Publisher homepage: www.universepg.com, ISSN: 2663-7804 (Online) & 2663-7790 (Print)

https://doi.org/10.34104/ajeit.024.070078

Australian Journal of Engineering and Innovative Technology

Journal homepage: www.universepg.com/journal/ajeit

A Method Based on Process Mining for Breast Cancer Diagnosis with Whale Optimization Algorithm and Support Vector Machine

Ali Mohammadiounotikandi[1]* and Somayeh Babaeitarkami[2]

[1]Department of Computer and IT Engineering, Faculty of Engineering, South Tehran Branch, Islamic Azad University (IAU), Tehran, Iran; and [2]Faculty of Art and Architecture, South Tehran Branch, Islamic Azad University (IAU), Tehran, Iran.

*Correspondence: ali.mohammadion@gmail.com (Ali Mohammadiounotikandi, Department of Computer and IT Engineering, Faculty of Engineering, South Tehran Branch, Islamic Azad University (IAU), Tehran, Iran).

ABSTRACT

Breast cancer is the second most common cancer among women and the second leading cause of death in the world. According to the statistics of the National Cancer Center, one out of every eight women in the United States is diagnosed with breast cancer. This cancer is the most common malignancy among Iranian women and the main focus of attention in Iran. The data shows that in recent years, the prevalence of the disease has been growing. All tumors are not cancerous and may be benign or malignant. Benign tumors grow abnormally but are rarely fatal. However, some benign breast masses can also increase the risk of breast cancer. The process mining is one of the methods used to diagnose or predict cancers. This method is one of the most popular approaches to breast cancer diagnosis. Process mining approaches can help doctors in better detection of breast cancer by reducing the number of false positive and negative results. The whale optimization algorithm is one of the new meta-heuristic algorithms and imitates the behavior of whale hunting. This algorithm starts with a set of random solutions, in each iteration the search agents update their position according to each of the search agents randomly or with the best solution obtained so far. In this research, using the whale algorithm method, a method to reduce cancer diagnosis error in a number of patients with 9 types of contamination has been investigated and presented. Therefore, in this research, with the help of MATLAB software and using the advantages of whale algorithm optimization, this number of diseases has been categorized, as a result of which the diagnosis error is reduced.

Keywords: Process mining, Cancer tumors, Breast cancer, Whale algorithm, and Support vector machine.

INTRODUCTION:

Breast cancer is the second most common cancer among women and the second leading cause of death in the world. Every year, more than 11,000 women around the world die due to this disease. According to the statistics of the National Cancer Center, one out of every eight women in the United States is diagnosed with breast cancer. Six percent of all deaths in the world are caused by this disease. The number of breast cancer patients in India is one in every 22 women. This cancer is the most common malignancy among Iranian women and the main focus of attention in Iran. In recent years, the prevalence of the disease has been growing and the data shows that the survival rate of patients up to five years after diagnosis was 88% and 10 years after diagnosis was 80%. In fact, all tumors are not cancerous and may be benign or malignant. Benign tumors grow abnormally but are rarely fatal. However, some benign breast masses can also increase the risk of breast cancer. Also, the risk of breast cancer has increased in some women with a history of biopsy of benign breast masses. Breast cancer is caused by abnormal growth of abnormal cells in the breast. Many factors are used to predict, diagnose and treat this disease, such as the presence of a tumor, the involvement of lymph nodes, indentation of the nipple, the occurrence of secretion in the breast, etc. The presence of similarities in the clinical and laboratory symptoms of breast cancer increases the possibility of misdiagnosis. A lump is the most common symptom of breast cancer, which is discovered accidentally by the patient in most cases, and is identified by the doctor during a clinical examination in the rest of the cases. This mass may be painful, but in most cases it is painless. In some cases, breast cancer appears as multiple masses. The similarity of clinical and laboratory symptoms of breast cancer increases the possibility of errors in diagnosis (Varsha *et al.,* 2022 & Essam *et al.,* 2022)

