SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

e-Health Journal 2025, 1(1), 1172.
https://doi.org/10.62617/ehj1172

Article

Diabetes prediction based on an improved whale optimization algorithm and support vector machine

Xinyi Yang, Jingyi Yang, Xiaoyan Liu, Qiang Hu*

College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China
*** Corresponding author:** Qiang Hu, huqiang200280@163.com

CITATION

Yang X, Yang J, Liu X, Hu Q. Diabetes prediction based on an improved whale optimization algorithm and support vector machine. e-Health Journal. 2025; 1(1): 1172. https://doi.org/10.62617/ehj1172

ARTICLE INFO

Received: 19 November 2024
Accepted: 20 December 2024
Available online: 27 December 2024

Abstract: In recent years, the high incidence of diabetes and its complications has posed a significant threat to public health. To enhance the early diagnosis rate, this paper proposes a risk prediction model based on an improved Whale Optimization Algorithm (WOA) and Support Vector Machine (SVM). The model first uses a Gaussian kernel function to address the issue of nonlinear data being linearly separable in high-dimensional feature space. Then, it optimizes the Whale Optimization Algorithm by incorporating Tent mapping, opposition-based learning, nonlinear functions, and an adaptive inertia weight strategy. Subsequently, SVM parameters are optimized to determine the optimal penalty factor and kernel function parameters, improving the efficiency of SVM kernel parameter search. Finally, the optimized PAWOA-SVM model is applied to the diabetes dataset of Iraqi Medical Institutions and the Alibaba Cloud Tianchi Precision Medical Contest dataset for diabetes prediction and validation, achieving a prediction accuracy of 93.54%. The prediction results of the PAWOA-SVM model are compared with other commonly used models such as BP neural networks, PSO-SVM, and decision trees, demonstrating the superior predictive performance of the PAWOA-SVM model. Both the model's recognition accuracy and computational efficiency show considerable improvement. Therefore, this research provides an effective model that can assist doctors in making early judgments about prediabetes, thereby improving the diagnosis rate of prediabetes.

Keywords: diabetes; support vector machine; tent mapping; opposition-based learning; adaptive inertia weight strategy; whale optimization algorithm

1. Introduction

With the aging population, changes in disease types, and dietary structures, chronic diseases are becoming increasingly prevalent. Among them, diabetes, one of the most common global chronic diseases, has seen a rising incidence rate. If diabetes is not detected and treated early, it can lead to numerous acute or severe long-term complications as the condition progresses. Once complications occur, they are difficult to reverse and cause further damage to the body. According to the 2021 10th edition of the "Global Diabetes Map", approximately 319 million people worldwide are in the prediabetic stage, with an estimated 72.8 million prediabetic individuals in China [1].

The best way to prevent diabetes is through early screening., Studies have shown that individuals at high risk of diabetes who undergo long-term, appropriate daily interventions can effectively prevent the onset of the disease. Traditionally, diabetes diagnosis has relied on doctors' years of accumulated personal experience and lab or instrument results. However, these methods often include subjective factors, potentially delaying treatment. Additionally, with the growing number of

patients, diagnostic workloads increase, leading to fatigue and a higher risk of misdiagnosis or missed diagnosis. Given the shortcomings of traditional methods, some scholars have proposed the use of mathematical models to predict diabetes risk [2], which not only aids doctors in treatment but also evaluates the effectiveness of disease prevention [3].

Common data mining methods used for diabetes analysis include neural networks [4], decision trees [5], and support vector machines (SVM) [6], with SVM receiving more widespread research attention. SVM is an excellent classification method for information processing, particularly for solving problems with small samples, nonlinearity, high dimensions, and generalization. Compared to deep learning models, SVM requires fewer training samples to quickly achieve optimal prediction results, making it superior in terms of model simplicity and training efficiency. However, SVM's accuracy in predicting diabetes is slightly lower, and its training time is relatively slower [7]. Therefore, this paper aims to optimize SVM to improve prediction accuracy and reduce training time. Since penalty factors and kernel function parameters significantly impact SVM accuracy, identifying suitable parameters is crucial for building an effective SVM model.

