Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8122895, 8 pages
https://doi.org/10.1155/2022/8122895

*Research Article*

# Breast Cancer Identification Using Machine Learning

**Xiao Jia,[1] Xiaolin Sun,[2] and Xingang Zhang[1]**

[1] *Henan Engineering Research Center of Intelligent Processing for Big Data of Digital Image, School of Computer Science and Technology, Nanyang Normal University, Nanyang, Henan 473061, China*
[2] *Information Management Center, Nanyang Normal University, Nanyang, Henan 473061, China*

Correspondence should be addressed to Xingang Zhang; cyny@nynu.edu.cn

Received 29 July 2022; Revised 15 September 2022; Accepted 19 September 2022; Published 3 October 2022

Academic Editor: Lianhui Li

Breast cancer is a cancer disease that seriously threatens women's health and occupies the first place in female cancer mortality. At present, the incidence rate of breast cancer in China is the first in the world and is on the rise. In view of the serious harm of breast cancer to life and health, researchers and institutions are making unremitting efforts to find a perfect diagnosis and treatment plan. With the improvement of computer performance and machine learning levels, intelligent algorithms have been able to replace human behavior and judgment in some fields. The traditional breast cancer diagnosis process requires medical experts to observe patient data repeatedly. In this case, the algorithm technology is used to quickly feedback a high probability reference result to doctors, which is particularly important to increase the diagnosis efficiency and reduce the burden of doctors. In order to improve the accuracy of existing breast cancer recognition methods, this paper proposes and implements a scheme based on a whale optimization algorithm to iteratively adjust the key parameters of the support vector machine to improve the accuracy of breast cancer recognition. In order to verify the performance of the WOA-SVM algorithm, this paper uses the Wisconsin breast cancer data in the UCI database for performance verification experiments. Experiments show that the WOA-SVM model has higher recognition accuracy than the traditional breast cancer recognition model.

## 1. Introduction

Breast cancer is the most common cancer and the leading cause of death in women. Around 1.4 million people worldwide are diagnosed with breast cancer every year. About 500000 people died of the disease, of which China accounted for 12.2% of the newly diagnosed cases and 9.6% of the deaths. Breast cancer not only destroys the bodies of patients but also seriously affects the image, psychology, and family relations of female patients [1–5]. In terms of the geographical distribution of the prevalence, the incidence rate of breast cancer in urban areas is significantly higher than that in rural areas. However, because the medical and health conditions and mental health levels of urban patients are better than those in rural areas, the mortality of urban patients is lower than that of rural patients. At present, the incidence rate of female breast cancer in most countries and regions is on the rise, especially in regions with low incidence rates of early breast cancer such as Asia, Africa, and South America. Therefore, the situation of breast cancer in the world is not optimistic.

The causes of breast cancer are complex, mainly focusing on sex hormone disorders, genetic factors, and viral infection. The incidence rate of breast cancer decreased significantly in women who had undergone ovariectomy, which confirmed that estrogen and lutein were directly related to the incidence of breast cancer. For the family with a history of breast cancer, especially for women with early menarche and dense breast glands, the prevalence rate will also increase significantly. It mainly focuses on the ageing process, unhealthy living habits, and genetic factors. In biological experiments, it has been proved that the virus causes cancer. When the virus enters the breast gland, it will induce a series of pathological changes in the organs and eventually convert to breast cancer. In addition, the pollution deposited by the Earth's environment with the development of society has also eroded human beings. Relevant studies have confirmed that in areas with serious industrial pollution, the incidence rate of breast cancer will be significantly higher than that in other areas.

The diagnostic process of breast cancer is divided into pathological diagnosis and imaging diagnosis. The pathological diagnosis of breast cancer relies on the reference of cell morphology. First, the cells of the breast mass are extracted by a fine needle, and the size, thickness, uniformity, and other data of the cells are counted in detail. Finally, the newly obtained data are classified according to the characteristics of the previous data. The imaging diagnosis methods of breast cancer are mainly divided into mammography, PET, MRI, CT, and ultrasound [6–11]. These methods have their own unique theories and tools to provide doctors with the diseased organs of patients which are invisible to the naked eye. Among them, breast ultrasound has a good resolution for soft tissue, can clearly show the layers between the thymus and the chest wall, and can accurately identify millimeter-level tumors. It is a non-radioactive way and has been a routine item for the detection of breast cancer.