Process mining has been used to discover learning traces by identifying learning trends by identifying bottlenecks, analyzing performance, and checking compliance and improving learning processes by recommending an appropriate trace (Dallagassa *et al.,* 2022). The focus of the studies is on different organizational settings (e.g. e-learning, training courses and law) and a conformity assessment approach will be used to measure the effect of metacognitive stimulation on the learning process, which is the subject of performance evaluation. The findings show the benefit of assessment where the sequential structure of learning processes is of interest. Process research is a business process management approach that is based on the analysis of process implementation in the real world to help managers in making decisions like a decision support system. In other words, process mining is a re-engineering of data-driven business processes. Currently, there are many techniques and tools for improving business processes, with the aim of process mining. These tools often include methods for discovering, analyzing and improving business processes, using data recorded in event graphs. In fact, unlike other business process management techniques, process mining is truth-oriented. In other words, process mining techniques, by following the process of executing processes in the real world, draw the model of the process implemented in reality. In this sense, it uses the data in the pictures, which were produced and recorded by the information systems during the execution of that process. After discovering the real model of the process and matching it with the original model, it identifies, analyzes and improves the bottlenecks and optimal paths. In fact, process mining is the meeting point of process model-based analyzes (such as simulation) and data-oriented analyzes (such as data mining) and answers the questions related to the process from the two perspectives of efficiency and compliance (De Roock & Niels: 2022 & Munoz-Gama *et al.,* 20222; Begum *et al.,* 2024)

Process discovery technique

This technique observes examples of system behaviors by examining the event diagram. In this sense, it first identifies the process under investigation and extracts its activities along with the time and resource of the activity. Then, with the help of one of the process discovery methods (such as alpha algorithm), it draws a real process model. This model shows the implemented process from the perspectives of work flow (procedure from beginning to end), social network (their roles and impact on the process), time, resources (executives) and item (file).

Adaptation technique

After discovering the model from the event diagram, the adaptation technique compares the process model obtained from the event diagram with the similar business process. The main task of matching techniques is to compare the alignment between the model and reality. Generally, they have a control aspect and check that the discovered model is real, registered in the diagram and are in accordance with the similar business model. In other words, this technique controls the degree of conformity of the business process model with the events registered in the database. In this way, the defects, errors, deviations and bottlenecks of the process are discovered along with their intensity and the desired path of the users is determined.

Upgrade technique

Upgrade and development techniques are used to expand or improve a process model. These intelligent and fact-based techniques examine the resulting information, correct the weak points and highlight the strong points, and produce a new process model. The process mining in practice is data science. This science tries to manage a process by using the available data analysis. In other words, process mining acts as a bridge between data mining analysis and process model analysis in order to manage business processes. Two life cycle models of business process management and CRISP cycle can be the basis of the process analysis model. It is possible to diagnose and predict all kinds of diseases using process mining techniques. Process mining in medicine is the process of extracting previously unknown, understandable and reliable information from medical databases and using it to predict, diagnose and help treat diseases. Discovering useful patterns between the disease and the patient's clinical and laboratory symptoms is one of the applications of process mining in medicine. A useful model is a model in data that expresses the relationship between a subset of patient data and disease diagnosis. Process mining is one of the methods used to diagnose or predict cancers. This method is one of the most popular approaches to breast cancer diagnosis. Process mining approaches can help doctors in better detection of breast cancer by reducing the number of false positive and negative results. One of the process mining algorithms is support vector machine (SVM). In the conducted research, SVM has been used to diagnose breast cancer and it has high classification accuracy compared to other existing artificial intelligence methods. SVM is a machine learning technique that was first introduced by Vepenek. SVM finds the maximum margin between two classes in a specific feature space. The Whale Optimization Algorithm (WOA) is one of the new meta-heuristic algorithms and imitates the hunting behavior of whales. This algorithm starts with a set of random solutions, in each iteration the search agents update their position according to each of the search agents randomly or with the best solution obtained so far. In this research, we aim to provide a process mining based method for breast cancer diagnosis with the help of whale optimization algorithm and support vector machine.

