Expert Systems With Applications 219 (2023) 119648

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

基于SOLSSA-catboost模型的心血管疾病风险评估

Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model

麻雀搜索算法

Xi Wei [a], Congjun Rao [a,*], Xinping Xiao [a], Lin Chen [b], Mark Goh [c]

[a] School of Science, Wuhan University of Technology, Wuhan 430070, China
[b] School of Management, Wuhan Institute of Technology, Wuhan 430205, China
[c] NUS Business School & The Logistics Institute-Asia Pacific, National University of Singapore, 119623, Singapore

ARTICLE INFO

Keywords:
CVD
Risk assessment
CatBoost model
SOLSSA
SOLSSA-CatBoost model

ABSTRACT

Cardiovascular disease (CVD) has become a significant public health problem affecting national economic and social development, and ranks among the top causes of death in the world. Thus, people pay increasing attention to the prevention, control, and risk assessment of CVD. In this paper, an improved sparrow search algorithm (SSA) is designed to optimize the parameters of Categorical Boosting (CatBoost) model, and it is applied to the risk assessment of CVD. The contributions of this research are mainly in the following aspects: (1) In the position update formula of the discoverer, the salp swarm algorithm is integrated, the global optimal solution of the previous generation is added to improve the global search ability and local development ability of SSA; (2) Using Opposition-based Learning (OBL) and Lateral mutation strategy to improve the search ability of the worst individual; (3) Sparrow search algorithm based on salp swarm algorithm, OBL and Lateral mutation strategy (SOLSSA) is used to optimize parameters of CatBoost to improve the prediction effect, and the experiments are carried out for the proposed model (SOLSSA-CatBoost) using two CVD data sets on Kaggle. The proposed model is compared with six machine learning models, including random forest (RF), logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), light gradient Boosting (LGB) and CatBoost, and is also compared with other four optimization algorithms (whale optimization algorithm (WOA), gray wolf algorithm (GWO), seagull optimization algorithm (SOA) and SSA) in optimizing the performance of the CatBoost. The experimental results show that compared with other comparison algorithms, SOLSSA-CatBoost has better prediction effect on test set, with F1-score reaching 90% and 81.51% in two CVD data sets respectively. The SOLSSA-CatBoost model in this paper can make a more accurate prediction of patients' disease risk, and provide a certain basis for doctors to judge the condition.

1. Introduction

The burdens of diseases in China have changed from infectious diseases and nutritional deficiency diseases to chronic diseases. As is known to all, chronic diseases last for a long time and are difficult to cure, which is related to the negative health effects in terms of quality of life and economic burden (Wang, Rao, Goh, & Xiao, 2023; Yach, Hawkes, Gould, & Hofman, 2004). Cancer, cardiovascular and cerebrovascular diseases and so on are common chronic diseases. According to statistics, people who die from chronic diseases account for up to 70 % of the total deaths worldwide, and cardiovascular and cerebrovascular diseases are the most important causes of death. Cardiovascular and cerebrovascular diseases are ischemic or hemorrhagic diseases occurring in the heart, brain and whole body tissues, which are caused by the combined action of many risk factors in many dimensions such as economy, society, behavior, environment and population. The mortality rate of cardiovascular disease (CVD) has been in the forefront for many years. Its mortality rate is higher than that of diseases with high mortality, such as tumors, and it has become the number one health killer, so it is urgent to strengthen the prevention and treatment of CVD. It is estimated that up to 90 % of CVD is preventable, and CVD is in a position to be identified by wearable sensors such as electrocardiogram (ECG) or heart sound sensors, or medical testing in hospital (Ali, El-Sappagh, Islam, Kwak, Ali, Imran, & Kwak, 2020; Saeed et al., 2018). However, in reality, CVD is a high-dimensional complex system accompanied by a large number of heterogeneous data, correct diagnosis is not easy to make and is often delayed due to the many factors complicating disease diagnosis. For example, clinical symptoms, functional, and pathologic

* Corresponding author.
E-mail addresses: xiweiwhut@163.com (X. Wei), cjrao@163.com (C. Rao), xiaoxp@whut.edu.cn (X. Xiao), linchen@wit.edu.cn (L. Chen), bizgohkh@nus.edu.sg (M. Goh).

https://doi.org/10.1016/j.eswa.2023.119648
Received 21 February 2022; Received in revised form 15 January 2023; Accepted 30 January 2023
Available online 3 February 2023

manifestations of heart disease are often associated with many other human organs besides the heart itself, and often heart disease may show diverse syndromes. Furthermore, different types of heart disease can have similar symptoms, further complicating diagnosis (Yan, Jiang, Zheng, Peng, & Li, 2006). Therefore, diagnosing CVD reliably is an arduous task.

At present, our country's ability to control CVD is weak. People don't know much about CVD, they don't know the potential pathogenic factors (diet, exercise, etc.), and they cannot timely deal with CVD. Therefore, it is necessary to predict and diagnose CVD, which is also one of the major challenges of modern medicine (Saba, Parodi, & Ganau, 2021). Risk assessment of CVD, on the one hand, can detect potential risks and trends in advance so as to screen out the high-risk groups and carry out risk factors intervention. On the other hand, it can better guide individuals and make high-risk groups change their life behaviors, thus reducing the risk of illness, which plays a certain guiding role in the prevention of CVD. At present, the misdiagnosis rate of manual diagnosis at the grass-roots level in China is high, in addition, due to the lack of suitable classification models, many established models cannot achieve high prediction accuracy. Therefore, establishing an appropriate classification model and a reliable new risk assessment mechanism can provide more scientific intervention measures and precise personalized treatment schemes for the prevention and treatment of CVD, and effectively improve the quality and efficiency of CVD risk management.

Recently, the research on the treatment of CVD has always been one of the fields of attention, whether it is drugs for treating CVD, medical devices to assist patients' rehabilitation, or the artificial intelligence with an increasingly profound influence on the medical field, more and more attention has been paid (Qian et al., 2022; Rosenson & Marcovina, 2018). In the early days, some researchers conducted research on the prediction of CVD based on limited and clear risk factors, such as age and hypertension. Goff et al. (2014) formulated a quantitative risk assessment method that could be used to guide nursing care, put forward and solved a few issues considered crucial to improve and adopt risk assessment in clinical practice; Long-term exposure to cigarette smoke will lead to an increase in the incidence of CVD, and the formation of thrombosis is the main cause of the occurrence and death of CVD. Therefore, Liu, Zhang, Wang, Xu, Li and Fang (2016) used atomic force microscopy single molecular force spectroscopy to study the main components in cigarettes and their metabolites and the influence of the interaction between thrombin and thrombomodulin; Comparison was made by Paynter et al. (2014) in a multi-ethnic non-smoking female cohort based on lifestyle and traditional CVD prediction, and they concluded that adding a healthy lifestyle to the traditional risk score could improve the overall degree of fitting of prediction model. However, these previous studies have insufficient predictive performance, and it is difficult to achieve personalized accurate prediction. With people's increasing emphasis on health, the development of national medical equipment and the progress of doctors' diagnosis, more and more CVD is diagnosed through various examinations (ECG and various imaging examinations, etc.) and clinical medical record data. However, CVD is a complex multidimensional system. To predict accurately, it is necessary to obtain a large number of case data, which depends on experienced physicians, but the number, energy and the speed of data processing of doctors are limited.

With the rapid development of emerging science and technology, we are facing the era of big data, and Machine Learning (ML) has developed rapidly in various fields (Tian, 2015). ML is an automatic and convenient mode that can help clinicians to deal with problems procedurally, and has been included in the clinical practice of CVD prediction (Zhao, Wood, Mirin, Cook, & Chunara, 2021). Commonly used ML methods include RF (Rao, Liu, Goh, & Wen, 2020), SVM (Cao & Tay, 2001; Rao et al. 2023), KNN (Saini, Singh, & Khosla, 2013; Wang, Pan, & Dong, 2022), and so on. In order to alleviate the above problems, experts have introduced several clinical decision support systems by using ML models to recognize various symptoms of cardiovascular patients and take corresponding actions in time. Ambale-Venkatesh et al. (2017) tested the ability of a ML technology (Random Survival Forest) to predict six cardiovascular outcomes compared with the standard cardiovascular risk score by using data from 6814 participants in the Multi-Ethnic Study of Atherosclerosis (MESA). The result showed that the Random Survival Forest technology performed better than established risk score. Tokodi et al. (2020) developed a risk stratification system based on ML to predict the mortality of patients undergoing cardiac resynchronization therapy (CRT). For the prediction of mortality at one to five years, among the trained classifiers, the RF performed best, with AUC values of 0.768, 0.793, 0.785, 0.776 and 0.803 respectively. Cikes et al. (2019) used complex ECG data and clinical parameters to group the phenomena of heart failure cohort by ML algorithm, and identified the patients who had beneficial responses to CRT. In 2021, network analytics and machine learning algorithms are employed to develop predictive models for CVD in patients with type 2 diabetes (T2D) conditions (Hossain, Uddin, & Khan, 2021). Tay, Poh, Van Reeth, and Kitney (2015) used SVM algorithm to establish a prediction model of myocardial infarction, and studied it on the cardiovascular health research dataset with good effect. Segar et al. (2021) developed a race-specific model using ML to predict the risk of heart failure within 10 years. In the process, RF interpolation was used to fill various data sets and maximum likelihood method was used to identify important predictors.

Boosting algorithm is an integrated algorithm. Boosting technology can play an important role in the parameter tuning system with limited training data, short training time and little professional knowledge (Alsahaf, Petkov, Shenoy, & Azzopardi, 2022; Rao, Gao, Wen, & Goh, 2022a). Commonly used Boosting include light gradient Boosting machine (LGB), extreme gradient Boosting (XGBoost) and so on (Bentejac, Csorgo, & Martinez-Munoz, 2021; Rao, Liu, & Goh, 2022b). Many researchers have employed Boosting to resolve problems. Finlay (2011) compared the performance of various models after integration, and compared with a single model. Finally, he found that the way of integrating combined models can obviously improve the generalization ability of models. At present, integrated learning technology has also been widely used in the medical field (Jiang et al., 2021; Ma, Meng, Yan, Yan, Chai, & Song, 2020). CatBoost is the latest representative so far, and as the latest model within the framework of gradient boosting decision tree (GBDT), it has advantages in dealing with classified variables and producing reliable results (Zhang, Zhao, & Zheng, 2020). Some scholars tried to use CatBoost to predict disease risk. Lee et al. (2021) used CatBoost model to predict the potential life-threatening intracranial hypertension (LTH) events in the future, and the continuous records of intracranial pressure and arterial blood pressure from 273 traumatic brain injury (TBI) patients were used as a development data set. Al-Absi et al. (2021) used CatBoost model distinguish healthy individuals from people having CVD. The results showed that CatBoost model achieved 93 % accuracy, and outperformed the existing model for the same purpose.