In the past half century, computer technology has brought a great driving force to social progress. With the change of times, computer technology has also scattered a wide range of research directions. As a star concept, artificial intelligence has become the direction of various technologies. Using machines to completely replace artificial is not only a beautiful demand but also a visible future. As an important category of artificial intelligence, machine learning technology can automate the process of data processing and use specific mathematical models to describe the data completely. Today, in the medical field, computer-aided diagnosis based on machine learning technology has been specifically implemented for various diseases [12–17]. Whether it is to use algorithms to classify and judge a large number of data or to predict new data based on past data, the existing algorithm models have achieved reliable results and have also been recognized and trusted by the majority of pathologists.

Although individual breast cancer can be judged in advance by miRNA sequence detection technology, breast cancer is a chronic cancer with no obvious early symptoms. Once the abnormality is detected, it is likely to be in the middle and late stages. Using machine learning technology to make accurate judgments can carry out targeted early diagnosis and treatment for patients. In order to improve the accuracy of existing breast cancer recognition methods, this paper proposes and implements a scheme based on a whale optimization algorithm to iteratively adjust the key parameters of the support vector machine to improve the accuracy of breast cancer recognition. In order to verify the performance of the WOA-SVM algorithm, this paper uses Wisconsin breast cancer data in the UCI database for performance verification experiments. Experiments show that the WOA-SVM model has higher recognition accuracy than the traditional breast cancer recognition model.

## 2. Related Work

*2.1. Diagnostic Techniques for Breast Cancer.* There are many research studies on machine learning technology in the intelligent diagnosis of breast cancer. The clinical data for breast cancer will be collected according to the patient's disease degree and period, including the patient's age, tumor cell morphology data, physical condition before and after the disease, survival time, and other aspects. Through these collected real data, various algorithm models are used to test and analyze. The researchers proposed using the sigmoid kernel support vector machine method for the auxiliary diagnosis of breast cancer. Under the 5-fold cross-validation, the average accuracy reached 96.24%, but this method ignored the further improvement of the results by accurate parameter optimization. Researchers use particle swarm optimization algorithms to optimize SVM parameters. The model has been applied to breast cancer data and achieved 97.28% high precision. It is proven that the optimization algorithm can significantly improve the SVM classification ability and has a certain practical value. Alickovic et al. [18] used a classification model based on the normalized multilayer perceptron neural network to classify the Wisconsin breast cancer (original) data set, used 80% of the data as a training set and 20% of data as a test set, and obtained 99.27% accuracy. Khadija et al. [19] used the Naive Bayes and support vector machine algorithms for the Wisconsin breast cancer diagnosis data set and used a 10-fold cross-validation method to detect the results. Among them, the Naive Bayes algorithm obtained 95.65% recognition accuracy, which is better than that of the support vector machine algorithm. The researchers used the support vector machine feature elimination algorithm to process the triple-negative breast cancer data set and then used the decision tree algorithm to test the classification, and the accuracy reached 97.8%. Polat et al. [20] proposed using least squares support vector machine (LS-SVM) to classify breast cancer data and achieved high accuracy. However, the accuracy results will vary greatly under different verification methods. Watkins et al. [21] proposed an immune-inspired supervised learning algorithm airs and applied it to the breast cancer data sample donated by Dr. William Wolberg of Wisconsin University Hospital in the United States. Through 10-fold cross-validation, the recognition rate reached 97%. At present, there may be some inaccuracies in the diagnostic research on clinical data using machine learning technology, but this is largely due to the limitations of data acquisition accuracy, computer hardware level, algorithm level, etc; these problems will be completely solved with the passage of time.