Problem Modeling and Formulation

Whale optimization algorithm

One of the largest mammals in the world is a whale. Among the 7 whales in the world, the most famous is the humpback whale. A mature humpback whale is about the size of a school bus. Favorite prey is whales, krill and groups of small fish. The most interesting thing about humpback whales is their special hunting method. This exploratory behavior is known as bubble-net feeding method. Humpback whales prefer to hunt groups of krill or small fish near the surface of the water. It has been observed that this exploration and hunting is done by creating indicator bubbles along a circle or paths. The WOA algorithm is one of the nature-inspired and population-based optimization algorithms that can be used in various fields (Yang *et al.,* 2022; Chakraborty *et al.,* 2022)

Wall WOA algorithm is performed in three stages or three phases, which are as follows:

- Siege hunting
- Exploitation Phase: The method of attacking the web bubble
- Discovery Phase: Search for prey

Encirclement hunting in the whale algorithm: whales can identify hunting locations and surround them. Since the location of the optimal design in the search space is not known through comparison, the algorithm assumes that the current best candidate solution is hunting the target or is close to the optimal state. After the best search agent is identified, other search agents try to update their location relative to the best search agent. This behavior is expressed through equations 1 and 2:

$$\vec{D} = \left| \vec{C}.\vec{X}^{*}(t) - \vec{X}(t) \right| \qquad (1)$$

$$\vec{X}(t+1) = \vec{X}^{*}(t) - \vec{A}.\vec{D} \qquad (2)$$

Where *t* represents the current iteration, *A* and *C* are coefficient vectors, X^* is the location vector of the currently obtained best solution, and X is the location vector. It should be noted that if there is a better solution, X^* should be updated in each iteration. Two vectors *A* and *C* are calculated as follows:

$$\overrightarrow{A} = 2\overrightarrow{a}.\overrightarrow{r} - \overrightarrow{a} \qquad (3)$$

$$\overrightarrow{C} = 2\overrightarrow{r} \qquad (4)$$

Where *a* decreases linearly from 2 to 0 during iterations (in both exploration and extraction phases) and *r* is a random vector between 0 and 1. Exploitation phase in WOA algorithm: In order to mathematically model the bubble behavior of tour walls, 2 methods have been designed:

Contraction blockade mechanism

This behavior is obtained by increasing the value of *a* in equation 3. The swing range of *A* is reduced by a. In other words, *A* is a random value between *a* and -*a* and *a* decreases from 2 to 0 during iterations. By choosing random values of *A* between -1 and +1, the new location of the search agent can be defined anywhere between the original location of the agent and the location of the current best agent. Spiral updating location: This method first calculates the distance between the wall located at X^* and *Y* coordinates of the bait located at X^* and Y^*. A helical equation is created between the position of the whale and the prey to mimic the helical motion of the humpback whale:

$$\overrightarrow{X}(t+1) = \overrightarrow{D'}(t).e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t) \qquad (5)$$

In this equation, *D*' refers to the distance between the 1st whale and the prey (the best solution obtained so far), *b* is a constant to define the shape of the logarithmic spiral, and is a random number between -1 and +1. It should be noted that the humpback whale swims around the prey along a contraction circle and at the same time in a spiral path. In order to model this simultaneous behavior, it has been assumed that the whale chooses one of the constriction blockade mechanism or the spiral model with a probability of 50% to update the position of the whales during the optimization. The mathematical model is as follows:

$$\overrightarrow{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t) - \overrightarrow{A}.\overrightarrow{D} & p < 0 \\ \overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t) & p \geq 0 \end{cases} \qquad (6)$$

Where, *p* is a random number between 0 and 1. In addition to the bubble net method, humpback whales search for prey randomly. The mathematical model of search is as follows. The discovery stage in WOA: A similar method based on the variation of vector *A* can be used for hunting (exploration). In fact, humpback whales search randomly according to each other's location. Therefore, vector A with random values greater than or less than -1 is used to force the search agent to move away from the reference whale. Unlike the extraction phase, in order to update the position of the search agent in the exploration phase, instead of using the data of the best search agent, random selection of the agent has been used. This mechanism, along with A>1, emphasizes discovery and allows the WOA algorithm to perform a global search. The mathematical model is as follows:

$$\overrightarrow{D} = \left|\overrightarrow{C}.\overrightarrow{X_{rand}} - \overrightarrow{X}\right| \qquad (7)$$

$$\overrightarrow{X}(t+1) = \overrightarrow{X_{rand}} - \overrightarrow{A}.\overrightarrow{D} \qquad (8)$$

In this equation, X_{rand} is a randomly selected position vector (random whale) from the current population. The WOA algorithm starts with a set of random solutions. In each iteration, the search agents update their position according to the randomly selected search agent with the current best obtained solution. The parameter *a* is reduced from 2 to 0, respectively, in order to provide exploration and extraction. A random search agent is selected in the case |A|>1, while the best solution is selected when the search agents position update is |A|<1. Depending on the value of *p*, the WOA algorithm has the ability to choose between circular or spiral motion. Finally, the WOA algorithm terminates by satisfying the termination conditions.