Although the ML methods such as KNN and SVM mentioned above have achieved good results in CVD prediction, they belong to a single global optimization model, which has limited performance, poor robustness and poor fault tolerance. Compared with a specific classifier, the classifier based on ensemble learning has better advantages in complex classification tasks. Therefore, in the prediction of CVD risk, CatBoost have gradually gained the favor of scholars because of its high efficiency and reasonable processing of category features, strong generalization ability and high accuracy. Li, Zhang, Xiong, Hu, Liu, Tu, and Yao (2022) used CatBoost predict hospital mortality in mechanically ventilated patients with congestive heart failure (CHF). However, CatBoost model is not highly explanatory, its different parameters combinations may have different prediction results. For this reason, some scholars have started to try to optimize CatBoost using swarm intelligence algorithms. Zhang, Chen, Zhang, Liu, Yu, Zhang, and Gao (2021) adopted difference-mutation brain storm optimization (DBSO) algorithm to optimize CatBoost model and got better prediction results.

Swarm intelligence optimization algorithm is a stochastic optimization technology, which opens up a new way for solving optimization problems (Liu, Guo, Wang, & Zhang, 2022). The research in the field of swarm intelligence is exploding, more and more researches combine ML methods with swarm intelligence optimization algorithms to improve the performance and efficiency of ML methods (Tharwat & Schenck, 2021). Every year, countless new swarm intelligence algorithms are proposed to solve many optimization problems, such as moth-flame optimization (MFO) (Mirjalili, 2015), whale optimization algorithm (WOA) (Mirjalili & Lewis, 2016), gray wolf algorithm (GWO) (Sm, Smm, & Al, 2014), seagull optimization algorithm (SOA) (Dhiman & Kumar, 2019) and so on. In 2018, particle swarm optimization-based support vector machine (PSO-SVM) classification model was employed to implement the identification of multiple fault condition of rolling bearing (Yan & Jia, 2018). Some scholars have further improved the optimization algorithm (Gupta & Deep, 2020; Jiang, Li, & Huang, 2013; Li, Xiong, Tseng, Yan, & Lim, 2022).

The sparrow search algorithm (SSA) is a simple and efficient swarm intelligence optimization algorithm proposed in 2020 (Xue & Shen, 2020). It conducts global and local searches by imitating the foraging and anti-predation behaviors of sparrows. Compared with other algorithms, SSA has the characteristic of higher solving efficiency, however, like other swarm intelligence algorithms, it is easy to fall into the situation of local extremum in the late iteration. Many scholars proposed different improvement strategies for the SSA. In 2021, Liu, Shu, Liang, Peng, and Cheng (2021) used a chaotic strategy to make the population more random and balance the convergence speed and detection ability of the algorithm. Tian and Chen (2021) proposed an improved SSA to optimize its parameters when using long and short-term memory (LSTM) model to predict wind speed of wind power plant. Fu and Liu (2022) proposed an improved SSA based on multi-strategy fusion. Firstly, the elite chaotic OBL strategy was used to generate the initial population, then the random following strategy of chicken swarm algorithm was used to optimize the followers in SSA, and Cauchy-Gaussian mutation strategy was used to maintain the population diversity. Finally, 10 benchmark test functions were used to test the superiority of the improved algorithm. He et al. (2022) used sinusoidal chaotic mapping to initialize sparrow population, introduced adaptive *T*-distribution to update the sparrow position, and optimized the parameters of variational modal decomposition (VMD) with the improved SSA. The results showed that the SSA had the characteristic of rapid convergence.

To sum up, CatBoost has advantages in dealing with classified variables and producing reliable results, disease data are often accompanied by classification characteristics, so CatBoost is suitable for cardiovascular disease risk prediction. Compared with single machine learning model, CatBoost is robust. In fact, for the CatBoost model, part of CatBoost's parameters are not highly interpretable, and CatBoost with different parameters has a great influence on the prediction effect of CatBoost. Thus, in order to make CatBoost model reach the optimal state after modeling, it is necessary to optimize parameters. SSA have the characteristics of fast search speed and wide search range, therefore, this paper tries to use SSA to optimize the parameters of CatBoost.

As mentioned earlier, SSA has a low population complexity and may fall into local convergence. Many scholars have made improvements to SSA, through further research, there are still some algorithms with insufficient optimization accuracy, and few literatures improve the positions of the relatively backward sparrow individuals. Therefore, we propose an improved sparrow search algorithm (SOLSSA) by improving individuals combining salp swarm algorithm, opposition-based learning (OBL) and Lateral mutation. In addition, in order to enhance the generalization ability of the prediction system, this paper uses K-fold cross-validation when optimizing parameters with SOLSSA. In summary, this paper proposes a CVD risk prediction method based on SLOSSA-CatBoost model to achieve better prediction effect. The main contributions of this paper are as follows:

(i) CatBoost is an ultramodern Boosting algorithm with high efficiency and strong generalization ability. CatBoost not only has a unique way to deal with categorical data, but also reduces the chance of overfitting and makes the model more universal. Therefore, it is preponderant to predict cardiovascular risk with CatBoost.
(ii) In the position update formula of the discoverer, the salp swarm algorithm is integrated, then the global optimal solution of the previous generation is added to coordinate the capabilities of local mining and global exploration.
(iii) Combining OBL and Lateral mutation strategy, perturbation mutation is carried out at the worst position to strengthen the information exchange between the worst individual and other better individuals.
(iv) Combining with relevant data sets, we compare the prediction performance of SOLSSA with other machine learning models, including RF, LR, KNN, SVM, LGB and CatBoost, the results show that SOLSSA proposed in this paper has good effect. We also compare the prediction performance of SOLSSA-CatBoost with other state-of-the-art optimization algorithms, including GWO-CatBoost, WOA-CatBoost, SOA-CatBoost and SSA-CatBoost. The results show that SOLSSA proposed in this paper has good effect, its optimization effect on CatBoost is the most prominent.

The rest of this paper is organized as follows: Section 2 puts forward a new method of CVD risk assessment based on SOLSSA-CatBoost model. Combined with two data sets of Kaggle website, the proposed SOLSSA-CatBoost model is empirically analyzed and compared with other models in Section 3. In Section 4, the experimental results and analysis are provided. Finally, Section 5 summarizes the work of this paper and puts forward the prospects.

2. Risk assessment method of CVD based on SOLSSA-CatBoost model

In this section, we propose a new CVD risk prediction model named SOLSSA-CatBoost based on ML and swarm intelligence algorithm. The paper uses CatBoost model distinguish healthy individuals from people having CVD. However, improper selection of parameters of CatBoost model will easily affect the prediction accuracy, so we propose an improved SSA (SOLSSA) to provide reliable prediction of CVD. When improving SSA, firstly, in the position update formula of the discoverer, the salp swarm algorithm is integrated. Secondly, the OBL strategy and Lateral mutation are used to improve the searching ability of the worst individual. At length, the proposed SOLSSA is applied to the parameters optimization of CatBoost model so as to obtain more accurate disease prediction results, which can help doctors to make more comprehensive and accurate decisions, thus reducing the misdiagnosis rate.

2.1. CatBoost model

CatBoost is a new machine learning model based on GBDT algorithm proposed by Dorogush et al. in 2017. Compared to earlier GBDT algorithm, CatBoost performs better in terms of speed and accuracy. CatBoost's main advantages lie in dealing with classified features efficiently and reasonably, and solving the problems of gradient deviation and prediction deviation, thus effectively avoiding the overfitting problem and further improving the accuracy and generalization ability.

CatBoost uses Ordered Boosting method to change the biased gradient estimation to unbiased in GBDT. The Ordered Boosting method is to randomly generate a $[1, n]$ arrangement σ to sort the original samples and initialize n different models $M_1, M_2, \ldots, M_n$, then each M_i only uses the top i samples in random arrangement to learn the model. Furthermore, in each iteration process, the unbiased gradient estimation of the j-th sample is obtained by the model $M_{j\text{-}1}$. CatBoost's processing method for categorical features is to scramble data set $D = \{(x_i, y_i)\}$, $i =$

1, 2, …, *n*., the new sequence is to traverse the sequence $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$. When the category feature of each sample is converted into numerical value, the value of this feature is the average value of the category feature of the first *p* records traversed, and the priority and the weight coefficient of priority are added at the same time.

$$\sigma_{p,k} = \frac{\sum_j^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_i,k}\right]\cdot Y_{\sigma_j} + a\cdot p}{\sum_j^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_i,k}\right] + a} \tag{1}$$

where *p* represents the prior term, and $a > 0$ represents the weight coefficient of the prior term. For the dichotomy problem of predicting CVD, the prior term is the average of patients in the CVD data set. $[\cdot]$ is an indicator function, and if the internal conditions are met, it will output 1, which means suffering from CVD. Otherwise, it will output 0, which means there is no illness.

In the classification and prediction of CVD, CatBoost model selects age, gender, resting blood pressure and other characteristics of the data set as the input of the model, whether the patient is ill or not is the output of the model. Among the variables closely related to CVD, most of them are categorical variables (gender, sports, etc.), and CatBoost has the unique advantage of automatic processing of categorical variables, so there is no need to quantify these variables and normalize the data, directly input them as categorical variables. However, the setting of different parameters in CatBoost has a certain influence on the prediction results of the model, so we can optimize the parameters to obtain better prediction results. The selection of parameters is based on three principles, namely, faster training speed, higher prediction accuracy and prevention of overfitting. Iterations of CatBoost represents maximum number of trees; Learning_rate of CatBoost can control the performance and complexity of the model; Depth of CatBoost is the main parameter that determines the prediction accuracy and helps prevent overfitting. Therefore, this paper chooses to optimize these three parameters.

2.2. SOLSSA

2.2.1. Sparrow search algorithm

Sparrow search algorithm (SSA) was proposed in 2020 (Xue & Shen, 2020), inspired by sparrows' predation behavior and anti-predation behavior. This algorithm is relatively novel with strong optimization ability and fast convergence speed. Assuming that there are *N* sparrows in the *D*-dimensional solution space, the position of the *i*-th sparrow in *D*-dimensional solution space is $X_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$, and the sparrow's fitness value is $F_{Xi} = f(x_{i1}, x_{i2}, \ldots, x_{id})$. Generally speaking, 10 %–20 % of the sparrows with high fitness values in each iteration are selected as discoverers, the rest are participants, and 10 %–20 % of individuals are randomly selected from the population as scouts. Discoverers usually have high energy and are responsible for searching for areas with abundant food throughout the population, acting as guides and providing foraging zones and directions for all participants. Scouts rely on the anti-predation strategy to avoid population falling into local optimum.

In the sparrow search algorithm, firstly, the population position is initialized by searching the upper and lower bounds of the space, and the formula is as follows:

$$x_{i,j}^0 = lb_j + (ub_j - lb_j)\cdot rand \tag{2}$$

where $x_{i,j}^0$ is used to represent the initial value of the *i*-th individual in the *j*-th dimensional space. ub_j and lb_j are the upper and lower bounds of the search space, *rand* is a random number from 0 to 1. However, with too much uncertainty, random distribution may occasionally produce an initial position that is unfavorable to individuals.