Compared with the number of studies on clinical data of breast cancer, there are relatively few studies on ultrasound images of breast cancer. Because ultrasonic images are often high noise images, modeling recognition or doctor judgment will produce high error values. Compared with other medical imaging methods, ultrasound images are cheap, noninvasive, and painless, and have high diagnostic feedback efficiency, so they are also widely welcomed and applied. Therefore, it is necessary to develop vigorously. According to the difference in geometric features between benign and malignant tumors in ultrasound breast tumor images, the researchers developed a diagnostic scheme. The final experimental accuracy was 72.90%. In addition, the ultrasound images of breast nodules can be used for experiments using the USNet model. The AUC values of benign nodules, malignant nodules, and normal glands are 93.67%, 94.34%, and 99.68% respectively. In addition, researchers used the AlexNet network model to recognize the ultrasound images of breast nodules provided by Beijing Friendship Hospital. Using the verification method of 80% of the training set and 10% of the verification set and test set, after being processed by the ACE algorithm group, 93.25% of the AUC value was obtained. Uniyal et al. [22] also introduced the gray level cooccurrence matrix into the feature extraction of ultrasonic RF signals, which further improved the accuracy of the feature description of breast lesions. The AUC value of the support vector machine classifier reached 86%. In addition, the researchers used a hierarchical binary tree SVM classifier to classify breast ultrasound RF data. In the process, the Shearlet transform was used to extract the multiscale and multidirectional features of breast ultrasound RF signals. Second, a multiscale directional binary pattern (MDBP) was used to reduce the data dimension without losing feature information. The experimental results showed that the AUC value was 78.40% and the accuracy reached 89.29%.

*2.2. Machine Learning.* Machine learning is the inevitable outcome of social development, and also the inevitable outcome of the development of artificial intelligence research to a certain extent. From the 1950s to 1970s, the research study on machine learning technology was conducted in the “reasoning period.” People at that time had realized that as long as the machine was endowed with logical reasoning ability, the machine would have wisdom. The original imagination is still so directional now. The representative products of this period mainly include A.Newell logic theorist program and later “general problem solving program” [23]. In the mid-1970s, researchers gradually felt that artificial intelligence could not be realized only by making machines have reasoning ability, but also by making machines have knowledge. During this period, many expert systems for solving practical problems appeared, with gratifying results. However, the expert system soon reached the bottleneck period, and some researchers thought that the way out for artificial intelligence was to let machines learn knowledge by themselves. In the 1980s, a decision tree was born, which directly simulated the human judgment process. Today, a decision tree is a commonly used machine learning technology. In 1986, the BP algorithm was also born, which laid the foundation for the hot in-depth learning at that time. In the mid-1990s, “statistical learning” was born, and its representative technology is the famous support vector machine. Its own algorithm and various derived kernel techniques are still widely used in almost all research fields. In the 21st century, there has been an upsurge in in-depth learning. The so-called in-depth learning refers to neural networks with more complex topology and more layers. At present, the application scenarios of in-depth learning are very rich, and in various competitions, in-depth learning has also made many outstanding performances.

Today, machine learning technology has been used in all aspects of life, such as weather forecast, stock trend, medical identification, etc. In addition, Gaode, Baidu Maps and other navigation platforms can recommend the best route according to the real-time traffic conditions, current location, and destination. In the 2012 US. election, there was an excellent machine learning data analysis team in the Obama camp, which conducted a three-dimensional analysis of various intelligence data, guided Obama to make timely and accurate election campaigns, and made a very important technical foreshadowing for the final election [24]. In the man-machine confrontation go game held in March 2016, AlphaGo [25] designed by the deep mind team based on reinforcement learning, also defeated lishiyu, the best chess player at that time, with a score of 4 : 1, which had a great shaking effect in the world, and also attracted great attention from all walks of life to machine learning technology. The auto drive system based on in-depth learning developed by Tesla is also becoming more and more perfect and will finally realize the situation of unmanned driving one day. The self-service shopping device that uses facial recognition to check out has been deployed nationwide by the Alipay team of Alibaba, which has reduced the waiting period of shopping queues, made it easy to use, and achieved a good response. In the field of machine learning, there are many excellent institutions dedicated to special research, including Stanford Research Institute, Google Lab, Aridamo Institute, etc. these organizations have done a lot of practical research, which has greatly promoted the extension and development of the field of science and technology.

*2.3. Intelligent Medical Technology.* At present, machine learning has been applied in many fields, such as data modeling, image analysis, natural language processing, audio recognition, social network filtering, and so on. The level of machine learning in the medical field is directly related to people’s life safety, so it is the key direction of intelligent technology development. In the middle of the 20th century, Ledley [26] first proposed applying the algorithm model to the medical scheme and using computers for auxiliary diagnosis. The concept of “intelligent medical care” was formally put forward in 2008. IBM took the lead in launching a medical wearable device that integrates Internet of things technology and artificial intelligence technology and achieved a good response. For the research and development of medicine, artificial intelligence technology can create a simulation environment based on the existing medical textbooks, patient medical records, treatment plans, and other huge amounts of data, from which the optimal composition ratio can be mined, greatly reducing the research and development threshold and cycle [27].

