration. As a result, the community has become more inclined toward using data mining techniques for the treatment of chronic disease such as breast cancer.

Data mining and data modeling could be exploited for the accurate diagnosis of high-risk cancer patients [4]. In this regard, the most common data mining techniques include random forest, SVM and decision trees, each of which provide variable accuracy rates in the prediction of breast cancer. However, some of these algorithms have limitations (e.g., computational complexity and long runtime), and their accuracy rate could be improved through their meticulous evaluation and combination in order to select their efficient features.

Feature selection is a preprocessing technique that could largely influence data mining techniques. Feature selection is performed to enhance classification accuracy through eliminating unnecessary and insignificant data from original datasets.

In the present study, feature selection was performed using GWO, which remarkably affects breast cancer diagnosis. In “[Review of previous methods](#)” section of the paper, a review of previously applied methods has been presented. In “[Procedure of proposed method](#)” section, the proposed method has been analyzed, and “[Data analysis](#)” section contains the analysis of the results. “[Conclusion](#)” section contains the conclusion of the study.

### **Review of previous methods**

In [5], the study aimed to diagnose breast cancer based on tumor features. Wisconsin Diagnostic Breast Cancer in (WDBC) from the University of California, Irvine, was also used. To extract useful information and diagnose a tumor, a hybrid of K-means and K-SVM was developed. The K-means algorithm is used to detect hidden patterns of benign and malignant tumors. The results not only supported the efficiency of the proposed method in the breast cancer diagnosis but also saved time during the training phase.

In [6], a new multi-layered method that was a combination of clustering and tree decision-making techniques was used to create a prediction system for cancer. The research used data mining technology such as classification, clustering, and prediction to diagnose cancer patients. The collected data was processed prior to processing, imported to the database, and classified to generate significant patterns using the decision tree algorithm. Then the data was divided into cancer and non-cancer data using the K-means clustering algorithm.

In [7], there was an investigation on the use of data mining techniques for the breast cancer diagnosis and treatment. Data was gathered from more than 16,000 patients and data mining techniques were implemented using Weka software. This study illustrates the data mining process and techniques as effective in contributing to the understanding of cancer outcomes as it converts large amounts of patient data into useful information and potentially valuable patterns.

In [8], the study was to investigate the probability of breast cancer as well as the probability of recurrent breast cancer using various data mining techniques. Cancer patient data was collected from Wisconsin dataset of the UCI machine learning. This dataset contains 35 features, which are selected using the feature selection methods and are computed using classification algorithms. According to the results, Naive Bayes algorithm and decision tree provide better and higher accuracy.

In [1], the breast cancer diagnosis was studied using various classification algorithms such as Rep Tree, J48, and Random Forest. The dataset used in this study was extracted from SEER repository with 762,691 samples and 134 features. Furthermore, data cleaning and dimensionality reduction techniques were used and seven features were selected for the final classification. Finally, different data mining techniques were compared and analyzed using the WEKA software. The results revealed that the Rep-Tree algorithm in the set of decision tree algorithms acted better in the breast cancer diagnosis and spent less time on creating the model.

In [3], a model was presented for cancer diagnosis using data mining algorithms. The algorithms used in this study were IBK, SMO, and Bf Tree. The dataset collected in this study was extracted from UCI Irvine repository, which contains 699 records and 9 features. WEKA software was run to compare the techniques, and it was found that the classifications accuracy of the SMO is 91.96%.

In [9], the breast cancer diagnosis was researched by using SVM technique. The benchmarks used in this study are accuracy, ROC, measurement, and computational time of training. The results confirmed the superior performance of the SVM algorithm to other classifications.

In [10], a new core (Hadamard) along with the SVMs was proposed to address the issue of breast cancer prediction using gene expression data. This study revealed the superiority of Hadamard. This core is flexible in terms of variable parameters and is effective in predicting breast cancer in terms of diagnosis.

In [11], a system was proposed for the breast cancer diagnosis based on RepTree, RBF network, and simple logistic using the WEKA software. In the test stage, a ten-fold validation method developed by the Medical Center, Ljubljana Institute of Oncology, Yugoslavia, was proposed for the evaluation of system performance. The accurate classification rate of the proposed system was 74.5%. According to this study, simple logistics can be used for dimensionality reduction and the proposed representative tree and the RBF model of the network can also be employed to reach automatic diagnostic systems for other diseases.

In [2], various data mining techniques were used to predict heart and cancer diseases. The dataset was taken from UCI repository. The results implied that the Naive Bayes algorithm presented in this study provided the highest accuracy in predicting the diseases.