This study proposes a prediction model based on an improved whale optimization algorithm (WOA) and SVM. The model uses a Gaussian kernel function to address the problem of making nonlinear data linearly separable in high-dimensional feature space. Then, the improved WOA [8] is introduced to determine the optimal penalty factor and kernel function parameters, enhancing the speed of parameter optimization in the SVM kernel. Finally, the optimized PAWOA-SVM model is applied to the diabetes dataset of Iraqi Medical Institutions and the Alibaba Cloud Tianchi Precision Medical Contest dataset for prediction and validation, achieving a prediction accuracy of 93.54%. The PAWOA-SVM model's prediction results are compared with other common models, including BP neural networks, PSO-SVM, and decision trees. The results demonstrate the superior predictive performance of the PAWOA-SVM model, which outperforms other models in both accuracy and efficiency. Therefore, this study establishes an effective model to assist doctors in making accurate judgments about prediabetes, thereby improving the diagnosis rate of prediabetes.

The rest of this paper is organized as follows: Section 2 introduces SVM. The Whale Optimization Algorithm is described in Section 3; in Section 4, we provide how to use the Whale Optimization Algorithm to optimize SVM to build a high-quality SVM model for diabetes prediction. Section 5 presents the experiments of this paper, verifying that the optimized SVM diabetes model is superior to the comparative model. Section 6 summarizes the entire paper.

2. Related models

2.1. Support vector machine

The Support Vector Machine (SVM) was first proposed in 1964 and saw rapid development after the 1990s, leading to a series of improvements and extensions. It has been widely applied in areas such as image recognition and text classification. SVM is a classification method based on a generalized portrait algorithm in pattern

recognition, with its early work stemming from Soviet scholars Vladimir N. Vapnik and Alexander Y. Lerner. In 1995, Corinna Cortes and Vapnik introduced a nonlinear SVM with a soft margin, applying it to handwritten digit recognition. Since its publication, this work has garnered significant attention and provided a valuable reference for SVM's development across various fields [9].

The core of SVM lies in the kernel function, which maps data into a high-dimensional space. The classification performance of SVM depends on the choice of kernel function [10], the kernel function parameter, and the penalty factor. The kernel function, mapping function, and feature space are inherently linked, with the choice of kernel function implicitly determining the mapping function and feature space. Adjusting the kernel function parameter alters the complexity of the sample data's subspace, thereby determining the maximum dimension for constructing the classification surface. The penalty factor adjusts the balance between the machine's confidence range and the kernel's empirical risk, ensuring that the trained SVM possesses strong generalization ability [11].

To achieve optimal SVM classification performance, it is crucial to select appropriate values for and. Traditional SVM parameter optimization methods include experimental methods, grid search [12], and gradient descent. However, these methods have become insufficient in terms of speed and accuracy in finding optimal solutions. In recent years, with the rise of artificial intelligence algorithms, many researchers have applied various metaheuristic algorithms to optimize SVM parameters, such as genetic algorithms [13] and PSO [14] algorithms, yielding promising results.

2.2. Whale optimization algorithm

Whales use three primary hunting behaviors [15]:

1) Encircling prey: The WOA assumes that the current optimal candidate solution is the prey, and other whales in the population slowly move towards this optimal position.
2) Bubble-net attack: Whales use spiral and upward movements to create bubbles, employing a bubble-net hunting strategy. By shrinking their encircling movements or spiraling, the whales get closer to the bait.
3) Random search: To enhance WOA's exploratory capability, whales no longer update their positions based on the current best prey but instead update their movements based on a randomly selected whale. If the random threshold is less than 1, the whale population searches for prey within a small range, finding more precise solutions locally. If the threshold is greater than 1, the search is global.