A good initial population will affect the process of optimization algorithm to find the global optimum, such as accelerating population convergence and improving the accuracy of the final solution. In this paper, the uniform function is used to initialize the population, so that individuals are evenly distributed in the population (see Fig. 1). Uniform distribution means that the individuals in the population are equidistantly distributed, or keep a certain uniform distance between individuals, and the evenly distributed population is beneficial to improve the early detection performance of SSA and hopefully it will converge to a better solution.

In SSA, discoverers have higher fitness values and faster access to food. They assume the responsibility of searching for food for the whole population and providing foraging directions for participants. The location update formula of the discoverers is:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}\cdot\exp\left(-\frac{i}{\alpha\cdot iter_{\max}}, R_2 < ST\right) \\ X_{i,j} + Q\cdot L, R_2 \geqslant ST \end{cases} \tag{3}$$

where $iter_{max}$ is the maximum number of iterations. $j = 1,2,\ldots,$ *PDNumber*, and *PDNumber* indicates the number of discoverers we set. α is a random number with a value in [0, 1]. R_2 ($R_2 \in [0,1]$) and ST ($ST \in [0.5,1]$) represent early-warning value and safety value respectively. *Q* is a random number that obeys normal distribution, and *L* represents a $1 \times d$ matrix, where all elements in the matrix are 1. When $R_2 < ST$, it means that there are no predators around the foraging environment at this time. On the contrary, it means that some sparrows have spotted predators and warned other sparrows in the population, at which point all sparrows will fly to other safe places to feed.

During the foraging process, some participants will always monitor the discoverers to prevent it from swallowing alone. Once they realize that the discoverers have found better food, they will immediately compete for food. If they win, they can get the food, otherwise they will fly off to other places to continue foraging. The location update of the participants is described as follows:

$$X_{i,j}^{t+1} = \begin{cases} Q\cdot\exp\left(\frac{X_{worst} - X_{i,j}^t}{i^2}\right), & i > n/2 \\ X_P^{t+1} + \left|X_{i,j} - X_P^{t+1}\right|\cdot A^+\cdot L, & otherwise \end{cases} \tag{4}$$

where *n* represents the number of participants, X_p is the best position currently occupied by the discoverer, and X_{worst} represents the worst position in the whole world at present. *A* represents a $1 \times d$ matrix with elements of 1 or −1, and $A^+ = A^T(AA^T)^{-1}$. When $i > 2/n$, this indicates that the sparrows with poor fitness value are very hungry and need to fly to other places to feed for more energy.

When sparrows are on the edge of the population, they are vulnerable to predators. If sparrows are in the current optimal position, they will flee to a position near themselves to be close to other sparrows to reduce the risk of being preyed on. Calculate the scouts' locations according to Eq. (5):

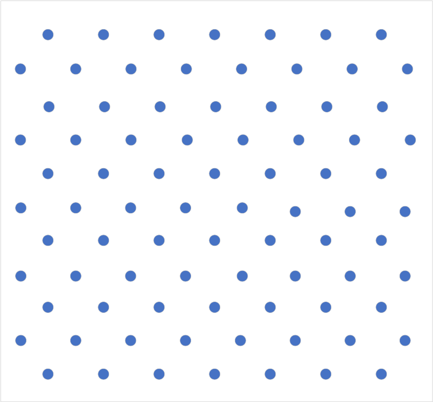


Fig. 1. Population initial distribution map.

$$X_{i,j}^{t+1} = \begin{cases} X_{best}^{t} + \beta \cdot \left| X_{i,j}^{t} - X_{best}^{t} \right|, f_i > f_g \\ X_{i,j}^{t} + K \cdot \left(\dfrac{\left| X_{i,j}^{t} - X_{worst}^{t} \right|}{(f_i - f_w) + \varepsilon} \right), f_i = f_g \end{cases} \tag{5}$$

where X_{best} is the current global optimal position and t represents the number of iteration. β is a step size control parameter, which is a random number that obeys the normal distribution with mean value of 0 and variance of 1. K is a random number between -1 and 1, f_i is the fitness value of the sparrow at present, f_g and f_w are the best and worst fitness values of the current global respectively. ε is a constant to avoid 0 value in the denominator.

However, the original SSA suffers from the disadvantages of reduced population diversity and the tendency to fall into local optimal solution. Therefore, this paper makes relevant improvement research work from two angles of the discoverers position update and the worst participant position update to improve the solution accuracy.

2.2.2. *The position update of the discoverers based on the salp swarm algorithm*

In the original SSA, the discoverers position update formula is only related to the position of the individual in the previous iteration and iteration times. For Eq. (3), in the case of $R_2 < ST$, the dimension of each discoverer in the sparrow population gradually decreases after multiplying by $\exp(-i/(\alpha \cdot iter_{\max}))$ (see Fig. 2), therefore the diversity of sparrow population becomes smaller and smaller in the later stage of the algorithm, which tends to cause the algorithm to fall into local optimum. To solve this problem, the location update method of leader in salp swarm algorithm is introduced into the location update formula of the discoverers in SSA to enhance the diversity of sparrow population and improve the optimization ability and convergence accuracy of the algorithm.

The salp swarm algorithm is a group intelligent optimization algorithm which simulates the group behaviors of the salp chain (Elattar & ElSayed, 2020; Zhang, Wang, Chen, Heidari, Wang, & Zhao, 2021). In the salp chain, it is divided into two types: the leaders and followers. During iteration, the leaders move towards food in a chain behavior, which is called the global search stage. The followers are guided by the leaders and then move to them, which is called the local search stage. Renew the locations of the two stages according to Eqs. (6) and (7):

$$X_{1,d} = \begin{cases} F_d + c_1((ub - lb)c_2 + lb), c_3 \geqslant 0.5 \\ F_d - c_1((ub - lb)c_2 + lb), c_3 < 0.5 \end{cases} \tag{6}$$

$$X_{i,d}^{'} = \frac{X_{i,d} - X_{i-1,d}}{2} \tag{7}$$

where $X_{1,d}$ and F_d are the position of the first salp (leader) and the position of food in d-th dimension, ub and lb are the corresponding upper and lower bounds, and c_1, c_2 and c_3 are the control parameters. $X_{i,d}$ and $X_{i\text{-}1,d}$ are the positions of two salps that are tightly connected to each other in the d-th dimension.

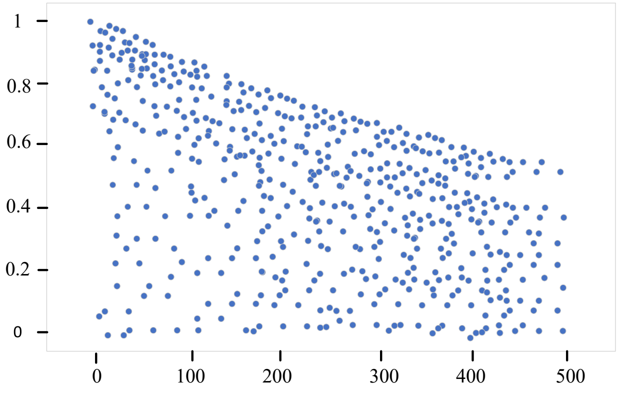


Fig. 2. Scatter diagram of detector location update coefficient.

In this algorithm, the leader constantly swims to the food (the global optimal solution), guiding the whole population to swim in a good direction. Inspired by this, this paper introduces the leader update method in the salp swarm algorithm into the detector position update formula, so that the detector position is not only affected by the previous generation detector position, but also affected by the global optimal position of the previous generation. In this case, sparrows can share information with the best individual in every iteration, so as to make the most of the information of the current global best solution and expand the search space. The improved location update formula is as follows:

(i) When $R_2 < ST$,

$$X_{i,j}^{t+1} = \begin{cases} X_{best}^{t} + c_1((ub[0,j] - lb[0,j])c_2 + lb[0,j]), c_3 \geqslant 0.5 \\ X_{best}^{t} - c_1((ub[0,j] - lb[0,j])c_2 + lb[0,j]), c_3 < 0.5 \end{cases} \tag{8}$$

(ii) When $R_2 \geq ST$,

$$X_{i,j}^{t+1} = X_{i,j}^{t} + Q \cdot L \tag{9}$$

2.2.3. *The position update of the worst participant based on OBL strategy and Lateral mutation*

In the SSA, the participants are updated according to Eq. (4), the location update of sparrows with poor energy storage will be affected by the worst individual in the whole world, the worse sparrows get less effective information in each iteration. Since each individual has the same chance to influence the others, an individual with particularly poor fitness may affect the performance of the whole swarm. In the extreme, the performance of the whole swarm might be strongly influenced by the individual with the worst fitness. Accordingly, the worst individual in the swarm should be further improved. We integrate the OBL strategy into location update formula of the worst individual. OBL was first proposed by Tizhoosh in 2005. Its main idea is to select the best solution by evaluating the current solution and the reverse solution. A number of optimization algorithms (SOA, WOA, etc.) use this method to improve their performance (Chen, Li, & Yang, 2020; Gao, Yang, Xiao, & Goh, 2022; Rao et al. 2022c). Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ be a feasible solution in D-dimensional space, and $x_{ij} \in [a_j, b_j]$, $j = (1, 2, \ldots, D)$, and its corresponding inverse solution is:

$$x_{ij} = a_{ij} + b_{ij} - x_{ij} \tag{10}$$

It selects the best solution by evaluating the current individual position and reverse position. To some extent, the introduction of OBL strategy improves the searching ability of high-quality solutions and accelerates the convergence speed, which is helpful to improve the diagnosis efficiency of CVD.

Lateral mutation is relatively simple and easy to execute, and has different search dynamics characteristics. Lateral mutation not only causes random disturbance to the last individual, but also adds the information of other excellent individuals, which enhances information exchange among sparrows. The update method is beneficial to maintain the diversity of population and improve the worst individual. Its implementation process is as follows:

$$x_{ij} = (1 - \alpha)^{*} x_{ij} + \alpha^{*} x_{kj} \tag{11}$$

Where $k \in \{1, 2, \ldots, N\}$, and $k \neq i, \alpha \in (0, 1)$ is a random number.

In order to improve the optimization ability of the algorithm, the OBL strategy and Lateral disturbance strategy are alternately executed under equal probability to dynamically update the position of the worst individual. On the one hand, in the OBL strategy, the reverse solution is obtained through the OBL mechanism to help the individual search for better solution. On the other hand, in Lateral mutation strategy, the disturbance mutation operation is carried out at the position of the worst solution, which enhances the information exchange between the worst individual and other individuals, and improves the searching ability of backward individuals. The calculation method is as follows:

$$x_{new} = \begin{cases} x_{ij}, R < 0.5 \\ x_{ij}^{'}, R \geqslant 0.5 \end{cases} \tag{12}$$

where R is a random number evenly distributed in the interval [0,1], and x_{new} is the updated position of the worst individual.