In [12], an automated diagnostic method for breast tumor was proposed using a hybrid of the SVM and a two-stage cluster technique. The UCI dataset contains 699 records and 10 features. This study was performed in MATLAB software. The results revealed that the SVM along with two-stage algorithms could significantly improve the speed of prediction accuracy and reduce subjective classification error in cancer.

In [13], three supervised learning classification algorithms were adopted to predict breast cancer, and they were compared using different parameters. The dataset was extracted from UCI repository. The results of the study suggested that the best technique for breast cancer prediction is to combine the three algorithms using the voting method.

In [14], various data mining techniques were used to predict cancer disease. Moreover, 10 cross-validation methods were also used to measure the unbiased function of the three prediction models for the performance comparison purposes. The instrument used in this study was WEKA. The results (based on the accuracy of breast cancer data) suggested that naive Bayes with an accuracy rate of 97.36% in the holdout sample was the best predictor. This accuracy rate is better than the reported values in the literature. The RBF network ranked second with an accuracy rate of 96.77% and J48 with an accuracy rate of 93.41% was in the third place.

In [15], there was an investigation on the feasibility of using Locality Preserving Projection (LPP)-based feature regeneration algorithm to build a new model of machine learning classification to predict the risk of short-term breast cancer. It was found that the proposed LPP-KNN model provides a higher accuracy in predicting this disease.

In a study [16] had an attempt to diagnose the breast cancer, using twelve data mining algorithms, including J48, Naïve Bayes and Random Forest. The data set used in this study is derived from UCI website, with 699 records and 10 attributes. The criteria discussed in this article are similar to, Kappa statistics, mean absolute error and accuracy. The results indicate that the classification algorithm of J48 is better in terms of precision, and the NB algorithm is weaker in terms of its performance compared to other algorithms.

In an investigation by [17] a fuzzy model is presented in order for evaluating the diagnosis of the breast cancer, one of the characteristics of which is displaying the potential relationship between the variables of age and size of the tumor. This system helps the breast cancer of the patient to be detected as soon as possible. In this study, such criteria as precision, detection, recall and the F criterion are considered, and this method is compared and contrasted to various data mining algorithms, such as Logistic, NB, Tree and so on. The results show that the proposed method is better than other algorithms in terms of precision.

In [18] research is an attempt to predict the period of staying alive among those diagnosed with breast cancer. Data are gathered from 5673 patients at Shiraz University of Medical Sciences. The proposed algorithm is Logistic in this study. The study has achieved the highest level of sensitivity and detection criteria, and it has used the smallest set of attributes like age, tumor size, the ratio of lymphatic glands involvement, invasion and so on. The results demonstrate that although the effect of age on the patient is a controversial matter, aging leads to a 60-month decrease of survival.

### **Procedure of proposed method**

In the present study, a method consisting of four main phases was proposed for the diagnosis of breast cancer using the GWO as feature selection (Fig. 1).

As is depicted in Fig. 1, the proposed method encompassed four main phases, which will be further discussed in the following sections.