In WOA, each whale's position represents a solution to the problem, and the prey's position represents the optimal solution. During the exploitation phase, each whale randomly chooses either to encircle the prey or perform a spiral bubble-net attack, spiraling upward and blowing bubbles to trap the prey. During the exploration phase, the whales roam and search for food, moving away from the current optimal reference position, giving the algorithm exploratory capabilities.

3. Method

A diabetes prediction model based on WOA-Optimized SVM is proposed in our study. The data samples involved in this study are linearly inseparable, so they first need to be transformed into linearly separable data. This is achieved through the Gaussian Radial Basis Function (RBF) kernel of SVM for classification. The performance of SVM mainly depends on the kernel function parameter and the penalty factor. To optimize these parameters, this study introduces an improved Whale Optimization Algorithm (WOA).

To address the deficiencies of the standard WOA, enhancements are made by introducing Tent mapping [16], opposition-based learning [17], nonlinear functions [18], and adaptive inertia weight strategies [19]. These improvements increase the likelihood of the whale population finding the optimal prey, thereby improving the convergence performance and search capabilities of the WOA. The specific implementation steps are as follows:

Step 1: Tent mapping is used to generate chaotic sequences for initializing the population. Next, opposition-based learning is applied to the population derived from Tent mapping. The combination of populations generated by Tent mapping and opposition-based learning is then formed. After calculating and ranking the objective function value for each individual in the population, the top N individuals with the best fitness are selected as the initial population.

Step 2: An improved encircling mechanism is employed by replacing the convergence factor with a nonlinear Sigmoid function, enhancing the algorithm's global search ability in the early iterations and local search ability in the later iterations.

Step 3: In conjunction with the dynamic changes of the convergence factor e, adaptive weight, which changes synchronously with the nonlinear convergence factor, is introduced during local position updates. Additionally, an adaptive weight coefficient is incorporated.

Step 4: Fitness values are calculated, and the historical best position of each whale as well as the global historical best position of the whale population are identified.

Step 5: Cauchy mutation is introduced to determine if the algorithm has reached the predetermined number of iterations or found the global optimal solution [20]. If so, the process ends, and the results are output. If not, and if the current optimal solution has not been updated for a long time, the algorithm uses Cauchy perturbation to generate a larger step size to escape the local optimum, after which fitness values are recalculated.

For diabetes prediction, the process of enhancing SVM using the Whale Optimization Algorithm is illustrated in **Figure 1**.

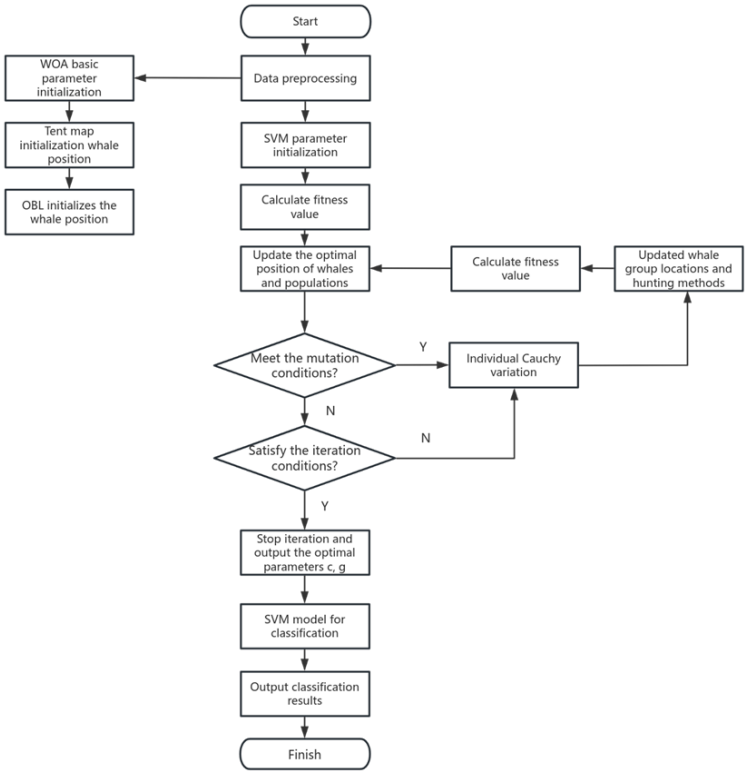

Figure 1. Process of improving the SVM-based diabetes prediction model using the whale optimization algorithm.