Machine learning-assisted diagnosis has the most important application in the medical field. According to statistics, more than 57 million cases in China are misdiagnosed to varying degrees every year, and the misdiagnosis rate is as high as 27%. In the future, with the improvement of machine learning level, the accuracy of patient data recognition and imaging diagnosis will be higher and higher, which will effectively reduce misdiagnosis and improve diagnosis efficiency, reduce the occurrence of various medical disputes, and provide high-quality medical services for the public. Intelligent medical technology has greatly eased the pressure on patients' treatment. With various public opinion surveys, medical staff and nonmedical staff have very high support and expectations for intelligent medical technology. At present, many departments need to work together in some disease fields, so the integration of intelligent medicine is an important goal.

The essence of intelligent diagnosis and treatment is to apply intelligent technology to the patient's course and treatment cycle, intelligently analyze some real-time test reports of the patient, and feed back the test results and treatment plans recommended by the machine. While intelligent medical technology requires hardware adaptation, of course, it cannot be separated from the support of data. Although we are now in the era of data sharing and there are some public data sets for researchers to study and analyze, institutions such as hospitals and medical research institutes have more data sets with a large number of samples. These data sets are more high-quality, but they cannot be provided to the outside world for various reasons. If all medical data sets can be interoperable, the current level of intelligent medicine will be greatly improved.

## 3. Proposed Breast Cancer Recognition Algorithm

*3.1. Whale Optimization Algorithm.* The whale optimization algorithm (WOA) is a new group optimization algorithm developed by Mirjalili et al. [28] in 2016 to establish a mathematical model by imitating the hunting behavior of humpback whales. Whales are considered to be the largest mammals in the world. In the whale brain, there are spindle cells similar to those in humans. These cells are responsible for judgment and social behavior. Therefore, they are also highly intelligent creatures. The humpback whale preys ingeniously. First, it swims upward in a spiral posture from a depth of 15 meters from the sea surface, spits out bubbles of different sizes, and makes all bubbles reach the water surface at the same time, forming a cylindrical and tubular bubble net, which narrows the range of activity of the prey and forces it towards the center of the bubble net. Then, it opens its big mouth almost vertically in the bubble net and swallows all the prey in the net.

WOA has the characteristics of a simple principle, few parameter settings, and strong optimization performance. It has been proved that WOA is superior to the traditional meta heuristic algorithm in both search accuracy and convergence rate in the extreme value optimization of standard test functions. The following describes three methods of WOA location update: surround prey, bubble attack, and random search.

(1) The stage of encircling prey. In the initial stage of predation, the whale will observe the approximate location of the prey and then surround the prey group. In the WOA algorithm, it is assumed that the problem solution or problem variable to be optimized is the location of the optimal whale. After the best prey position is defined, other whale individuals will approach this position and gradually surround the food. In contrast, in the WOA algorithm, the distance between the individual and the optimal whale needs to be calculated first:

$$\overrightarrow{D} = \left|\overrightarrow{C}\overrightarrow{X^*}(t) - \overrightarrow{X}(t)\right|, \tag{1}$$

where $t$ is the current iteration number. $\overrightarrow{X^*}(t)$ is the individual position of the t-genera tion whale, which will be continuously revised with each iteration. The swing factor is defined as follows:

$$\overrightarrow{C} = 2\overrightarrow{r}. \tag{2}$$

The individual position in the whale group will be updated according to the best individual position, and the expression is as follows:

$$\overrightarrow{X}(t+1) = \overrightarrow{X^*}(t) - \overrightarrow{A}\overrightarrow{D}, \tag{3}$$

where $\overrightarrow{A}$ is the convergence factor, which is defined as follows:

$$\overrightarrow{A} = 2\overrightarrow{a}\overrightarrow{r} - \overrightarrow{a}, \tag{4}$$

where $r$ is a random number between 0 and 1.