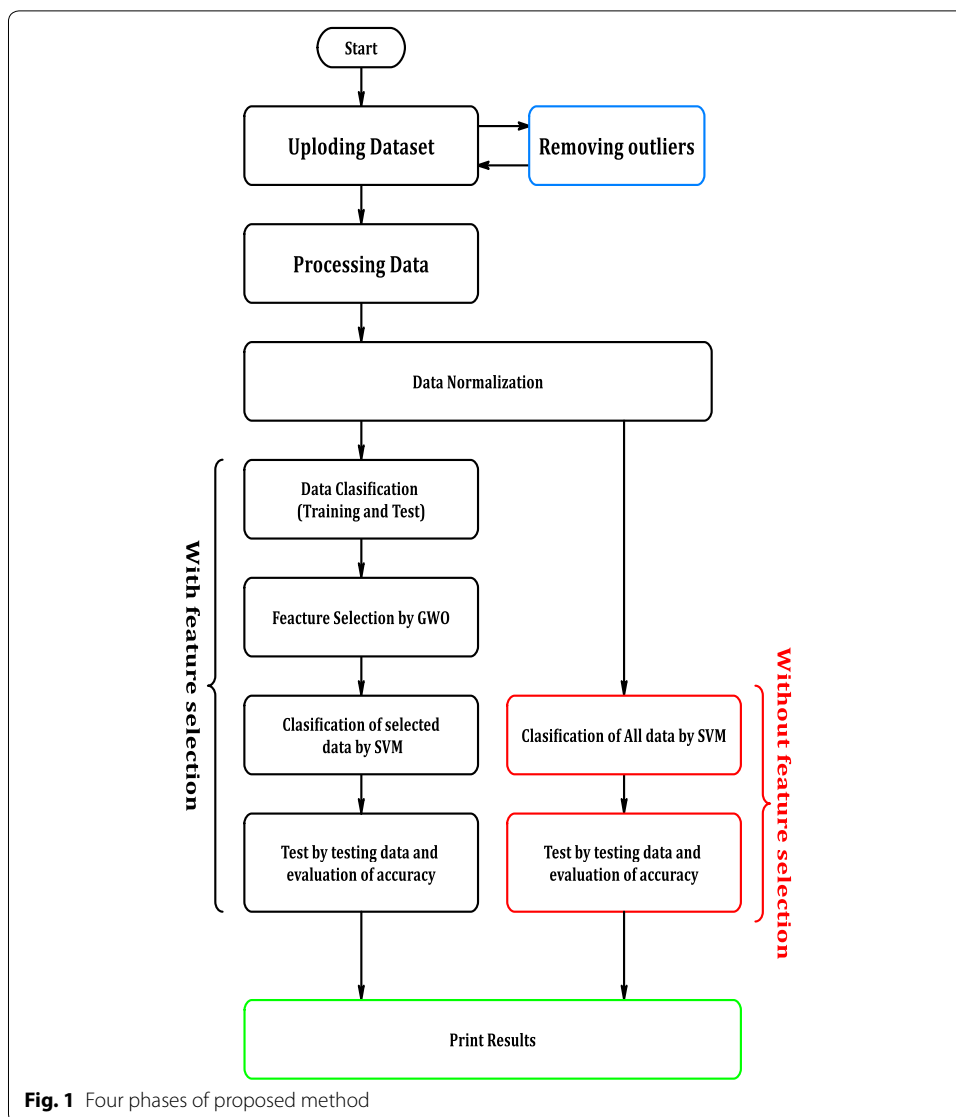


Fig. 1 Four phases of proposed method

**Phase I: data preprocessing**

Initially, data refinement was performed. In addition, data clearing and filtering were carried out in order to prevent the formation of inappropriate rules and patterns. In this paper, the breast cancer dataset was preprocessed, and the outliers were eliminated using the outer line method.

**Phase II: data normalization and classification**

In this phase, repetitive records were eliminated using the min–max normalization method. Afterwards, the data were divided into two groups of training and testing. At this stage, 16 records were excluded from 699 records, and the remaining data were divided into two categories of training and testing. Following that, 80% of the data (samples) was randomly assigned to the training group, and the remaining 20%

(samples) was selected for the testing group. The training data were trained using the proposed SVM–GWO method and tested based on the testing group data that were not used in the training phase.

**Phase III: use of GWO for feature selection and classification**

In this phase, feature selection was employed to detect the significant features of a particular result using GWO. Following that, SVM was used for the classification and evaluation of the benchmarks.

**Gray Wolf Optimization (GWO)**

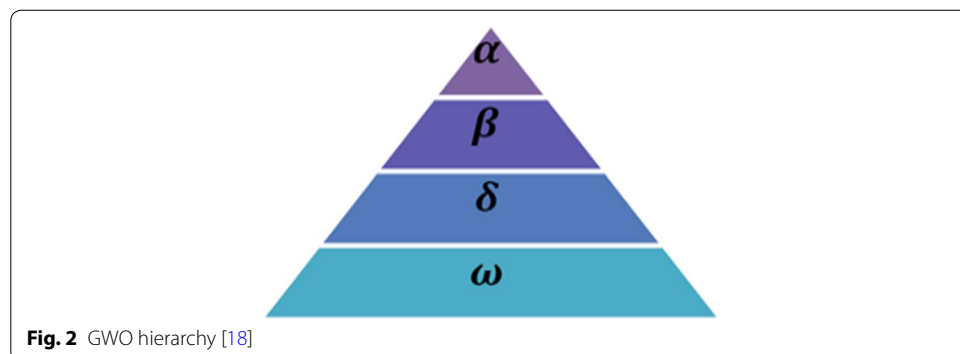
GWO [19] is a metaheuristic algorithm inspired by the hierarchical structure and social behavior of the wolves in hunting. This algorithm is population-based with a simple setup process and could be easily extended to large-scale issues. GWO was introduced in 2014. Each herd has four main ranks that are modeled as a pyramid (Fig. 2) [19].