4. Experiment

4.1. Data collection and preprocessing

The experiments were conducted on two datasets: The diabetes dataset of Iraqi Medical Institutions and the Alibaba Cloud Tianchi Precision Medical Contest dataset.

The dataset on diabetes in Iraqi medical institutions was derived from research and surveys conducted in 2018 by the laboratories of the Iraqi Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. This dataset comprises 582 records, each containing 12 feature variables and one target outcome variable. The feature variables include:

1) Age and Gender of the individuals.
2) Creatinine Ratio (Cr), measured in mg/g.
3) Body Mass Index (BMI).
4) Urea and Cholesterol (Chol), measured in mmol/L.
5) Fasting Lipid Profile indicators: LDL, VLDL, Triglycerides (TG), and HDL Cholesterol, all measured in mmol/L.
6) HBA1C, expressed as a percentage (%).

The target outcome variable, Class, indicates whether an individual has diabetes: “Y” for diabetic and “N” for non-diabetic. **Table 1** summarizes these variables.

Table 1. Overview of the diabetes dataset of Iraqi medical institutions.

Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	Class
M	58	20.8	56	9.1	6.6	2.9	1.1	4.3	1.3	33	Y
M	60	2.1	45	7.6	3.3	1.7	0.9	1.7	0.8	36.6	Y
F	56	4	55	9.2	4.1	0.6	1.3	1.4	0.9	30	N
F	43	2.1	84	5.7	4.7	5.3	0.9	1.7	2.4	25	N
M	31	7	72	8.1	6	2.2	1.4	4	1	28	Y
F	42	10.3	46	11.5	4.4	2.1	2	2.5	0.9	26	N

The Tianchi Precision Medical Contest dataset was collected by Alibaba Cloud from September to October 2017. It contains a total of 5642 data points, with 41 feature variables. These variables include patients' basic information, liver function indicators, biochemical markers, five hepatitis B-related indicators, blood routine data, and blood glucose as the target variable. The description related to Tianchi Precision Medical Contest dataset can be referred to in **Tables 2** and **3**.

Table 2. Feature variable categories in the Tianchi precision medicine competition dataset.

Feature category	Feature Name
Basic situation indicators	Gender, age, and date of medical examination
Liver function indicators	Aspartate aminotransferase, alkaline phosphatase, r-glutamyl transferase, alanine aminotransferase, total protein, globulin, albumin, white sphere ratio
Biochemical indicators	Triglycerides, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, total cholesterol, creatinine, uric acid, urea,
Five indicators of hepatitis B	Hepatitis B surface antigen, hepatitis B surface antibody, hepatitis B e antigen, hepatitis B e antibody, hepatitis B core antibody
Blood routine indicators	White blood cell count, red blood cell count, hemoglobin, average hemoglobin content of red blood cells, hematocrit, red blood cell volume distribution width, platelet count, average red blood cell volume, platelet volume distribution width, average hemoglobin concentration of red blood cells, average platelte volume, platelet specific bvolume, eosinophil%, lymphocyte%, neutrophil%, Monocytes%, basophils%

Table 3. Overview of the Tianchi precision medicine competition dataset.

albumin	triglyceride	creatinine	uric acid	total protein	total cholesterol	alkaline phosphatase	blood sugar
49.6	1.31	77.25	349.39	76.88	4.43	99.59	6.06
47.76	2.81	87.12	486.78	79.43	4.06	67.21	5.39
48	0.99	78.19	452.07	86.23	4.13	63.69	5.59
44.02	1.06	61.46	368.85	70.98	6.89	74.08	4.3
41.83	0.97	66.66	383.87	78.05	5.37	75.79	5.42

Different types of data often have varying magnitudes, and the differences in values can be significant. If raw data is directly input into the model, it may disproportionately influence the model's performance due to these variations in

scale. Therefore, to effectively utilize the data, normalization [21] is required. All 768 input records were normalized using the zero-mean method, mapping the original data into a dataset with a mean of 0 and a variance of 1. The normalization formula is as follows:

$$\chi = \frac{\chi - \mu}{\sigma}$$

where and represent the mean and standard deviation of the original dataset, respectively.