(2) Bubble attack phase. In this order, according to the behavior of the humpback whale to spit out bubbles for predation, two strategies, contraction and spiral update, are designed to achieve the purpose of local optimization of the algorithm. In the process of shrinking and encircling mechanisms, the whale group shrinks and encircles. When a is less than 1, the whale will approach the whale with the best current position. The process of spiral position updating is to calculate the distance between the individual humpback whale and the current optimal whale, and then shrink the swimming in a spiral way. In the process of food search, the mathematical model of spiral contraction mode is as follows:

$$\begin{aligned}\overrightarrow{X}(t+1) &= \overrightarrow{D'}e^{bl}\cos(2\pi l) + \overrightarrow{X^*}(t),\\ \overrightarrow{D'} &= \left|\overrightarrow{X^*}(t) - \overrightarrow{X}(t)\right|,\end{aligned} \tag{5}$$

where $\overrightarrow{D'}$ represents the distance vector between the individual and the currently optimal whale; $L$ is a random number between 0 and 1; and B is a constant that limits the shape of the logarithmic spiral. In order to be able to both maintain contraction and swim to food along the spiral path.

(3) Random search phase. Humpback whales capture prey by controlling the a vector to swim. When a is greater than 1, individual whales randomly search according to each other's positions, which can promote individual humpback whales to conduct global search and obtain the global optimal solution.

*3.2. WOA-SVM Algorithm.* At the beginning of the birth of each metaheuristic algorithm, the most intuitive performance is demonstrated through its extreme value optimization ability of the standard test function. Therefore, in the face of practical problems to be solved, WOA's fitness optimization is generally divided into the maximum problems and minimum problems. The biggest problem is that the classification accuracy is generally used as the fitness; for the minimum problem, the error rate is generally used as the fitness. In this paper, the misclassification rate of the SVM model in the sense of cross-validation is taken as the fitness value, and the two important parameters of SVM are iteratively optimized by the WOA algorithm to continuously improve the generalization accuracy of the SVM model.

Reasonable parameter setting is very important to improve the classification performance of SVM. In this paper, the WOA-SVM algorithm is used to recognize breast cancer data. The WOA-SVM algorithm first initializes the population, that is, initializes the population number and the individual search dimension. Since the parameters that have the greatest impact on the accuracy of the SVM model are the penalty factor C and the kernel function parameter $g$, the individual search dimension is set to 2. The data set used in this paper is relatively small, and the dimension of a single data sample is small. The initial position of the individual population can be set randomly, which will not affect the convergence speed and final accuracy of the algorithm.

In each iteration of the WOA algorithm, all individuals in the population will be updated according to their search results. Individuals beyond the search boundary will be reinitialized. For individuals that do not exceed the boundary, local optimization or global optimization will be performed according to the parameter setting conditions. When the iteration times of the whole algorithm model reach the set value or get perfect fitness results, the whole test process ends. The flowchart of the WOA-SVM algorithm for identifying breast cancer is shown in Figure 1.

## 4. Experiments

### 4.1. Experimental Setup

*4.1.1. Experimental Environment and Parameter Setting.* The experimental research platform and tools are Microsoft Windows10 and MATLAB r2018a. The LIBSVM toolbox is used for some functions of the research work. The Toolbox provides many application functions related to SVM models and their source codes, which make it convenient for users to conduct data research. In order to show the performance advantages of the WOA-SVM, BP neural network, traditional SVM, and PSO-SVM are selected as comparison methods. The BP model is established by using the new function in the MATLAB neural network toolbox. The hidden layer is set to 10 neurons, the training times are set to 500, the learning rate is set to 0.01, and the default 1e-5 is used as the training target value. Traditional SVM model parameter settings: $C = 0.1$, $g = 0.005$. The population size of PSO and WOA is set to 10 and the number of iterations is 100.

In this paper, the clinical data of breast cancer are studied by using the WBCD data set in UCI, which is taken from the nuclear features extracted from the fine needle of a breast mass. The data set contains 699 samples, each of which has 9 characteristic attributes and 1 category label. In the sample data inspection, it is found that 16 sample data are missing, and the samples with missing data have been eliminated. The resulting data set is shown in Table 1.