As can be seen in Fig. 2, alpha wolves (the best selective features) dominate beta wolves (features less valued than the best features). The beta wolves assist the alpha wolves in the decision-making process and have the potential to be replaced by the alpha wolves [18]. Moreover, the alpha wolves dominate the delta wolves, including old wolves, hunters, and wolves taking care of the baby wolves (i.e., factors specifying superior features). On the other hand, omega wolves are affected by all the high-ranking wolves. In fact, they have a minimum right compared to the other group members. In GWO, this category includes the minor features or all the features. Omega wolves eat after the other members of the herd and are not involved in decision-making processes [19]. The hierarchical structure and social behavior of gray wolves are divided into three main phases, as follows:

1. Tracking and approaching;
2. Encircling;
3. Attacking.

GWO contains a significant parameter known as *a*, the value of which is within the range of 0–2. This value and two random parameters that are subsequently generated could be used to update the coefficient vectors *A* and *C* at each stage [19].

$$a = 2 - t \frac{2}{\text{MaxIter}} \tag{1}$$



**Fig. 2** GWO hierarchy [18]

$$\vec{A} = 2a \cdot \vec{r}_1 - a \quad (2)$$

$$\vec{C} = 2\vec{r}_2 \quad (3)$$

Given the boundary of each wolf group, three responses are achieved, according to which the final response is specified. In other words, the three first groups of the wolf hierarchy participate in decision-making operations based on Eq. 5 [19].

$$\vec{D} = \left| \vec{C} \cdot \vec{X}_p(t) - \vec{X}(t) \right| \quad (4)$$

$$\begin{aligned} \vec{X}_1 &= \left| \vec{X}_\alpha - \vec{A}_1 \cdot \vec{D}_\alpha \right| \\ \vec{X}_2 &= \left| \vec{X}_\beta - \vec{A}_2 \cdot \vec{D}_\beta \right| \\ \vec{X}_3 &= \left| \vec{X}_\delta - \vec{A}_3 \cdot \vec{D}_\delta \right| \end{aligned} \quad (5)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (6)$$

where  $t$  indicates the current iteration,  $\sim A$  and  $\sim C$  are coefficient vectors,  $\sim a$  are linearly decreased from 2 to 0 over the course of iterations and  $r_1, r_2$  are random vectors in  $[0, 1]$ .

#### Phase IV: data classification and comparison of the results

Data were classified using algorithms svm, in this phase, the selected data and all data are obtained results were analyzed and compared.

#### Data analysis

In this section, we assessed the data and analyzed the results and the evaluated criteria.

#### Datasets

Datasets used in this study were selected from UCI reservoir [20] and included 699 records, 9 features, and two cancer classes (0 class related to healthy individuals and 1 class related to breast cancer patients) (see in Table 1).

#### Preferences and criteria for comparison

The results must be assessed by a set of criteria to evaluate the performance of the proposed method. In general, the confusion matrix is exploited to evaluate the position and efficiency of disease classification and diagnosis systems. Analysis of confusion matrix in classification and diagnosis of diseases lead to four modes of positive truth, negative truth, positive false and negative false. Table 2 shows the position of expressed parameters in the confusion matrix.

That:

TP: true positives: number of examples predicted positive that are actually positive.

FP: false positives: number of examples predicted positive that are actually negative.

**Table 1 Dataset input features**

Column	Feature	Description
1	Mass thickness	Benign cells tend to be in one layer, whereas cancer cells tend to be in multiple layers
2	Uniformity of cell size	Cancer cells have different sizes while benign cells are similar in size
3	Cell shape uniformity	Cancer cells have different shapes, whereas benign cells are similar in shape
4	Marginal bond	Benign cells apt to attach mutually but cancer cells tend to misplace this capability
5	Epithelial cell size	Epithelial cells that are significantly large may be a malignant cell
6	Bare nuclei	Bare nuclei are a term used for nuclei that are not surrounded by cytoplasm. They are usually observed in benign cells
7	Bland chromation	In benign cells, the nucleus (chromatin) is in the form of uniform texture, but cancer cells are not this way
8	Normal nucleoli	Normal nucleoli are small structures observed in cell nucleus. These nuclei are very small in benign cells (observable). However, they are large and distinctive in cancer cells
9	Mitosis division	Mitosis is the division of the cell nucleus into two daughter cells. Pathologists calculate the grade of cancer by counting the number of mitosis divisions

**Table 2 Confusion matrix**

Actual values	Predicted values	
	Unhealthy	Healthy
Unhealthy	TP (true positives)	FN (false negatives)
Healthy	FP (false positives)	TN (true negatives)

FN: false negatives: number of examples predicted negative that are actually positive.