Although the above two perturbation strategies can enhance the searching ability of the worst individual, in order to ensure the effectiveness of perturbation, after the update, the greedy strategy is introduced to compare the fitness values of the new location and the old location. If the fitness of the new position is better, replace the original position, so as to ensure that the worst individual position after improvement is better than that before improvement. The corresponding formula is as follows:

$$x_{new}^{'} = \begin{cases} x_{new}, f(x_{new}) \geqslant f(x_{ij}) \\ x_{ij}, f(x_{ij}) > f(x_{new}) \end{cases} \tag{13}$$

where $f(x)$ is the fitness function.

To sum up, the method proposed in this paper makes two improvements to the traditional SSA: Aiming at the detector in the basic SSA, the salp swarm algorithm is introduced, then the optimal individual information is added to make up for the deficiency of the dimensionality reduction of the detector and further enhance the global search ability of the algorithm. Aiming at the worst participant in basic SSA, in order to reduce adverse influence of the global worst individual on whole swarm, OBL strategy and Lateral mutation strategy are introduced to strengthen the connection between the worst participant and other better individuals, and promote the sparrow population to fly and search in a better direction. In addition, in order to prevent overfitting in the optimization process, an approach based K-fold cross validation is implemented to determine the best classifier parameters.

The specific steps of the SOLSSA are as follows:

Step 1: Initialize the population and define related parameters, including sparrow population size, the maximum number of iterations, etc.

Step 2: Calculate and sort the fitness values of sparrows, find out the best and worst fitness values and their corresponding positions, and initially set the position of sparrow with the best fitness as the current food position, where the individual position of sparrow is the parameter value of the corresponding model (such as iterations, learning_rate and depth of Catboost model).

Step 3: Select some sparrows with better positions as discoverers, update their positions according to Eqns. (8) and (9). The rest are participants, and update their positions according to Eqn. (4).

Step 4: OBL and Lateral mutation strategies are carried out alternately with equal probability on the current worst sparrow.

Step 5: Use the greedy strategy to judge the worst individual position. If the new individual is better, the current individual will be replaced, otherwise, not be replaced.

Step 6: Select some sparrows randomly from the whole population as scouts and update their positions according to Eqn. (5).

Step 7: Calculate the fitness values of the updated whole sparrow population, and update the position of the globally optimal sparrow, namely, the food source position.

Step 8: Repeat Step 3 to Step 7 until it reaches a certain number of iterations, and output the global optimal solution, that is, the optimal combination of parameters of the model.

2.3. Risk assessment process of CVD

CatBoost is used as the main model in cardiovascular risk assessment. In order to achieve better prediction effect, this paper uses the proposed SOLSSA to optimize the initial iterations, depth and learning_rate of CatBoost. Many experiments use two data sets for training and testing phases: training and test, which is defined as percentage split cross-validation. This kind of approaches is not enough to reduce risk of overfitting the data in training and validating predictive models. To prevent overfitting, an approach based K-fold cross validation is implemented to determine the best classifier parameters. The data set is divided into training set and special test set. The training process of the classifier in this paper is performed on the training set with K-fold cross-validation. To achieve this, the training set is randomly divided into k subsets of equal size. Among the k subsets, one subset is designated as the validation data of the validation model, and the other k-1 subsets are used as the training data. After that, each of the k subsets is used as validation data exactly once, and records F1-score on the verification set. The cross-validation process is repeated k times, k values of F1 can be obtained in each iteration of SOLSSA, and finally the average of these k values of F1 is taken as the fitness value of SOLSSA. By using a search process with cross-validation, the process is repeated for each combination of the parameters of the classifier, and the parameters are searched by maximizing average "F1″. The solution with the highest fitness value is the best parameter combination of CatBoost.

The steps of risk assessment of CVD based on SOLSSA-CatBoost model are as follows:

Step 1: Input CVD data and reprocess the data.

Step 2: Divide the training dataset into k subset $S_1, S_2, \cdots, S_k$, such that no same samples is contained in any other subset.

Step 3: The classifier is trained using optimal parameters to present while the K-fold cross validation F1-score is computed as: Step 3.1: initialize $j = 1$. Step 3.2: subset S_j as verification set, else as train set. Step 3.3: compute the F1 score of S_j, let $j = j + 1$ and repeat Step 3.2 - Step 3.3 until $j = k$. Step 3.4: compute average value of F1-score, that is K-fold cross validation F1-score.

Step 4: Let K-fold cross validation F1-score be value of fitness. By continuous iterations of SOLSSA, the optimal parameters combination of CatBoost model is updated.

Step 5: Check whether the algorithm meets the termination conditions. If the termination condition is reached, the optimal solution is output. Otherwise, skip to Step 4.

Step 6: After this training process, the model having the highest average F1-score is used to predict CVD.

The implementation flow of the above steps is shown in Fig. 3.

3. Experimental design

In order to verify the effectiveness of the proposed SOLSSA-CatBoost, we compare it with some machine learning classifiers, mainly including RF, LR, KNN, SVM and LGB. In addition, we also compare it with the results of other swarm intelligence optimization algorithms to optimize CatBoost, including WOA-CatBoost, GWO-CatBoost, SOA-CatBoost and SSA-CatBoost. In order to verify the rationality of the proposed model, two data sets on the ML competition website Kaggle are used. The dataset #1 is a combined heart failure data set, which includes 11 features such as age, gender, chest pain type, resting blood pressure, etc. There are 918 cases in total, of which 508 cases are sick and 410 cases are not. The specific feature information is shown in Table 1. Details of dataset #2 are detailed in Section 4.4.

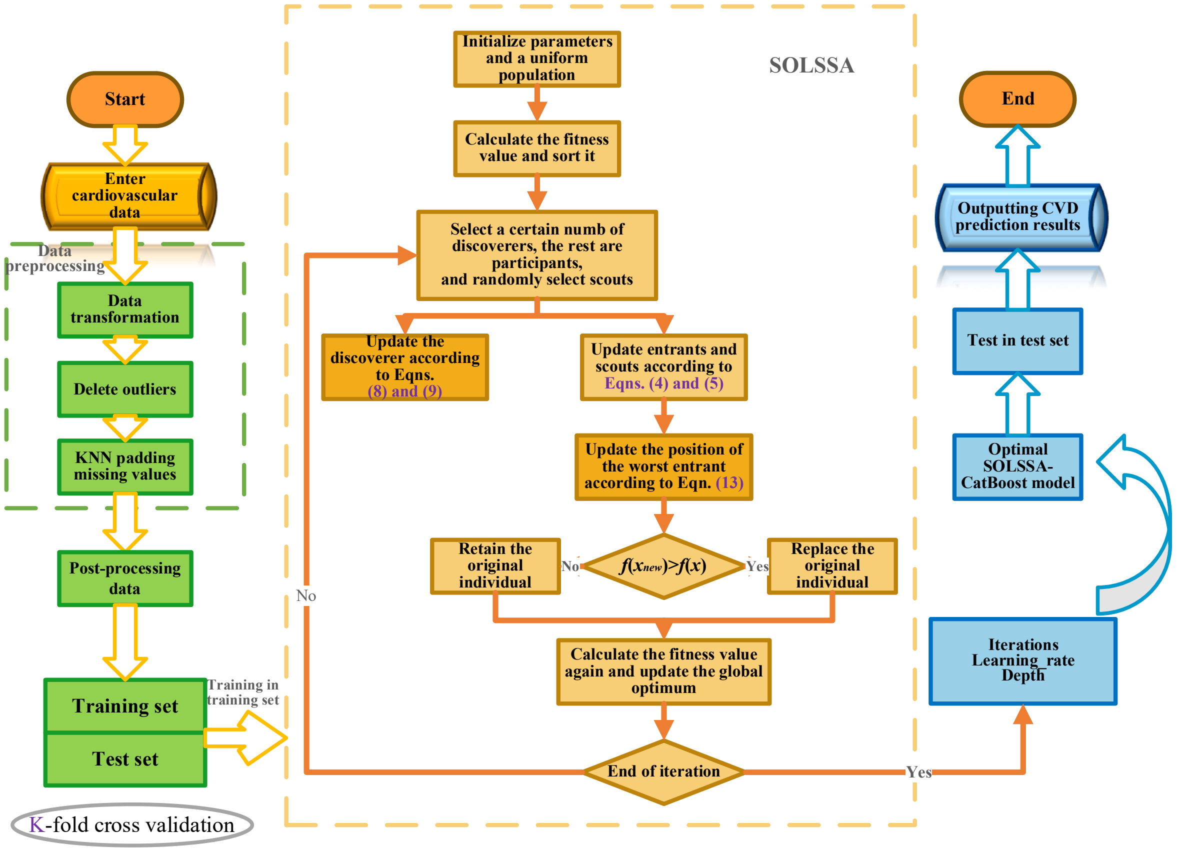

Start
Enter cardiovascular data
Data preprocessing
Data transformation
Delete outliers
KNN padding missing values
Post-processing data
Training set
Test set
Training in training set
K-fold cross validation
SOLSSA
Initialize parameters and a uniform population
Calculate the fitness value and sort it
Select a certain numb of discoverers, the rest are participants, and randomly select scouts
Update the discoverer according to Eqns. (8) and (9)
Update entrants and scouts according to Eqns. (4) and (5)
Update the position of the worst entrant according to Eqn. (13)
Retain the original individual
No
$f(x_{new})>f(x)$
Yes
Replace the original individual
Calculate the fitness value again and update the global optimum
End of iteration
No
Yes
Iterations Learning_rate Depth
Optimal SOLSSA-CatBoost model
Test in test set
Outputting CVD prediction results
End

Fig. 3. The flowchart of SOLSSA-CatBoost model.

Table 1
Feature description of data set.

Feature	Name	Feature description
X_1	Age	Age.
X_2	Sex	Gender, M stands for male, recorded as 1; F stands for female, recorded as 0.
X_3	ChestPain Type	Type of chest pain, TA: Typical, recorded as 1; ATA: Atypical, recorded as 2; NAP: Non-angina pectoris, recorded as 3; ASY: Asymptomatic, recorded as 4.
X_4	RestingBP	Blood pressure at rest (mmHg).
X_5	Cholesterol	Serum cholesterol level (mm/dl).
X_6	FastingBS	Fasting blood glucose, which is 1 when the blood glucose level is >120 mg/dl, otherwise it is displayed as 0.
X_7	RestingECG	Resting ECG results, Normal: Normal is recorded as 0; ST: ST-T wave anomaly is recorded as 1; LVH:Estes standard shows possible or definite left ventricular hypertrophy, which is recorded as 2.
X_8	MaxHR	The maximum heart rate.
X_9	ExerciseAngina	Exercise angina pectoris, Y means yes, recorded as 1; N means no recorded as 0.
X_{10}	Oldpeak	Compared with rest, ST segment depression caused by exercise.
X_{11}	ST_Slope	Slope of ST segment at the peak of movement; Up: Upward slope, recorded as 1; Flat: Flat, recorded as 2; Down: Downward slope, recorded as 3.
y	HeartDisease	1 is sick, 0 is normal.