4.2. Model validation and analysis

To validate the accuracy of the PAWOA-SVM model, 80% of the diabetes dataset of Iraqi Medical Institutions was used as the training set, and the remaining 20% as the test set. And similarly, 80% of the Tianchi Precision Medicine Competition dataset was used as the training set, and the remaining 20% as the test set. The Gaussian Radial Basis Function (RBF) kernel was selected as the kernel function. The performance of the PAWOA-SVM model was compared with that of BP neural networks, PSO-SVM, and decision tree models. In each experiment, the training and test sets were randomly generated in proportion, and 30 rounds of training and testing were conducted. To compare the models, four commonly used evaluation metrics were adopted: Accuracy, recall, AUC (Area Under the Curve), and F1 score. A detailed comparison of the four models is provided in **Table 4**.

As shown in **Tables 4** and **5**, the PAWOA-SVM model demonstrates the highest reliability, with the highest accuracy rate and AUC value, indicating the best classification performance. The model's prediction results are highly accurate. Overall, the PAWOA-SVM model outperforms other models, making it a suitable tool for diabetes prediction.

Table 4. Comparison of results from four models on Iraqi dataset.

Method	Accury	Recall	AUC	F1
BP Neural Network	88.51%	86.45%	0.864	0.89
Decision Tree	87.43%	84.62%	0.857	0.84
PSO-SVM	92.17%	86.28%	0.844	0.84
PAWOA-SVM	92.53%	92.71%	0.912	0.90

Table 5. Comparison of results of four models based on Tianchi dataset.

Method	Accury	Recall	AUC	F1
BP Neural Network	88.36%	88.43%	0.843	0.86
Decision Tree	87.19%	84.57%	0.816	0.84
PSO-SVM	92.22%	88.21%	0.812	0.84
PAWOA-SVM	93.25%	92.68%	0.935	0.88

5. Discussion

Although we have conducted relevant research to improve the whale optimization algorithm and the support vector machine prediction model, which has

led to improved prediction accuracy, there are still some limitations. The following issues can be further explored and studied:

The datasets used in this paper are static. To gain a more comprehensive and in-depth understanding of the development and changes in diabetes, future research should focus on enhancing the ability to mine time series data characteristics. This would provide more accurate and comprehensive data support for the early prediction of diabetes.

In the future, the prediction model proposed in this paper could be applied to the development of a diabetes auxiliary diagnosis and management system. A system platform could be established to create a medical database, facilitate data collection, updates, and interactions, enabling real-time monitoring and prediction of diabetes patients and high-risk groups, thereby improving diabetes management in China.

The model presented in this article can be applied to disease prediction in datasets with similar features, such as cancer prediction based on these features. By effectively utilizing big data technology and analyzing patients' medical data with their knowledge and consent, we can help identify and eliminate potential disease risks for patients.

6. Conclusion

The high incidence of diabetes and its complications poses a serious threat to public health. To achieve early detection, diagnosis, and treatment of diabetes and ultimately reduce mortality rates, this study proposed a diabetes risk prediction model based on an improved Whale Optimization Algorithm (WOA) and Support Vector Machine (SVM). Using the diabetes dataset of Iraqi Medical Institutions and Tianchi Precision Medicine Competition dataset, the model optimized the penalty factor and kernel parameters of SVM. Tent mapping, opposition-based learning, nonlinear functions, and adaptive inertia weight strategies were introduced into the standard WOA, effectively reducing the likelihood of the whale population getting trapped in local optima and improving the accuracy of the optimization algorithm.