TN: true negatives: number of examples predicted negative that are actually negative.

The results must be assessed by a set of criteria to evaluate the performance of the proposed method. Presented in Eqs. 7 to 15, the set of criteria is used as the most important evaluation criteria.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity = \frac{TN}{FP + TN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$F\_measure = \frac{2 * Recall * Precision}{Precision + Recall} \tag{11}$$

$$RMSE = \sqrt{\sum_{t=1}^n (y - \hat{y})^2 / n} \tag{12}$$



$$MSE = \sum_{t=1}^n (y - \hat{y})^2/n \tag{13}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

$$Kapa = \frac{[PA - PE]}{[1 - PE]} \tag{15}$$

$$PE = (TP + FN) * (TP + FP) * (TN + FN)/N^2$$

$$PA = (TP + TN)/N$$

**Results and discussion**

In the present research, the main objective was improving the performance of classification and increasing diagnosis accuracy by reducing the dimensions of features using the GWO. Different situations were considered based on the following scenarios to more accurately assess the proposed method.

1. First scenario: 60% of data for training and 40% remaining for testing.
2. Second scenario: 70% of data for training and 30% remaining for testing.
3. Third scenario: 80% of data for training and 20% remaining for testing.

- Selected features

The features selected by GWO, which were evaluated in the three scenarios, are presented in Table 3.

As observed in Table 3, number 1 is indicative of selecting a feature, whereas number 0 shows the lack of selection of the feature. With regard to the selected features, the results of the proposed method in each scenario are shown in Table 4.

As seen in Table, the performance of GWO algorithm improved in selecting an optimal subset of features and SVM class and the evaluation criteria had a better performance in the third scenario considering that more data were included for training. All criteria evaluated are described, as follows:

- Accuracy criterion

Accuracy with feature selection method by GWO and without feature selection is compared and the results are shown in Fig. 3. According to xmpirical results, cancer data with feature selection had a higher accuracy. The GWO algorithm only requires a

**Table 3 Selected features**

Features	1	2	3	4	5	6	7	8	9	10
First scenario	0	1	1	1	1	1	0	1	0	0
Second scenario	1	1	0	1	1	0	0	1	0	1
Third scenario	0	1	0	0	0	0	1	1	1	1



two-parameter initialization compared to algorithms such as Bat algorithm. In addition, this algorithm has high convergence speed and is capable of generalization to n-dimensional space. Therefore, the use of this algorithm in the selection of an optimal subset of features along with the SVM provides high accuracy in this regard.

- ROC curve

This curve is a general and valid criterion for assessing classification performance, which is mainly assessed on the basis of two criteria of detection and recall/sensitivity. In this respect, detection and recall/sensitivity are used as criteria in the performance of the negative and positive sections, respectively. In the ROC diagram, the true positives (TP) are in the horizontal axis while the false positives (FP) are in the vertical axis. The ROC curve acts more accurately in determining whether a person has breast cancer or not, compared to other criteria. In our proposed method, the curve is established with feature selection and without feature selection, as shown in Figs. 4 and 5.