State vector machine

SVM is another simple algorithm that provides relatively good performance with less computational cost. In regression, SVM works by finding a hyper plane in an N-dimensional space (N number of features) that fits the multidimensional data while considering an outlier. In classification, the same hyper surface is computed, but only for individual classification of data points while considering a margin. There are many possible hyper pages to choose from. However, the goal is to find the hyper plane with the maximum margin, that is, the maximum distance between the target classes (Manoharan *et al.,* 2022; De Farias Silva *et al.,* 2022)

Proposed Strategy

The WOA method is one of the new meta-heuristic algorithms and imitates the behavior of hunting whales. This algorithm starts with a set of random solutions, in each iteration the search agents update their position according to each of the search agents randomly or with the best solution obtained so far. In this paper, the purpose of performing simulations is that we want to design a system that can determine whether the cancer is benign (1) or malignant (-1) with high accuracy by having 9 medical parameters from a patient suspected of breast cancer. In fact, by this system, patients referring to this system are supposed to be divided into two classes: benign (1) and malignant (-1). Therefore, as mentioned above, this issue can be considered as one of the important issues in the field of classification. Also, as mentioned, in this paper we are going to use the SVM neural network to implement the Classifier system. Meanwhile, in order to perform classification and prediction by SVM with a very low and negligible error rate, one of the optimization methods is also needed. In this paper, WAO algorithm is used and using parameters of the classification strategy, i.e. SVM, which are C and Rbf_Sigma, should be selected in such a way that the prediction error or classification is minimized. On the other hand, as we know, any neural network, including SVM, in order to achieve its goals, i.e. classification or prediction, must go through the training process at the beginning of creating this strategy, and after the neural network is trained and matured like an intelligent human brain, in order to evaluate, the network also needs to be tested. As a result, by using a series of primary data extracted from reference articles, the SVM neural network is trained and tested, and during the training process, the parameters of the SVM neural network are optimized by the WOA whale meta-heuristic optimization algorithm.

Model Evaluation

The investigated features in the mechanism for unsupervised domain adaptation detection include precision, recall, accuracy, and score criterion, which are given in equations 9 to 12, respectively.

$$\text{Precision} = \frac{tp}{tp + FP} \tag{9}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{10}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{11}$$

$$\text{F}_{score} = \frac{2tp}{2tp + fp + fn} \tag{12}$$

In the above equation:

TP represents the number of records whose true category is positive and which the algorithm correctly recognized as positive.

FN represents the number of records whose category is positive and which the algorithm mistakenly recognized as negative.

FP represents the number of records whose real category is negative, but the algorithm mistakenly recognized its category as positive.

TN represents the number of records whose real category is negative and which the algorithm correctly recognized as negative.

RESULTS AND DISCUSSION:

The primary data that have been extracted and used are as follows:

- The number of patients whose medical information has been collected about their cancerous tumor is equal to 683 people.
- The number of medical parameters obtained from each patient is equal to 9 parameters, and each parameter can assign a degree of cancer contamination from 1 to 10.

In **Table 1**, the degrees of cancer contamination considered for the considered parameters, which are 9 in number, are shown.

Table 1: Parameters and degrees of cancer contamination.

Item	Parameter	contamination degree
1	Clump Thickness	1-10
2	Uniformity of Cell Size	1-10
3	Uniformity of Cell Shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelial Cell Size	1-10
6	Bare Nuclei	1-10
7	Bland Chromatin	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10

Also, the number of patients for each degree of cancer infection in each of the above parameters is presented in **Table 2**.

Table 2: Number of patients in each degree of contamination.