Author contributions: Conceptualization, XY; methodology, XY; writing, XY; experiment, JY and XL, supervision, QH. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Whiting DR, Guariguata L, Weil C, et al. IDF diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes research and clinical practice. 2011; 94(3): 311–321.
2. Jiang R, Xin Y, Chen Z, et al. A medical big data access control model based on fuzzy trust prediction and regression analysis. Applied Soft Computing. 2022; 117: 108423.
3. Jiang R, Han S, Yu Y, et al. An access control model for medical big data based on clustering and risk. Information Sciences. 2023; 621: 691–707.
4. Srivastava S, Sharma L, Sharma V, et al. Prediction of diabetes using artificial neural network approach. Springer Singapore Publishing; 2019. pp. 679–687.
5. Dudkina T, Meniailov I, Bazilevych K, et al. Classification and Prediction of Diabetes Disease using Decision Tree Method. IT&AS Publishing; 2021. pp. 163–172.

6. Akbulut FP, Akan A. Support vector machines combined with feature selection for diabetes diagnosis. IU-Journal of Electrical & Electronics Engineering. 2017; 17(1): 3257–3265.
7. Lin P. Research on prediction model of diabetes based on SVM. Jilin University; 2022.
8. Liu L, Zhang R. Multistrategy improved whale optimization algorithm and its application. Computational Intelligence and Neuroscience. 2022; 1: 3418269.
9. Veisi H. Introduction to SVM. In: Rad JA, Parand K, Chakraverty S (editors). Learning with Fractional Orthogonal Kernel Classifiers in Support Vector Machines: Theory, Algorithms and Applications. Springer Nature Singapore Publishing; 2023. pp. 3–18.
10. Azzeh M, Elsheikh Y, Nassif AB, et al. Examining the performance of kernel methods for software defect prediction based on support vector machine. Science of Computer Programming. 2023; 226: 102916.
11. Dai T, Dong Y. Introduction of SVM related theory and its application research. In: Proceedings of the 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE); 24–26 April 2020; Shenzhen, China.
12. Yao L, Fang Z, Xiao Y, et al. An intelligent fault diagnosis method for lithium battery systems based on grid search support vector machine. Energy. 2021; 214: 118866.
13. Resmini R, Silva L, Araujo AS, et al. Combining genetic algorithms and SVM for breast cancer diagnosis using infrared thermography. Sensors. 2021; 21(14): 4802.
14. Choubey DK, Tripathi S, Kumar P, et al. Classification of Diabetes by Kernel based SVM with PSO. Recent Advances in Computer Science and Communications. 2021; 14(4): 1242–1255.
15. Chakraborty S, Saha AK, Chakraborty R, et al. An enhanced whale optimization algorithm for large scale optimization problems. Knowledge-Based Systems. 2021; 233: 107543.
16. Godbole V, Gaikwad S. An Exponential Map-based Whale Optimization Algorithm (Exp-WOA) for Optimization. ECTI Transactions on Computer and Information Technology. 2024; 18(4): 443–455.
17. Wang X, Hu J, Hu J, et al. A modified equilibrium optimizer using opposition-based learning and teaching-learning strategy. IEEE Access. 2022; 10: 101408–101433.
18. Pan Z, Gu Z, Jiang X, et al. A modular approximation methodology for efficient fixed-point hardware implementation of the sigmoid function. IEEE Transactions on Industrial Electronics. 2022; 69(10): 10694–10703.
19. Zhang J, Wang JS. Improved whale optimization algorithm based on nonlinear adaptive weight and golden sine operator. IEEE Access. 2020; 8: 77013–77048.
20. Zhao X, Fang Y, Liu L, et al. A covariance-based Moth–flame optimization algorithm with Cauchy mutation for solving numerical optimization problems. Applied Soft Computing. 2022; 119: 108538.
21. Singh D, Singh B. Investigating the impact of data normalization on classification performance. Applied Soft Computing. 2020; 97: 105524.