3.1. Data set analysis

3.1.1. Data preprocessing

The pretreatment process in this paper mainly includes data conversion, outliers processing, missing values processing. In view of the fact that language variables can't be quantitatively analyzed in the process of data analysis, combined with the common coding methods in UCI related data sets, it will be transformed into a form that can participate in the classification model, as shown in Table 1. Referring to the basic data set standard of Chinese adult health examination (HRC00.04), the data set is cleaned and the abnormal values are eliminated. For those variables with more outliers, if they are directly removed, the model may be affected, so this paper treats them as missing values. The KNN filling method is more extensive, simple and extensible, so this paper uses KNN filling processing to compensate incomplete data and make up for the impact of missing features on data distribution (Ma, Tian, Liu, & Zhang, 2020).

3.1.2. Correlation analysis

In this work, SOLSSA-CatBoost is applied to the heart failure data set consisting of 11 features. At first, we analyze the data set to understand the relationship among theses 11 features. Moreover, if there is multicollinearity among variables, the model may produce unreasonable results. Therefore, we need to identify and remove variables with strong correlation. Pearson correlation coefficient can describe the close relationship between two fixed distance variables well, this paper uses it calculate the correlation between features in a heat map, as shown in Fig. 4.

Correlation matrix represents features and analyzes relationships. The correlation matrix value is between −1 and 1, if the matrix value is close to 1 or −1, it means that features influence each other, and any one of them can be independently selected for the proposed algorithm. It can be seen from Fig. 4 that all the correlation coefficient values are below 0.6, that is, there are no highly correlated features, so all these features will be input into the proposed method without feature selection.

3.1.3. Descriptive analysis

We perform a simple statistical analysis of the experimental data, first of all, analyze the common relationship between gender and age.

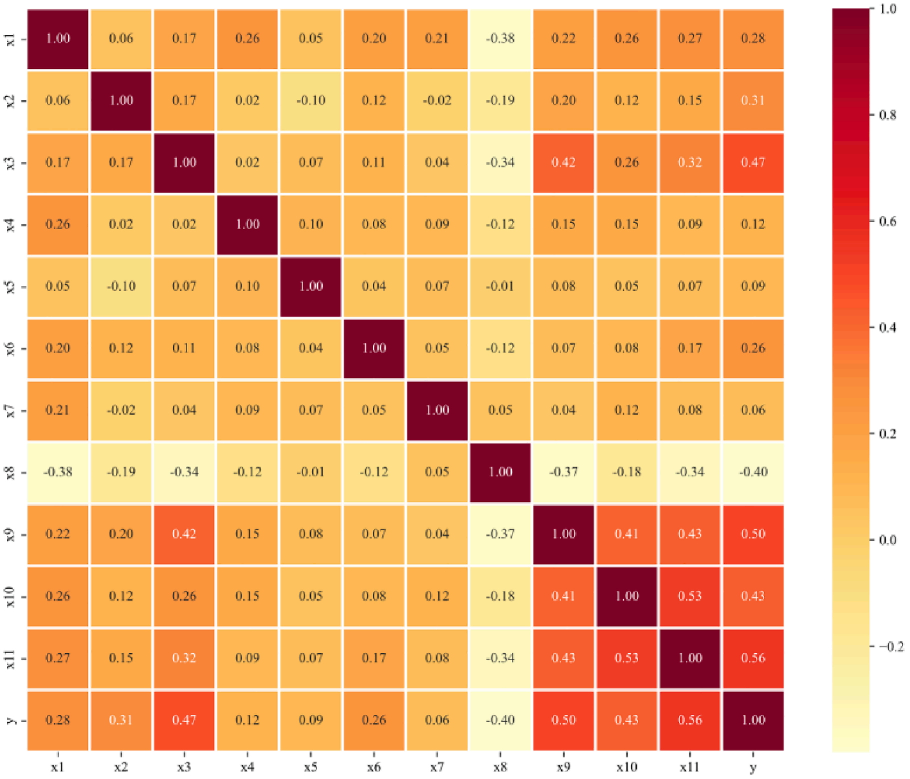


Fig. 4. Correlation matrix of the data set characteristics.

From the box chart in Fig. 5, we can see that most female patients are aged around 50–60, and most male patients are aged around 45–65, that is, most of the sick people are concentrated in middle-aged and elderly people.

Then, we briefly analyze the relationship between Oldpeak and ST_Slope in Fig. 6 with 0.53 relative value.

According to the data we have, they show such a relationship: With the increase of ST_Slope value, the Oldpeak value also gradually increases, and we can find that the Oldpeak of patients is higher than that of healthy people.

3.2. Evaluation index of model performance

In the field of ML, the evaluation indicators based on confusion matrix metrics are generally used as the evaluation indicators of classification algorithm, such as accuracy, precision, recall, F1-score and area under the Receiver Operating Characteristic (ROC) curve (AUC). This paper focuses on the accuracy, F1-score and AUC value of model prediction.

3.2.1. Classification evaluation indicators based on confusion matrix

Confusion Matrix is often used as an evaluation metric for prediction model accuracy. The confusion matrix of the binary classification results is shown in Table 2, where *TP* (True Positive) represents the number of people who are correctly judged to be sick. *FN* (False Negative) indicates the number of people who are wrongly judged as normal. *TN* (True Negative) indicates the number of people who are correctly judged as normal. *FP* (False Positive) indicates the number of people who are wrongly judged to be sick. And *TP* and *TN* measure the ability of the model to predict whether CVD exists or not.

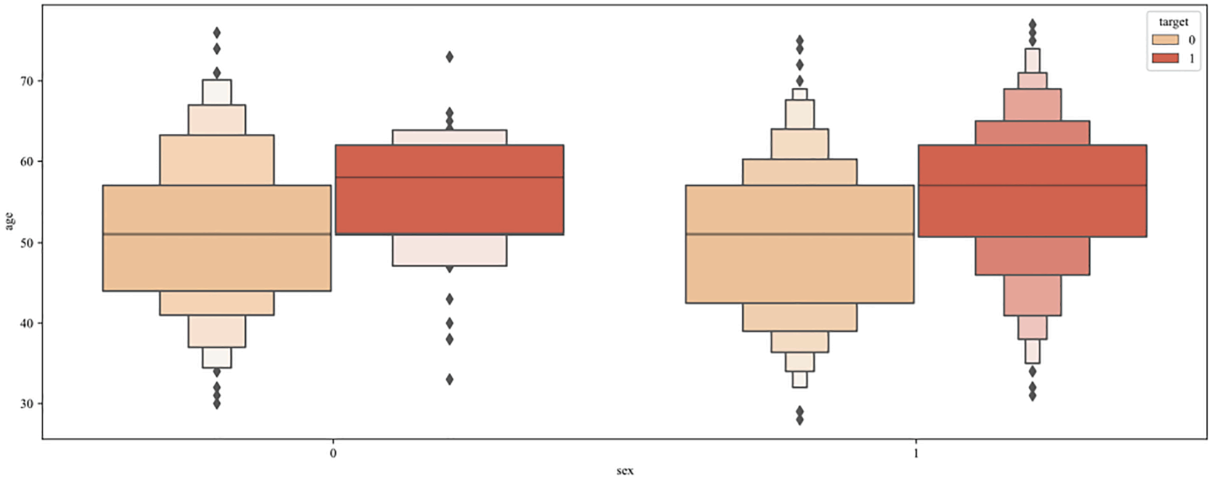


Fig. 5. Box diagram between Sex and Age.

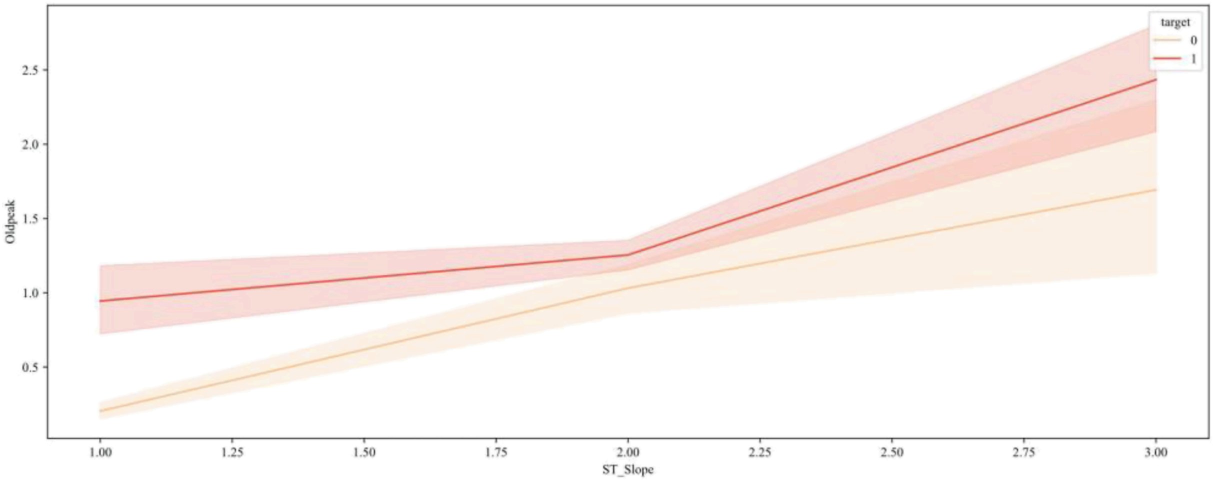


Fig. 6. Lineplot between ST_Slope and Oldpeak.

Table 2
Confusion matrix.

	Predicted Positive	Predicted Negative
Real Positive	*TP*	*FN*
Real Negative	*FP*	*TN*

(1) Accuracy

Based on the quality of the solution, the proposed method is tested and analyzed on the basis of accurate calculation. Accuracy represents the ratio between the number of correctly predicted samples (including those with and without CVD) and the total number of samples in the data set. Accuracy (%) is calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

(2) Precision

The precision rate refers to the proportion of the number of people who are correctly judged to be sick to the number of people who are predicted to be sick. The calculation precision is based on the following equation.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

(3) Recall

The recall rate refers to the proportion of patients with CVD that are correctly identified in all patients with CVD. Classifiers with higher recall rate will pay as much attention to patients with CVD as possible, so as to avoid mistaking patients with CVD as normal people without CVD.

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

(4) (4) F1-score

F1-score (%) is the average of the precision rate and recall rate, and its value is [0,1]. The calculation formula is shown as follows.

$$F1\text{ - }score = 2\cdot\frac{Precision\cdot Recall}{Precision + Recall} \tag{17}$$

3.2.2. ROC curve and AUC value

ROC curve is a (1-TNR, TPR) trajectory of coordinate points formed by confusing the misjudgment rate and recall rate on the matrix. Among them, the vertical axis TPR indicates the proportion of all patients with CVD that are correctly judged to be ill in all patients with CVD: TPR = $TP/(TP+FN)$. On the horizontal axis, FPR indicates the proportion of people who are wrongly judged to be ill in all normal people without CVD: FPR = $FP/(FP+TN)$. Generally, scholars use AUC value as an indicator to measure the performance of the model, which is defined as the area enclosed by the coordinate axis under the ROC curve. The value of AUC is [0.5,1], and the closer it is to 1, the higher the authenticity of the detection method is. When it is equal to 0.5, the authenticity is the lowest and it has no application value.