*4.1.2. Experiment Verification Method Experiment.* Two verification methods are adopted. The first verification method adopts the traditional set-aside method, which first divides the data set into two mutually exclusive sets, assuming that the two sets are *s* and *t*, respectively. Set *s* is used to train the algorithm model, and set *t* is used to test the generalization ability of the algorithm. In order to ensure the consistency of data category distribution, the training set and test set should be selected randomly from the middle proportion of the two categories, that is, assuming that 70% of the data samples are required as the training set; then, 70% of the samples from benign and malignant are randomly selected for model training, and the remaining samples are used for evaluation. The second verification method adopts 10-fold cross-validation. The data set is divided into 10 parts, on average, 9 of which are used as training data and 1 as test data in turn. In this paper, benign and malignant samples are divided into 10 equal parts respectively. Each time, one benign data sample and one malignant data sample are selected as the verification set, and the rest of the data samples are used as the test set for simulation experiments.

*4.2. Performance Evaluation Index.* The evaluation index is a quantitative index of the quality of the algorithm or parameters given by inputting the same data into different algorithm models or the same algorithm model with different parameters. In many evaluation indicators, most of them can only reflect part of the performance of the model. If the evaluation indicators are not used reasonably, the problems with the model itself will not be found, and even wrong conclusions will be drawn. Therefore, in the process of evaluating the performance of the algorithm model, it is often necessary to combine a variety of different indicators for comprehensive analysis.

Accuracy is the most direct and original indicator for evaluating model performance. It is defined as the percentage of samples correctly identified by the algorithm model within the total samples. The calculation formula is as follows:

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}. \tag{6}$$

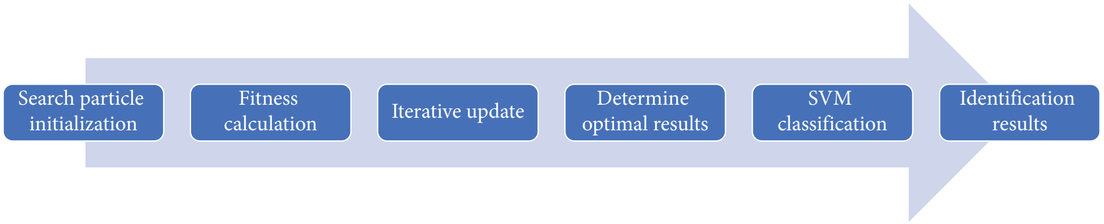

Figure 1: Flowchart of the recognition algorithm.

Table 1: WBCD data set information table.

| Data set name | WBCD |
|---|---|
| Number of samples | 683 |
| Benign | 444 |
| Malignant | 239 |
| Characteristic number | 9 |
| Category | 2 |

Table 2: Comparison of simulation results of tumor recognition rate (select 70% of the training set).

| Method | Accuracy | Benign diagnosis rate | Malignant diagnosis rate |
|---|---|---|---|
| WOA-SVM | 99.02 | 98.51 | 100 |
| PSO-SVM | 98.54 | 99.25 | 97.22 |
| SVM | 93.20 | 96.27 | 87.50 |
| BP | 98.06 | 99.25 | 95.83 |

Precision refers to the proportion of samples that are predicted to be true and actually true in the total number of true samples. Its formula is as follows:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{7}$$

Sensitivity refers to the ability of the model to correctly judge the data of malignant patients. The higher the sensitivity of the algorithm model, the lower the missed diagnosis rate of the model. The formula is as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \tag{8}$$

Specificity indicates the ability of the model to correctly judge nonpatients. The higher the specificity value, the lower the misdiagnosis rate of the representative model. The formula is as follows:

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

*4.3. Experimental Result.* It can be seen from Table 2 that WOA-SVM and PSO-SVM algorithms are significantly better than SVM and BP in terms of accuracy and benign/malignant accuracy of WBCD data sets. Compared with the SVM model with default parameters, the accuracy of the two optimization algorithms is improved by 5.34% and 5.82%, which shows that the value of key parameters has a great impact on the performance of the SVM model. When 70% of the training set samples are taken, the training results of the test set show that WOA-SVM is 0.48% higher than PSO-SVM in accuracy. Although PSO-SVM is 0.71% higher than WOA-SVM in benign diagnosis rate, the recognition rate of WOA-SVM for malignant samples has reached 100% perfect accuracy, 2.78% higher than PSO-SVM.

In order to more fully reflect the real performance of each algorithm for WBCD data set classification, the number of samples in the test set is increased, that is, the samples used for training are decreased step by step. Due to the reduction of learnable samples, the accuracy results of each algorithm also show a decreasing trend. The performance of the WOA-SVM algorithm in accuracy is still better than the other three algorithms, which further shows the effectiveness and stability of WOA-SVM.