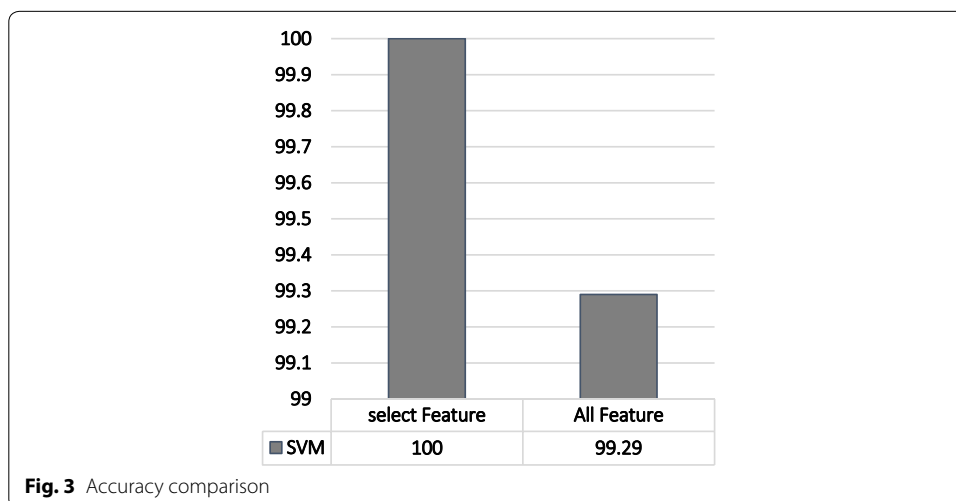
- AUC criterion

This criterion contributes to the provision of a general measure of the performance of all classes. Therefore, it ranges from 0 to 1. In the proposed method, the value of this criterion was 1 or 100%.

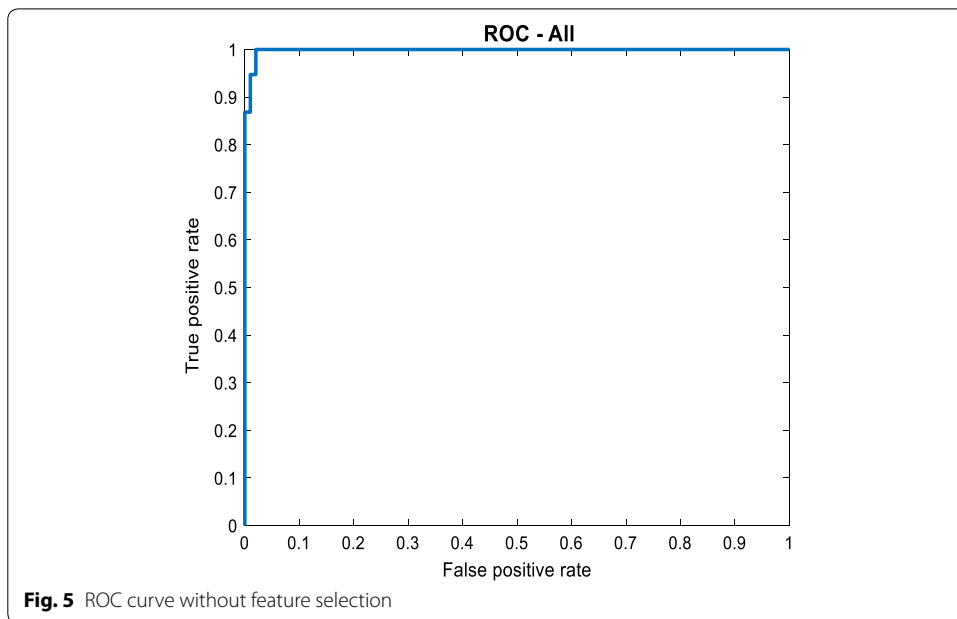
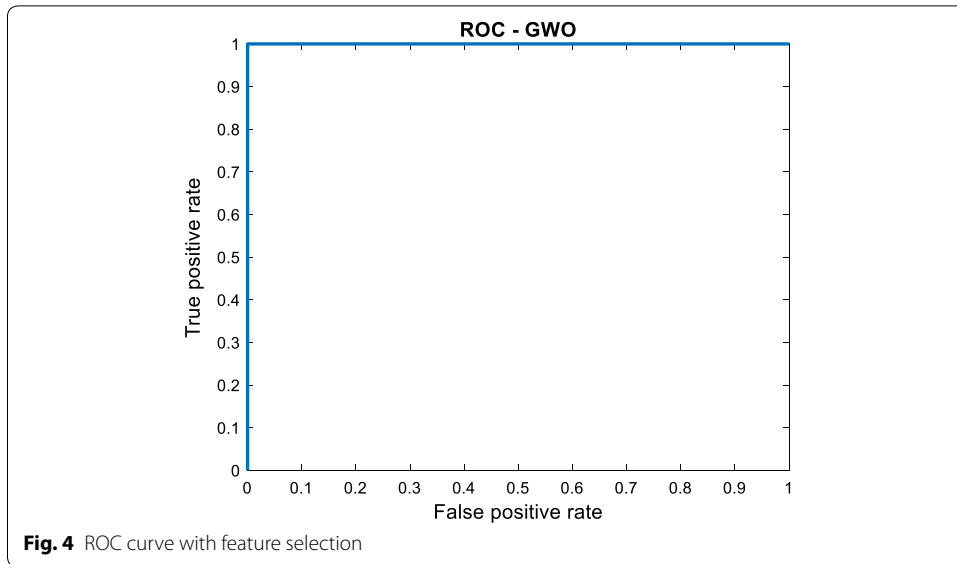
- Mean square error

Mean square error is one of the statistical tools to find prediction accuracy in modeling. In the proposed method, the extent of this criterion for the three scenarios is presented in Table 5.