3.3. Experimental setting

In the model evaluation scheme, the performance on the test set can better measure the quality and generalization ability of the trained model than performance on the training set. Many experiments use two data sets for training and testing phases: training and test, which is defined as percentage split cross-validation. This kind of approaches is not enough to reduce risk of overfitting the data in training and validating predictive models. In order to train, test, and tune the specific parameters of the classifiers and prevent overfitting in the optimization process, this study adopts 5-fold cross validation and chooses 5-fold cross validation F1-score as the fitness function in the optimization algorithms. The data set is divided into training set and special test set according to 4: 1. Divide the training dataset into 5 subsets $S_1, S_2, \cdots, S_5$, first, S_1 is designated as verification data, and the remaining four subsets are used as training data. Based on the parameters searched by SOLSSA, CatBoost trains the model in four subsets, records F1-score on the verification set S_1, and repeats the cross-validation process 5 times, that is, $S_1, S_2, \cdots, S_5$ are used as verification sets respectively. At this time, 5 values of F1 can be obtained in each iteration of SOLSSA, and finally the average of these 5 values of F1 is taken as the fitness value of SOLSSA. By continuous iterations of SOLSSA, the optimal solution is updated (The optimal solution is the parameters of the CatBoost model with the highest average F1-score). After this training process, train the model to get the final model and the model with the highest average F1-score is evaluated on the special test set.

The parameters and search space setting of SOLSSA-CatBoost are listed in Table 3. The regularization sub-parameter is the coefficient of L2 regularization term of the cost function, and it is set to 3. Other parameters are set as default parameters. Since the hyper-parameters of SOLSSA are fixed, its search space is not applicable.

Table 3
Parameters and search space setting.

Type	Parameters	Initial value	Search space
SOLSSA	maximum number of iterations	100	N.A.
	population size	50	N.A.
	proportion of discoverers	0.2	N.A.
	variable number	3	N.A.
	early warning value	0.6	N.A.
CatBoost	iterations	100	100–2000
	depth	4	4–10
	learning_rate	0.01	0.01–0.5

4. Experimental results and analysis

In this section, two data sets are used, and three evaluation indicators are applied to test the performance of the proposed method and algorithm. In order to evaluate the performance of the proposed model, two types of models are used to compared with the proposed model. On the one hand, we compare it with some machine learning classifiers. On the other hand, we also compare it with the results of other swarm intelligence optimization algorithms to optimize CatBoost.

4.1. Compared with machine learning models

Before analyzing the proposed method, six widely used classifiers, i.e., RF, LR, KNN, SVM, LGB and CatBoost, were tested on the test set of data set #1. Furthermore, the SOLSSA-CatBoost was tested under the same experimental settings. Prediction results and corresponding prediction indicators are indicated in Table 4. To compare the performance of the methods more intuitively, all the results are presented in the form of histograms in Fig. 7.

As can be seen from Table 4 and Fig. 7, for accuracy index, SOLSSA-CatBoost achieves the best results on all models, which is 1.54 %–21.46 % higher than others. The accuracy rates of RF, LR, KNN, SVM, LGB and CatBoost are 86.18 %, 84.30 %, 66.38 %, 70.60 %, 84.52 % and 86.30 % respectively. Among them, KNN has the worst performance in accuracy, which is 66.38 %. CatBoost has the highest accuracy rate of 86.30 % among these basic classification models, and can obtain accuracy value similar to SOLSSA-CatBoost. SOLSSA-CatBoost achieves 87.84 % accuracy on the data set #1, which means that 158 out of 180 people will be correctly classified. Experiments show that the proposed SOLSSA can optimize the CatBoost model with parameters, and the accuracy of model is enhanced.

F1-score is the average of precision and recall. The above Table 4 and Fig. 7 show that the F1-score obtained by SOLSSA-CatBoost is the highest. The F1-score of RF, LR, KNN, SVM, LGB and CatBoost are 87.69 %, 85.88 %, 70.30 %, 74.74 %, 86.18 % and 87.80 % respectively. Among them, KNN has the worst performance in F1-score, CatBoost has the highest F1-score, reaching 87.80 %. This indicates that CatBoost has certain advantages in predicting CVD compared with other classifiers. Furthermore, the F1-score of CatBoost reaches 90.00 % after being optimized by SOLSSA proposed in this paper, which is 2.20 %–19.70 % higher than others. It indicates the superiority of SOLSSA-CatBoost in CVD classification.

Table 4
Compared with classification models.

Model	Accuracy (%)	F1-score (%)	AUC value (%)
RF	86.18	87.69	85.75
LR	84.30	85.88	84.04
KNN	66.38	70.30	65.79
SVM	70.60	74.74	69.74
LGB	84.52	86.18	84.15
CatBoost	86.30	87.80	85.87
SOLSSA-CatBoost	87.84	90.00	87.41

The AUC of RF, LR, KNN, SVM, LGB and CatBoost are 85.75 %, 84.04 %, 65.79 %, 69.74 %, 84.15 % and 85.87 % respectively. It shows that CatBoost model has good classification performance among the commonly used standard classification models, which proves the rationality of choosing CatBoost as the basic model in the classification prediction of CVD in this paper. As can also be seen from Table 4 and Fig. 7, SOLSSA-CatBoost has the highest AUC value of 87.41 %, which is 1.54 %–21.62 % higher than others. It can also be seen from Fig. 8 that the area under the ROC curve of the SOLSSA-CatBoost model is the largest, that is, the AUC value is the largest. To sum up, the SOLSSA algorithm proposed in this paper is practical and meaningful, and can effectively improve the prediction effect of CatBoost.

4.2. Comparison with state-of-the-art optimization algorithms

In addition to the SSA, some classic swarm intelligence optimization algorithms are often used to optimize models, such as WOA, GWO. On this basis, the improved sparrow search algorithm proposed in this paper is compared with other four optimization algorithms (WOA, GWO, SOA and original SSA) in optimizing the performance of the CatBoost to further verify the effectiveness of the SOLSSA. They are all carried out under the same experimental conditions.

4.2.1. Best fitness evaluation

The best fitness results of the optimization algorithms are compared in Table 5. It can be seen that the SOLSSA approach outperforms other optimization algorithms and the fitness value of the second-best optimization algorithm (SSA) is 89.06 %. In addition, according to the search result of the SOLSSA, its optimization effect on CatBoost is the most prominent, when iterations = 101, depth = 6, learning_rate = 0.43, the fitness value reached 90.00 %. Fig. 9 systematically shows the optimization process of these two optimization algorithms with strong optimization ability. As can be seen from Fig. 9, the results of SOLSSA-CatBoost and SSA-CatBoost in training set and verification set are good, the models fit well, and SOLSSA-CatBoost can search for more parameter values and get better results. In addition, when the result on the training set is getting better and better, it means that the model has learned the training set well and detected random noise and changes in the training set. A significant gap between training and validation score is a sign of overfitting, and the degree of this gap is evidence of the magnitude of overfitting. SOLSSA-CatBoost and SSA-CatBoost mainly choose the turning point when the score of the validation set begins to decline, but the result of the training set is getting higher and higher as the optimal choice of the parameter, which effectively prevents overfitting. In other words, the result of the searched parameters is relatively reasonable.

Convergence analysis of the proposed SOLSSA relative to the compared algorithms was performed based on the CatBoost classifier. Fig. 10 presents their convergence curves, which can be clearly seen that SOA algorithm has little improvement on the prediction accuracy of CatBoost model, and the fitness function value varies from 88.48 % to 88.59 %; WOA and GWO keep searching for optimization in this process, but finally they only stay at the fitness values of 88.8 % and 88.87 %. SSA keeps the fitness value of 88.9 % after 12 iterations. Then, after the 47th iteration, it falls into the local optimum, keeping the fitness value of 89.06 % unchanged. However, SOLSSA proposed in this paper has an F1 value of 88.9 % after 12 iterations, and then jumps out of the local optimum twice, and converges after the optimum value reaches 90.00 %, which is significantly better than other algorithms. From this result, we can conclude that the SSA shows good performance, which is mainly due to its unique sparrow population classification mechanism. Furthermore, the improved sparrow search algorithm proposed in this paper further enhances its global search and the ability to jump out of the local optimum on this basis, which makes the proposed SOLSSA algorithm a promising optimization algorithm.

The best prediction model is determined by the searched final parameters, and the model having the highest average F1-score is tested on

SOLSSA-CatBoost	87.84	90.00	87.41

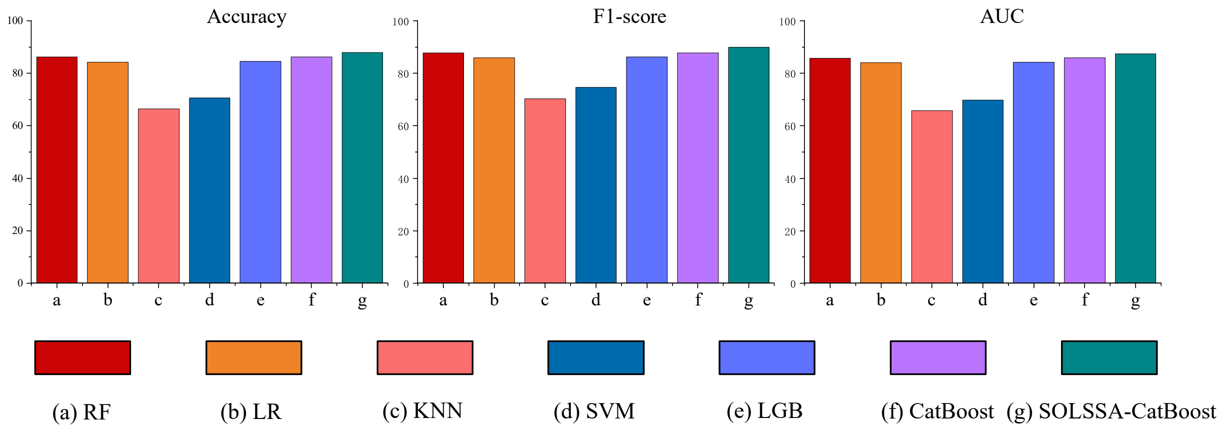


Fig. 7. The histograms of performance comparison between SOLSSA-CatBoost and ML models.