The framework of using a group optimization algorithm to optimize SVM model parameters adopts the average accuracy of the model in the sense of 10-fold cross validation as the fitness value. When the model training is completed, the corresponding parameter values under the best fitness are transferred to the SVM model, and then tested in a corresponding way. Here, this paper directly derives the optimal fitness value, that is, the highest average accuracy value of the SVM model under 10-fold cross-validation, and transmits the corresponding sensitivity (SEN) and specificity (SPE). The results are shown in Table 3. Compared with the retention method, the accuracy of each algorithm model has significantly decreased in the accuracy results. The accuracy of the BP neural network model has decreased by 6.3% compared with the retention method of 70% training set and 30% test set, indicating that the BP model has poor overall control ability for the relatively rigorous validation method of 10 -fold cross-validation. In the SVM without an optimization algorithm and only using default parameters, the accuracy value is still not ideal. The WOA-SVM algorithm has higher accuracy and sensitivity than PSO-SVM. Therefore, in the 10-fold cross validation mode, WOA-SVM is still the optimal model compared with the other three algorithms.

## 5. Conclusion

As one of the most common cancers, breast cancer has a high incidence in women, and the incidence rate is increasing year by year. For patients, if they can accurately judge their illness and timely take targeted treatment plans to follow-up treatment, the survival rate of patients will be significantly improved. In order to reduce the possibility of misdiagnosis or missed diagnosis when doctors judge the

Table 3: Comparison of recognition performance under ten-fold cross-validation.

| Method | Accuracy | Benign diagnosis rate | Malignant diagnosis rate |
|---|---|---|---|
| WOA-SVM | 97.50 | 99.54 | 94.19 |
| PSO-SVM | 97.21 | 98.42 | 95.39 |
| SVM | 91.91 | 94.79 | 87.35 |
| BP | 91.76 | 92.72 | 93.40 |

patient data or breast pathological images, in this paper, machine learning technology is used to recognize WBCD data set obtained from fine needle aspiration of breast mass. The new group optimization algorithm WOA is proposed to intelligently adjust the parameters of the SVM model, maximize the fitting ability of the SVM model to the WBCD data set, and optimize the recognition results. Then, the simulation test is carried out by using the set aside method and the 10-fold cross-validation method. The experimental results show that the performance of the WOA-SVM model is significantly better than the traditional breast cancer recognition model and has a practical value.

The data sets used in this paper are not from the Asian region. Due to the great differences in the physique of women from all continents, the research results have certain limitations in terms of reference significance for Asian women. In addition, all the recognition experiments in this paper are carried out using traditional machine learning algorithms. Although they have achieved good results, they do not use the current popular deep learning algorithms. In the future, they are ready to extend the knowledge system to the field of deep learning.

## Data Availability

The data set can be accessed upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Acknowledgments

The authors thank the Key Scientific Research Projects of Colleges and Universities in Henan Province of China, Previous Normal-Dose CT Image Induced Perfusion CT Image Reconstruction (grant no. 20A520031).

## References

[1] A. G. Waks and E. P. Winer, “Breast cancer treatment: a review,” *JAMA*, vol. 321, no. 3, pp. 288–300, 2019.

[2] Y. S. Sun, Z. Zhao, Z. N. Yang et al., “Risk factors and preventions of breast cancer,” *International Journal of Biological Sciences*, vol. 13, no. 11, pp. 1387–1397, 2017.

[3] G. N. Sharma, R. Dave, J. Sanadya, P. Sharma, and K. K. Sharma, “Various types and management of breast cancer: an overview,” *Journal of Advanced Pharmaceutical Technology & Research*, vol. 1, no. 2, pp. 109–126, 2010.

[4] T. J. Key, P. K. Verkasalo, and E. Banks, “Epidemiology of breast cancer,” *The Lancet Oncology*, vol. 2, no. 3, pp. 133–140, 2001.

[5] J. G. Elmore, K. Armstrong, C. D. Lehman et al., “Screening for breast cancer,” *JAMA*, vol. 293, no. 10, pp. 1245–1256, 2005.

[6] M. Morrow, J. Waters, and E. Morris, “MRI for breast cancer screening, diagnosis, and treatment,” *The Lancet*, vol. 378, no. 9805, pp. 1804–1811, 2011.