As observed in the table, the lower the mean square error, the higher the accuracy of our model. According to the results obtained in the third scenario, the prediction accuracy of our model was higher in the mode of feature selection.



**Fig. 3** Accuracy comparison

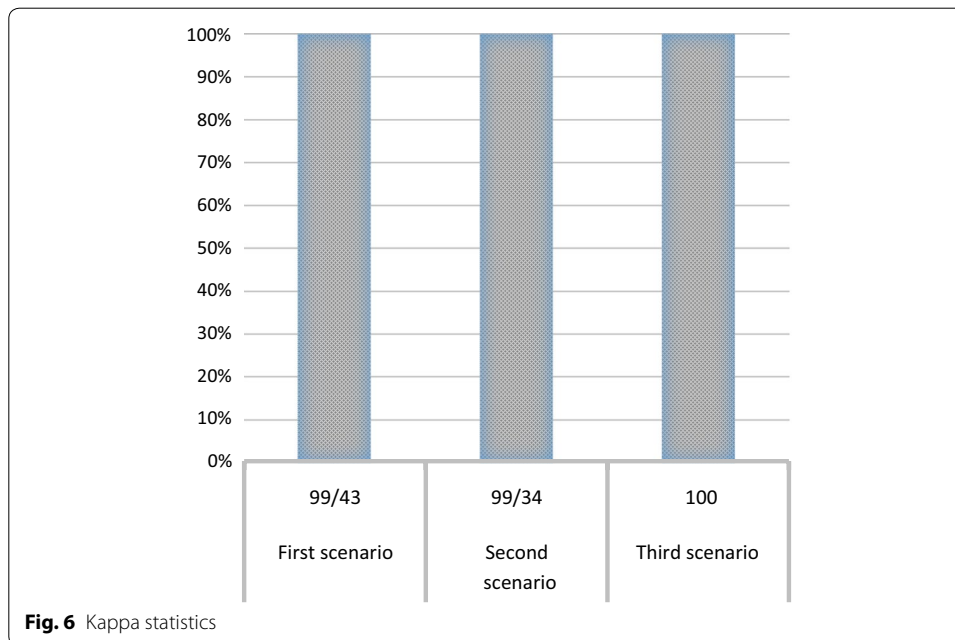


**Table 5** Mean square error

	With selected features	Without selected features
First scenario	6.0357	5.8893
Second scenario	6.1571	6.0905
Third scenario	5.5357	5.6214

**Table 6** Root mean square error

	With selected features	Without selected features
First scenario	2.4568	2.4268
second scenario	2.4814	2.4679
third scenario	2.3528	2.371



**Fig. 6** Kappa statistics

- Root mean square error (RMSE)

The root mean square error (RMSE) for the proposed scenarios is presented in Table. This parameter is more used to estimate the difference between the predicted values using the model with observed values. The accuracy of the model is higher when the mean square error for a specific model is lower than another model. In the proposed method, the extent of this criterion for the three scenarios is presented in Table 6.

- Kappa statistics

The Kappa statistics is a criterion that compares the expected accuracy with the observed accuracy. The results of calculating this criterion in the three scenarios are presented in Fig. 6.

For further evaluation, the proposed method was compared with several algorithms used in the diagnosis of breast cancer.

As observed in Table 7, the accuracy, sensitivity and detection criteria in the proposed method have had considerable progress in the diagnosis of breast cancer,

**Table 7 Comparison of the results for the proposed method**

Algorithm	Research	% accuracy	% sensitivity	% specificity
PSO-SVM	[21] in 2019	93	93	94
SVM	[22] in 2013	93	92	100
SVM	[23] in 2018	85.89	86	76
Proposed model	–	100	100	100

compared to other algorithms, due to the use of the GWO algorithm in selecting a useful and optimal subset of features over the years.

## Conclusion

In this paper, the GWOs and SVM were used for feature selection and data classification in order to increase the accuracy of breast cancer diagnosis, respectively. Experimental results were obtained using the MATLAB and UCI dataset. The best results were obtained from a hybrid of the SVM algorithm and the GWO to select the subset of the efficient features. In the proposed method, the accuracy, sensitivity, and specificity were 100, 100, and 100, compared to the other algorithms. Future studies can adopt this perspective in the diagnosis of heart diseases, diabetes, and others.