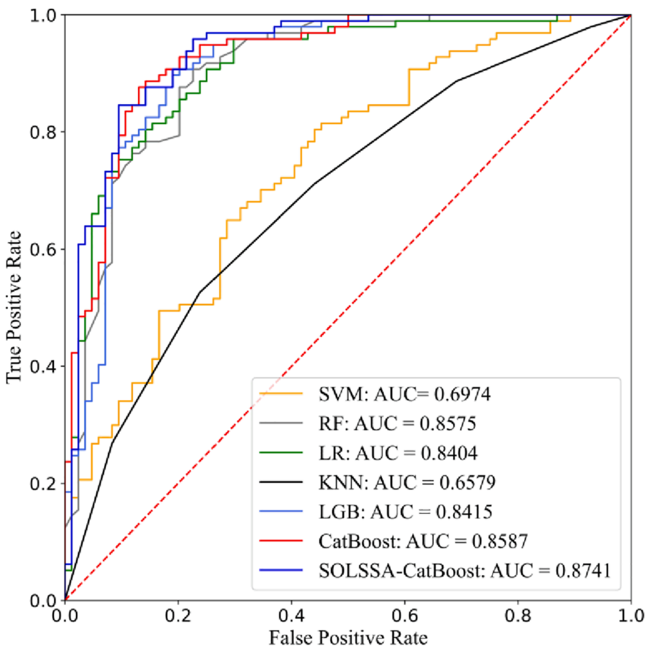


Fig. 8. ROC curves of seven models.

Table 5
Optimal solution and fitness values.

Model	Iterations	Depth	Learning_rate	Fitness
SOLSSA-CatBoost	101	6	0.43	90.00
SSA-Catboost	100	5	0.48	89.06
GWO-CatBoost	179	7	0.26	88.87
WOA-CatBoost	100	5	0.15	88.80
SOA-CatBoost	106	4	0.07	88.59

a special test set. The results were shown in the following Table 6. It can be seen that the F1-score of SOLSSA-CatBoost is still the highest in the test set, which is 1 %–2.05 % higher than other models. It fully shows that SOLSSA-CatBoost is effective and has strong generalization ability. In this sense, using SOLSSA-CatBoost model to predict CVD can effectively reduce the misdiagnosis rate, thus halting, reversing and reducing the spread of CVD and health hazards.

4.2.2. Accuracy and AUC evaluation

Table 7 compares the accuracy and AUC of SOLSSA and the other optimization algorithms based on the CatBoost classifier in the verification set. To compare the performance of the methods more intuitively, accuracy and AUC values in the process of optimization are presented in Fig. 11. In the iterative process, the SOLSSA-CatBoost approach achieves the highest accuracy (87.84 %), followed by SSA-CatBoost, which has the accuracy of 87.62 %. The SOLSSA-CatBoost approach achieves the highest AUC (87.41 %), followed by SSA-CatBoost, which has the AUC of 87.16 %. Among them, SOA-CatBoost has the worst search effect, with the final accuracy of 87.18 % and AUC of 86.78 %. From this result, we can conclude that the SSA shows good performance, the improved sparrow search algorithm proposed in this paper further enhances its global search and achieves better results. As a risk prediction model of CVD, it has a huge number of patients in practice, and the improvement of model performance means that it will be better able to judge whether patients are suffering from diseases, thus reducing the risk of CVD in advance.

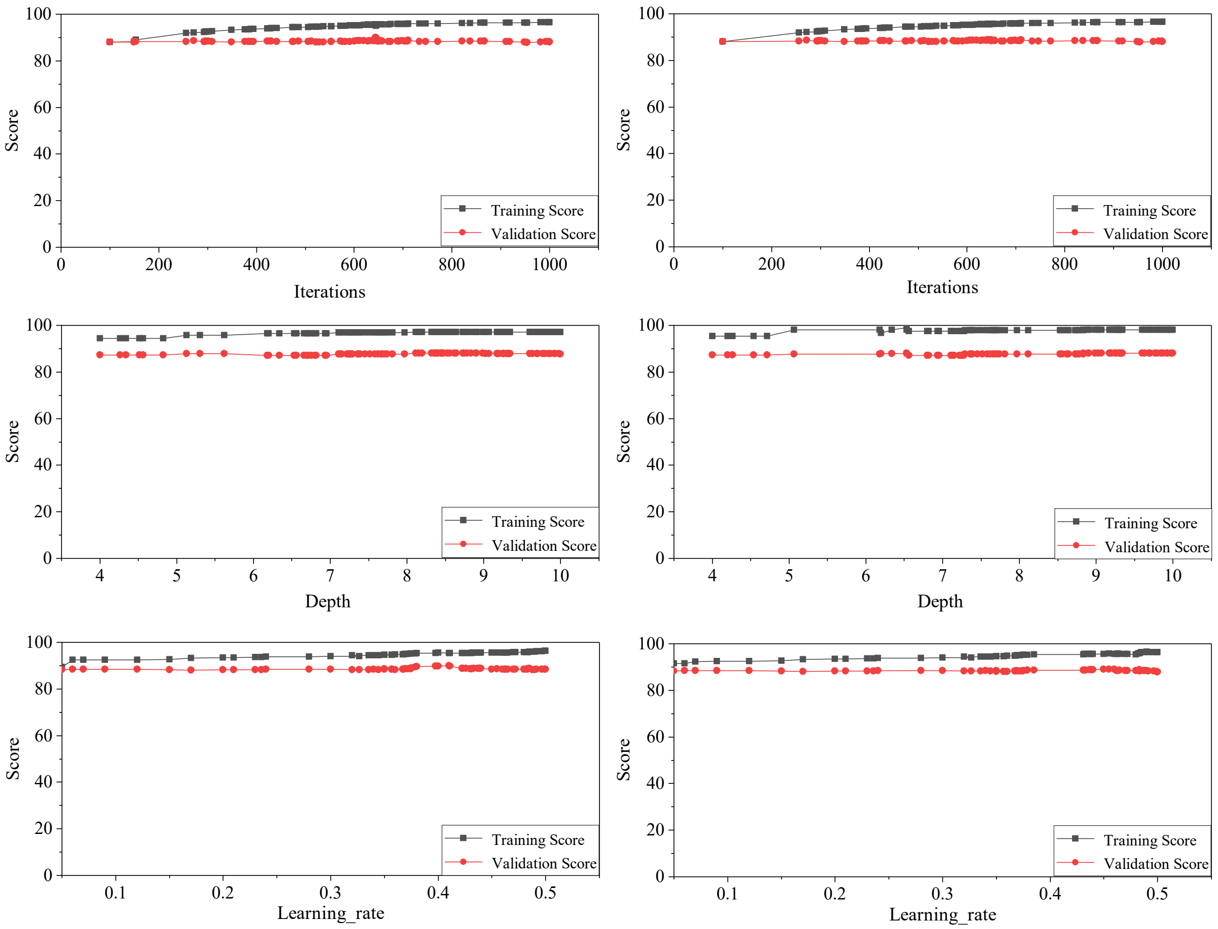
4.3. Effectiveness analysis of two improvement strategies

In order to verify the effectiveness and necessity of the two improvements made in this paper, we have carried out two separate experiments: 1) Only introducing salp swarm algorithm; 2) Only combining OBL and Lateral mutation strategies. We compare its performance with the work done in this paper, and the results are shown in Table 8.

As can be seen from Table 8, these two improvements can get better results than SSA-CatBoost. In the experiment 1), the salp swarm algorithm is integrated, then the global optimal solution of the previous generation is added. Its accuracy, F1-score, AUC are 87.70 %, 89.28 % and 87.00 % respectively, which are better than SSA's in general. It improves the global search ability of SSA from the perspective of the discoverers. In the experiment 2), perturbation mutation is carried out at the worst position to strengthen the information exchange between the worst individual and other better individuals, and improve the searching ability of the worst individua. Its accuracy, F1-score, AUC are 88.06 %, 89.51 % and 87.25 % respectively, which are all better than SSA's. It is also found that introducing these two improvements into SSA at the same time can achieve better prediction results.

4.4. Further verification on other data sets

In order to further reflect the performance of the proposed model, the



(a) SOLSSA-CatBoost

(b) SSA-CatBoost

Fig. 9. Validation curve for the SOLSSA-CatBoost and SSA-CatBoost.

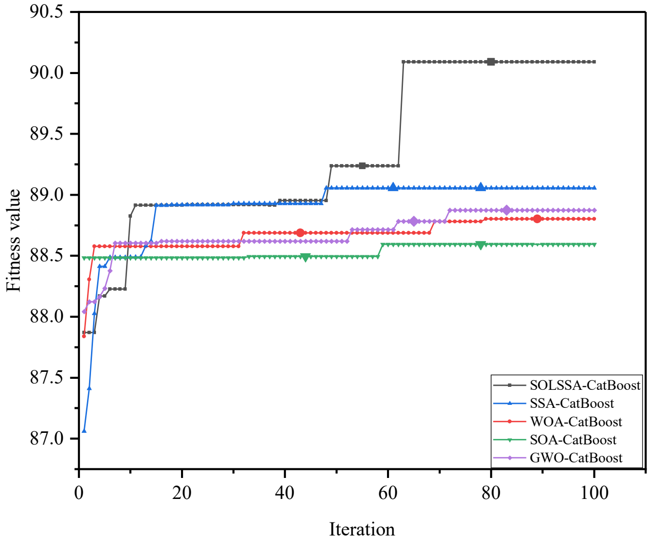


Fig. 10. Convergence progress of five models.

Table 6
F1-score of different optimization algorithms on test set.

Model	F1-score (%)
SOLSSA-CatBoost	90.00
SSA-Catboost	89.00
GWO-CatBoost	88.80
WOA-CatBoost	88.72
SOA-CatBoost	87.95

Table 7
Accuracy and AUC comparison of SOLSSA and other optimization algorithms based on CatBoost classifier in the verification set.

Model	Accuracy (%)	AUC (%)
SOLSSA-CatBoost	87.84	87.41
SSA-CatBoost	87.62	87.16
GWO-CatBoost	87.40	86.92
WOA-CatBoost	87.40	86.98
SOA-CatBoost	87.18	86.78

SOLSSA-CatBoost model is applied to other relevant data sets. In this section, another heart failure data set in Kaggle data platform is selected. We do the same pretreatment for this data set, and evaluate the performance of the proposed model on the test set. We compare SOLSSA-CatBoost with some machine learning models, also compare SOLSSA-CatBoost with other four optimization algorithms (WOA, GWO, SOA and original SSA) in optimizing the performance of the CatBoost.