[7] E. A. Morris, “Breast cancer imaging with MRI,” *Radiologic Clinics of North America*, vol. 40, no. 3, pp. 443–466, 2002.

[8] C. D. Lehman, C. Gatsonis, C. K. Kuhl et al., “MRI evaluation of the contralateral breast in women with recently diagnosed breast cancer,” *New England Journal of Medicine*, vol. 356, no. 13, pp. 1295–1303, 2007.

[9] B. Zangheri, C. Messa, M. Picchio, L. Gianolli, C. Landoni, and F. Fazio, “PET/CT and breast cancer,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 31, no. 0, pp. S135–S142, 2004.

[10] E. L. Rosen, W. B. Eubank, and D. A. Mankoff, “FDG PET, PET/CT, and breast cancer imaging,” *RadioGraphics*, vol. 27, no. suppl_1, pp. S215–S229, 2007.

[11] S. K. Yang, N. Cho, and W. K. Moon, “The role of PET/CT for evaluating breast cancer,” *Korean Journal of Radiology*, vol. 8, no. 5, pp. 429–437, 2007.

[12] Z. Jia, Y. Lin, J. Wang et al., “Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1977–1986, 2021.

[13] H. Samuel and O. Zaïane, “MedFact: towards improving veracity of medical information in social media using applied machine learning,” *Canadian Conference on Artificial Intelligence*, pp. 108–120, Springer, Cham, 2018.

[14] S. Chen, D. Bergman, K. Miller, A. Kavanagh, J. Frownfelter, and J. Showalter, “Using applied machine learning to predict healthcare utilization based on socioeconomic determinants of care,” *American Journal of Managed Care*, vol. 26, no. 1, pp. 26–31, 2020.

[15] Z. Jia, J. Junyu, X. Zhou et al., “Hybrid spiking neural network for sleep EEG encoding,” *Science China Information Sciences*, vol. 65, 2022.

[16] C. E. Brodley, U. Rebbapragada, K. Small, and B Wallace, “Challenges and opportunities in applied machine learning,” *AI Magazine*, vol. 33, no. 1, pp. 11–24, 2012.

[17] Z. Jia, X. Cai, and Z. Jiao, “Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging,” *IEEE Sensors Journal*, vol. 22, 2022.

[18] E. Alickovic and A. Subasi, *Normalized Neural Networks for Breast Cancer classification[C]//International Conference on Medical and Biological Engineering*, pp. 519–524, Springer, Cham, 2019.

[19] A. Khatija and N. Shajun, “Breast cancer data classification using SVM and Naive Bayes techniques,” *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 12, pp. 21167–21175, 2016.

[20] K. Polat and S. Güneş, “Breast cancer diagnosis using least square support vector machine,” *Digital Signal Processing*, vol. 17, no. 4, pp. 694–701, 2007.

[21] A. Watkins, J. Timmis, and L. Boggess, “Artificial immune recognition system (AIRS): an immune-inspired supervised learning algorithm:an immune-inspired supervised learning

algorithm," *Genetic Programming and Evolvable Machines*, vol. 5, no. 3, pp. 291–317, 2004.

[22] N. Uniyal, H. Eskandari, P. Abolmaesumi et al., "Ultrasound RF time series for classification of breast lesions," *IEEE Transactions on Medical Imaging*, vol. 34, no. 2, pp. 652–661, 2015.

[23] K. A. Mainzer, "Short history of the AI," *Artificial Intelligence-When Do Machines Take over?*, pp. 7–13, Springer, Berlin, Heidelberg, 2020.

[24] J. Baldwin-Philippi, "Data ops, objectivity, and outsiders: journalistic coverage of data campaigning," *Political Communication*, vol. 37, pp. 1–20, 2020.

[25] O. Kwon, "Very simple statistical evidence that AlphaGo has exceeded human limits in playing GO game," 2020, https://arxiv.org/abs/2002.11107.

[26] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis," *Science*, vol. 130, no. 3366, pp. 9–21, 1959.

[27] J. B. O. Mitchell, "Artificial Intelligence in Pharmaceutical Research and Development," *Elsevier Public Health Emergency Collection*, vol. 26, 2018.

[28] V. Vapnik, *The Nature of Statistical Learning theory*, Springer Science &Business Media, Berlin, Germany, 2013.