### Abbreviations

GWO: Grey Wolf Optimizer; SVM: support vector machine.

### Acknowledgements

Not applicable.

### Authors' contributions

All the corresponding authors contributed equally to the conduct of the present study. All authors read and approved the final manuscript.

### Funding

Not applicable.

### Availability of data and materials

<https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin>.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup> Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran. <sup>2</sup> Department of Computer, Neyshabur Branch, Islamic Azad University, Neyshabur, Iran.

Received: 28 May 2019 Accepted: 9 September 2019

Published online: 16 October 2019

## References

- Hamsagayathri P, Sampath P. Decision tree classifiers for classification of breast cancer. *Int J Curr Pharm Res*. 2017;9(2):31.
- Osman AH. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. *Int J Adv Comput Sci Appl*. 2017;8:158–65.

3. Chaurasia V, Pal S. A novel approach for breast cancer detection using data mining techniques. In: International journal of innovative research in computer and communication engineering (an ISO 3297: 2007 certified organization), vol. 2; 2017.
4. Shawe-Taylor J, Sun S. A review of optimization methodologies in support vector machines. *Neurocomputing*. 2011;74(17):3609–18.
5. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst Appl*. 2014;41(4):1476–82.
6. Ramachandran P, Girija N, Bhubaneswar T. Early detection and prevention of cancer using data mining techniques. *Int J Comput Appl*. 2014;97(13):48–53.
7. Lu J, Keech M. Emerging technologies for health data analytics research: a conceptual architecture. In: 2015 26th international workshop on database and expert systems applications (DEXA). IEEE; 2015. p. 225–9.
8. Pritom AI, Munshi MAR, Sabab SA, Shihab S. Predicting breast cancer recurrence using effective classification and feature selection technique. In: 2016 19th international conference on computer and information technology (ICCIT). New York: IEEE; 2016. p. 310–4.
9. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. *PLoS ONE*. 2017;12(1):e0161501.
10. Chaurasia V, Pal S. Data mining techniques: to predict and resolve breast cancer survivability. *Int J Comput Sci Mob Comput (IJCSMC)*. 2014;3(1):10–22.
11. Chaurasia V, Pal S. Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease. *Rev Res*. 2014;3(8).
12. Kumar UK, Nikhil MS, Sumangali K. Prediction of breast cancer using voting classifier technique. In: 2017 IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM). New York: IEEE; 2017. p. 108–14.
13. Vijaya Lakshmi IV, Krishnaveni G. Performance assessment by using SVM and ANN for breast cancer mammography image classification. *Int J Eng Technol Sci Res*. 2017;4(9):620–6.
14. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. *J Algorithms Comput Technol*. 2018;12(2):119–26.
15. Heidari M, Khuzani AZ, Danala G, Mirniaharikandehi S, Qian W, Zheng B. Applying a machine learning model using a locally preserving projection based feature regeneration algorithm to predict breast cancer risk. In: Medical imaging 2018: imaging informatics for healthcare, research, and applications, vol. 10579. International Society for Optics and Photonics; 2018. p. 105790T.
16. Kumar V, Mishra BK, Mazzara M, Verma A. Prediction of malignant & benign breast cancer: a data mining approach in healthcare applications. 2019. arXiv preprint [arXiv:1902.03825](https://arxiv.org/abs/1902.03825).
17. Dutta S, Ghatak S, Sarkar A, Pal R, Pal R, Roy R. Cancer prediction based on fuzzy inference system. *Smart innovations in communication and computational sciences*. Springer: Singapore; 2019. p. 127–36.
18. Nourelahi M, Zamani A, Talei A, Tahmasebi S. A model to predict breast cancer survivability using logistic regression. *Middle East J Cancer*. 2019;10(2):132–8.
19. Mirjalili S, Mirjalili SM, Lewis A. Grey wolf optimizer. *Adv Eng Softw*. 2014;69:46–61.
20. <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin>.
21. Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans Comput Biol Bioinform (TCBB)*. 2019;16(3):841–50.
22. Kumar GR, Ramachandra GA, Nagamani K. An efficient prediction of breast cancer data using data mining techniques. *Int J Innovat Eng Technol (IJET)*. 2013;2(4):139.
23. Sangaiyah I, Kumar AV. Improving medical diagnosis performance using hybrid feature selection via relief and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction. *Clust Comput*. 2018. <https://doi.org/10.1007/s10586-018-1702-5>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

---