Domain	1	2	3	4	5	6	7	8	9	10	Sum
Clump Thickness	139	50	104	79	128	33	23	44	14	69	683
Uniformity of Cell Size	373	45	52	38	30	25	19	28	6	67	683
Uniformity of Cell Shape	346	58	53	43	32	29	30	27	7	58	683
Marginal Adhesion	393	58	58	33	23	21	13	25	4	55	683
Single Epithelial Cell Size	44	376	71	48	39	40	11	21	2	31	683
Bare Nuclei	402	30	28	19	30	4	8	21	9	132	683
Bland Chromatin	150	160	161	39	34	9	71	28	11	20	683
Normal Nucleoli	432	36	42	18	19	22	16	23	15	60	683
Mitoses	563	35	33	12	6	3	9	8	0	14	683

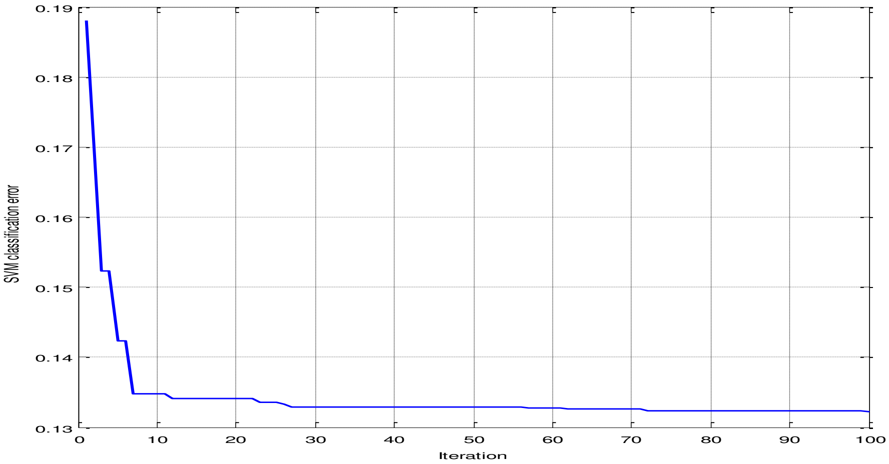

Fig. 1: Reducing the final error by increasing the iteration of WOA algorithm.

Optimization results and neural network error percentage after applying the test data to evaluate the performance of SVM network can be seen in **Fig. 1** and **2**. **Fig. 1** shows the optimization of cancer diagnosis error in patients by the proposed method. **Fig. 2** also shows the close relationship between the two parameters used in SVM for optimization and simulated by the whale algorithm.

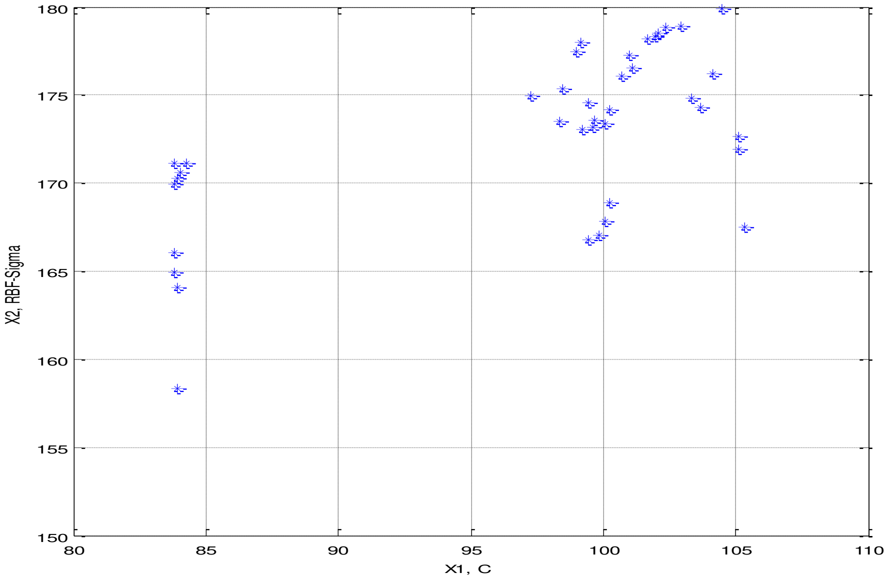

Fig. 2: Close relationship between two optimization parameter.

In this section, the results obtained on the test data will be given based on the classifications used. Accordingly, the disturbance matrix will also be given.

Following are the disturbance matrices for the three classifiers used after feature selection on the test data. **Fig. 3** and **4** shows the disturbance matrix related to tree method and SVM, respectively.

Table 3: Results obtained on test data.

	Classified tree		SVM	
	Before	After	Before	After
Precision	71.67	71.67	78.67	79.33
Recall	83.33	83.33	90.00	91.43
Accuracy	77.78	77.78	81.47	81.36
Score criterion	80.46	80.46	85.52	86.10

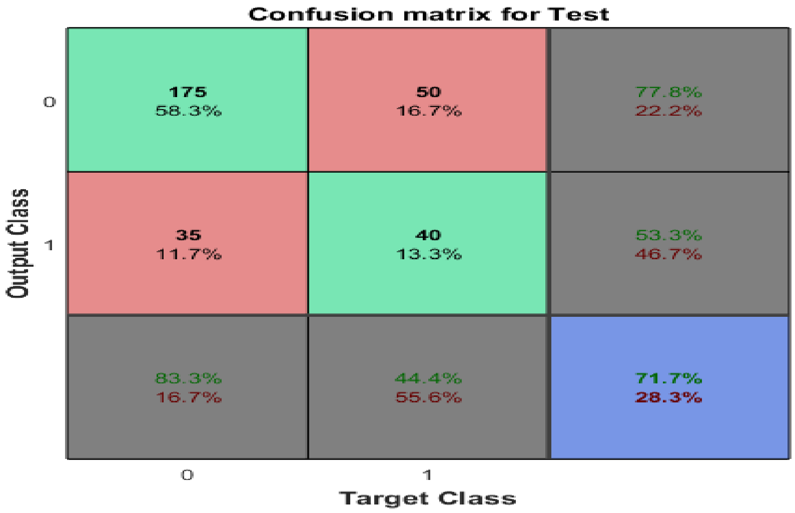

Fig. 3: Disturbance matrix on test data and decision tree classification.

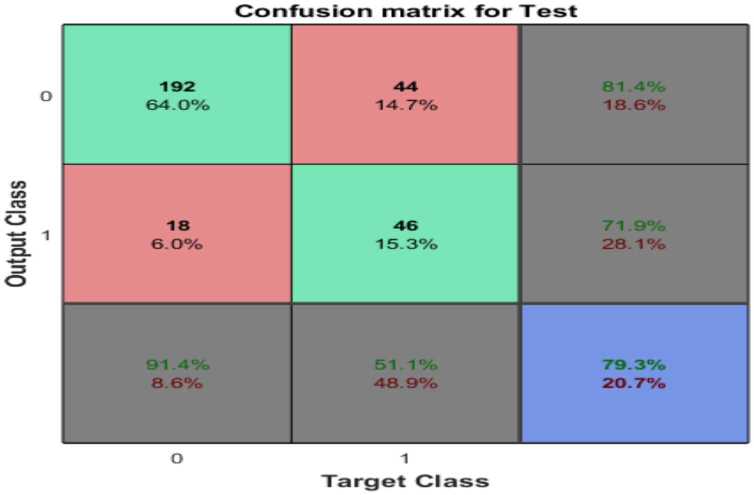

Fig. 4: Disturbance matrix on test data and SVM classification.

In this section, the obtained results are analyzed based on the tables and figures given, it can be said:

1) Feature selection can improve the efficiency of classification algorithms.
2) The efficiency of the models on the test data was almost close to each other.

The current study was carried out with the aim of optimizing and evaluating a system based on support vector machine to help the doctor to determine the type of breast cancer masses. The system presented in this study was successful in detecting benign and malignant masses and performed the classification with 97% accuracy. The advantage of SVM-WOA's proposed method was the use of feature reduction using the principal component analysis method and the appropriate selection of the evolutionary optimizer, the results of which indicated greater speed and better generalizability while increasing accuracy compared to other cases implemented in this issue. Based on this, this method can be a very suitable tool to help doctors diagnose the disease or be used as a second opinion for the final diagnosis. The use of such accurate and fast methods increases the hope of implementing an intelligent breast cancer diagnosis system. The simulation results showed that the proposed system has reached an accuracy of 97 on the data set of breast cancer patients, which is higher than similar researches on this data set. In addition, one should pay attention to the value of the false negative parameter. The percentage of this parameter is very important in prediction models in the field of medicine because a sick person is wrongly considered healthy, which can have very dangerous consequences. In the prediction model presented in this research, this value was zero for the 30-70 categories, which is another advantage of the proposed method. It is worth mentioning that this model was proposed for breast cancer detection, which is a two-class problem. Therefore, its behavior in other issues related to multi-class classification needs to be investigated. In addition, the proposed model was applied only on quantitative data, which can be applied to the behavior of this algorithm on the images and signals studied in this field. Also, since the lack of comprehensive internal clinical data is one of the limitations of the present study, it is suggested for future studies to localize this system to train the network with hospital data sets and predict this disease.

CONCLUSION:

In this paper, the WOA algorithm was used to optimize the output of the SVM to classify the type of breast cancer into two categories, benign and malignant. First, the data used were pre-processed by normalization and using independent component analysis. Then the network was trained and validated with training and validation samples and finally tested with test samples. One of the reasons for the high sensitivity and specificity in this paper is the pre-processing of the input data and the appropriate selection of the SVM optimizer for this purpose. In fact, the method proposed in this paper is superior to other methods due to its high speed and accuracy and good generalizability. Such studies can be reviewed and used for future studies, and on the other hand, they will be very economical due to the low cost and high speed of the process.

ACKNOWLEDGEMENT:

We are grateful to all the dear professors for providing their information regarding this research.

CONFLICTS OF INTEREST:

The authors are declared obviously in the manuscript and have no conflict of interest.

REFERENCES:

1) Begum A, Mamun MAA, and Begum M. (2024). Effective stroke prediction using machine learning algorithms. *Aust. J. Eng. Innov. Technol.*, **6**(2), 26-36. https://doi.org/10.34104/ajeit.024.026036
2) Chakraborty, S., Sharma, S., & Saha, A. (2022). A novel improved whale optimization algorithm to solve numerical optimization and real-world applications. *Artificial Intelligence Review*, 1-112. https://doi.org/10.1007/s10462-021-10114-z
3) Dallagassa, M. R., dos Santos Garcia, C., & Carvalho, D. R. (2022). Opportunities and challenges for applying process mining in healthcare: a systematic mapping study. *J. of Ambient Intelligence and Humanized Computing*, 1-18.
4) De Farias Silva, C. E., Costa, & Tonholo, J. (2022). Application of machine learning to predict the yield of alginate lyase solid-state fermentation by Cunninghamella echinulata: artificial neural networks and support vector machine. *Reaction Kinetics, Mechanisms and Catalysis*, **135**(6), 3155-3171.
5) De Roock, E., & Martin, N. (2022). Process mining in healthcare-An updated perspective on the state of the art. *J. of biomedical informatics*, **127**, 103995.
6) Hanna, K., Krzoska, E., & Speirs, V. (2022). Raman spectroscopy: Current applications in breast cancer diagnosis, challenges and future prospects. *British j. of cancer*, **126**(8), 1125-1139.
7) Houssein, E. H., Emam, M. M., & Ali, A. A. (2022). An optimized deep learning architecture for breast cancer diagnosis based on improved marine predator's algorithm. *Neural Computing and Applications*, **34**(20), 18015-18033.
8) Manoharan, A., Begam, K. M., & Sooriamoorthy, D. (2022). Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review. *J. of Energy Storage*, **55**, 105384.
9) Munoz-Gama, J., Martin, N., & Zerbato, F. (2022). Process mining for healthcare: Characteristics and challenges. *J. of Biomedical Informatics*, **127**, 103994.
10) Nemade, V., Pathak, S., & Dubey, A. K. (2022). A systematic literature review of breast cancer diagnosis using machine intelligence techniques. *Archives of Computational Methods in Engineering*, **29**(6), 4401-4430. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10968939/
11) Yang, W., Xia, K., & Feng, Y. (2022). A multi-strategy Whale optimization algorithm and its application. *Engineering Applications of Artificial Intelligence*, **108**, 104558.

Citation: Mohammadiounotikandi A., and Babaeitarkami S. (2024). A method based on process mining for breast cancer diagnosis with whale optimization algorithm and support vector machine. *Aust. J. Eng. Innov. Technol.*, **6**(3), 70-78. https://doi.org/10.34104/ajeit.024.070078