Table 9 shows the evaluation indicators of the prediction performance of different machine learning models. Based on AUC indicator, although CatBoost is not as good as LGB, the overall performance still shows a certain degree of competitiveness. It is not difficult to find that the prediction effect of the model proposed in this paper is still the best. Concretely, (i) the accuracy of SOLSSA-CatBoost is 85.00 %, which is

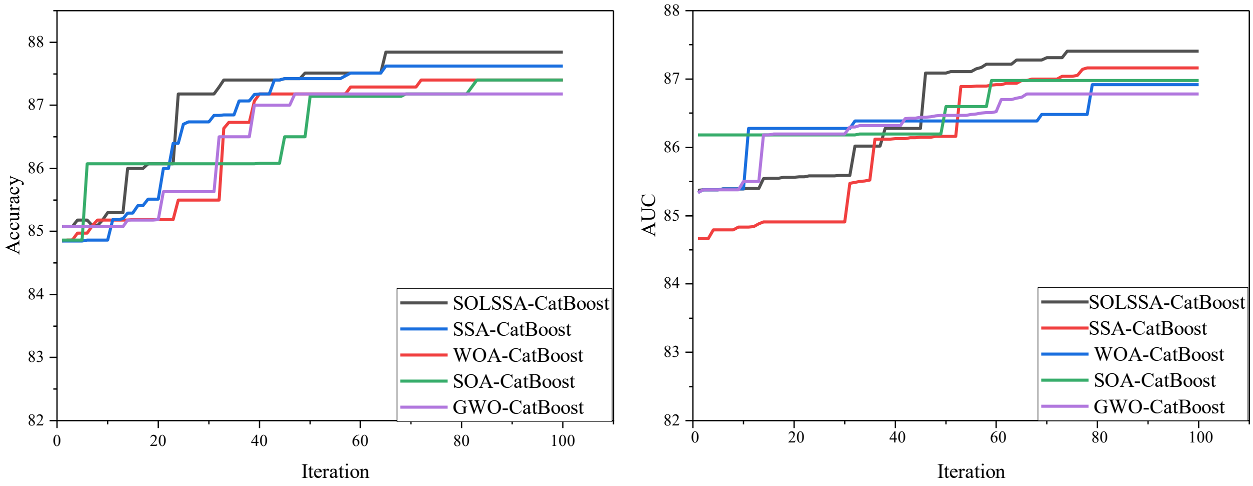


Fig. 11. Iterative accuracy and AUC values of five models.

Table 8
Experimental results presented by different improvements.

Model	Accuracy (%)	F1-score (%)	AUC (%)
Experiment 1)	87.70	89.28	87.00
Experiment 2)	88.06	89.51	87.25
SSA-CatBoost	87.62	89.06	87.16
SOLSSA-CatBoost	87.84	90.00	87.41

Table 9
Prediction results of different models on test set.

Model	Accuracy (%)	F1-score (%)	AUC (%)
RF	82.00	72.34	79.70
LR	79.00	65.21	74.02
KNN	63.33	63.57	50.70
SVM	68.00	57.00	50.00
LGB	83.67	75.17	81.64
CatBoost	83.68	77.25	81.57
SOLSSA-CatBoost	85.00	81.51	84.21

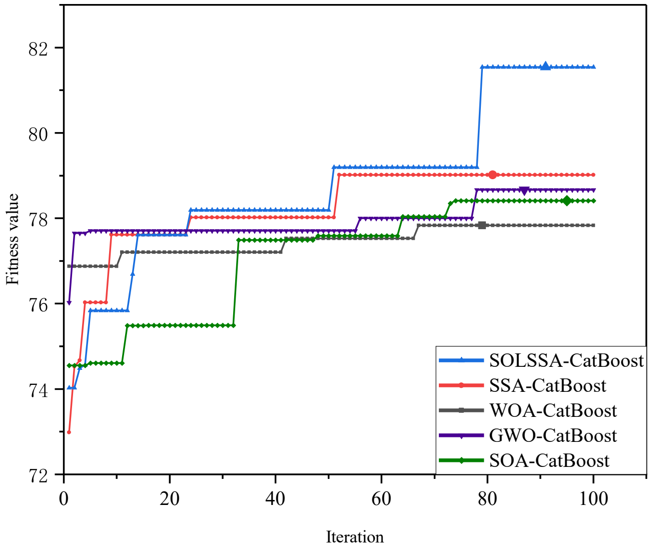


Fig. 12. The dynamic convergence curves of different models.

1.32 % higher than CatBoost; (ii) the F1-score of SOLSSA-CatBoost is 81.51 %, which is 4.26 % higher than CatBoost; (iii) the AUC of SOLSSA-CatBoost is 84.21 %, which is 2.64 % higher than CatBoost.

Fig. 12 shows dynamic convergence curves of different models on verification set. It can be clearly seen that compared with other swarm intelligence optimization algorithms, the sparrow search algorithm shows better performance, which is mainly due to its unique sparrow population classification mechanism. The fitness value of SSA-CatBoost model remained constant after the 54th iteration, while SOLSSA-CatBoost jumped out of the local optimum in the later period, and finally got a higher fitness value, which was 2.53 % higher than SSA-CatBoost. Therefore, the improved sparrow search algorithm proposed in this paper further enhances its global search and the ability to jump out of the local optimum on this basis, which verifies the necessity and effectiveness of the proposed improvements. Table 10 shows the final accuracy and AUC evaluation indexes results of each model on verification set. It can be found that SOLSSA-CatBoost proposed in this paper has the best prediction performance, with accuracy, F1-score and AUC reaching 85.01 %, 81.55 % and 84.30 % respectively.

5. Conclusion

In this paper, an improved sparrow search algorithm is proposed to optimize parameters of CatBoost, named SOLSSA-CatBoost. We mainly improve SSA from two aspects: First of all, in the position update formula of the discoverers, we fuse the salp swarm algorithm and SSA to enhance the information sharing between the discoverers and the globally optimal individual and help maintain the diversity of the population; Secondly, the OBL strategy and the Lateral disturbance strategy are introduced to disturb the position of the worst sparrow individual and help improve the search ability of the worst individual. Finally, we use the improved algorithm to optimize the parameters of CatBoost, and verify it with two cardiovascular data sets. We made a comparative analysis from two aspects. On the one hand, we compared SOLSSA-CatBoost model proposed in this paper with other commonly used

Table 10
The results of performance comparison SOLSSA and other optimization algorithms based on CatBoost classifier.

Model	Accuracy (%)	F1-score (%)	AUC (%)
SOLSSA-CatBoost	85.00	81.51	84.21
SSA-CatBoost	84.33	79.02	83.78
WOA-CatBoost	84.67	77.83	83.69
GWO-CatBoost	84.00	78.67	83.29
SOA-CatBoost	85.00	78.41	83.31

machine learning models, including RF, LR, KNN, SVM, LGB and CatBoost. On the other hand, we compared SOLSSA with other four optimization algorithms (WOA, GWO, SOA and original SSA) in optimizing the performance of the CatBoost. The training and validation sets were used to tune the CatBoost model parameters during the training epochs by varying the predictive models with SOLSSA and K-fold cross validation. After obtaining optimized model, a dedicated test set was used to estimate the F1-score for the model. The final results show that SOLSSA-CatBoost model has a good prediction effect in cardiovascular disease risk prediction.

We believe that the research done in this paper is meaningful in CVD risk prediction. For patients, it can better guide individuals, make high-risk groups change their life behaviors, thus reducing the risk of illness. For doctors, it can help doctors better identify potential risks and trends, so as to screen out high-risk groups and reduce the misdiagnosis rate.

However, this study still has some limitations and future research findings. First of all, there are a variety of structural forms of health care big data that can be used for CVD risk assessment, such as structured, semi-structured and unstructured. However, the method proposed in this paper has only been verified on data sets that may contain only one structure. Secondly, this paper selects CatBoost as the main model, and verifies that SOLSSA is effective in optimizing the parameters of CatBoost, but it is unknown whether SOLSSA is effective for other classifiers. In future work, we will consider the evaluation data set with heterogeneous characteristics and consider how to design a CVD risk assessment mechanism based on various structural data. In addition, the SOLSSA can be applied with other classifiers, such as SVM, RF, and LR. These studies can evaluate the generality of SOLSSA and its promotion ability in different classifiers.

CRediT authorship contribution statement

Xi Wei: Methodology, Software, Writing – original draft. **Congjun Rao:** Conceptualization, Methodology, Data curation. **Xinping Xiao:** Supervision, Writing – review & editing. **Lin Chen:** Visualization, Investigation, Validation. **Mark Goh:** Formal analysis, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 72071150, 71871174).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2023.119648.

References

Al-Absi, H. R. H., Refaee, M. A., Rehman, A. U., Islam, M. T., Belhaouari, S. B., & Alam, T. (2021). Risk factors and comorbidities associated to cardiovascular disease in Qatar: A machine learning based case-control study. *IEEE Access, 9*, 29929–29941.

Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion, 63*, 208–222.

Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A framework for feature selection through Boosting. *Expert Systems with Applications, 187*, Article 115895.

Ambale-Venkatesh, B., Yang, X., Wu, C. O., Liu, K., Hundley, W. G., Mcclelland, R., Gomes, A. S., Folsom, A. R., Shea, S., & Guallar, E. (2017). Cardiovascular event prediction by machine learning the multi-ethnic study of atherosclerosis. *Circulation Research, 121*(9), 1092–1101.

Bentejac, C., Csorgo, A., & Martinez-Munoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937–1967.

Cao, L., & Tay, F. E. (2001). Financial forecasting using support vector machines. *Neural Computing & Applications, 10*(2), 184–192.

Chen, H., Li, W., & Yang, X. (2020). A whale optimization algorithm with chaos mechanism based on quasi-opposition for global optimization problems. *Expert Systems with Applications, 158*, Article 113612.

Cikes, M., Sanchez-Martinez, S., Claggett, B., Duchateau, N., Piella, G., Butakoff, C., … Bijnens, B. (2019). Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *European Journal of Heart Failure, 21*(1), 74–85.

Dhiman, G., & Kumar, V. (2019). Seagull optimization algorithm: Theory and its applications for large-scale industrial engineering problems. *Knowledge-Based Systems, 165*, 169–196.

Elattar, E. E., & ElSayed, S. K. (2020). Probabilistic energy management with emission of renewable micro-grids including storage devices based on efficient salp swarm algorithm. *Renewable Energy, 153*, 23–35.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*(2), 368–378.

Fu, H., & Liu, H. (2022). Improved sparrow search algorithm with multi-strategy integration and its application. *Control and Decision, 37*(1), 87–96.

Gao, M. Y., Yang, H. L., Xiao, Q. Z., & Goh, M. (2022). COVID-19 lockdowns and air quality: Evidence from grey spatiotemporal forecasts. in press *Socio-Economic Planning Sciences*. https://doi.org/10.1016/j.seps.2022.101228.

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., … Wilson, P. W. F. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association task force on practice guidelines. *Journal of the American College of Cardiology, 63*(25 Pt B), 2935–2959.

Gupta, S., & Deep, K. (2020). Enhanced leadership-inspired grey wolf optimizer for global optimization problems. *Engineering with Computers, 36*(4), 1777–1800.

He, D., Liu, C., Jin, Z., Ma, R., Chen, Y., & Shan, S. (2022). Fault diagnosis of flywheel bearing based on parameter optimization variational mode decomposition energy entropy and deep learning. *Energy, 239*, Article 122108.

Hossain, E., Uddin, S., & Khan, A. (2021). Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Systems with Applications, 164*, Article 113918.

Jiang, H. L., Mao, H. F., Lu, H. M., Lin, P. Y., Garry, W., Lu, H. J., Yang, G. Q., Rainer, T. H., & Chen, X. H. (2021). Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *International Journal of Medical Informatics, 145*, Article 104326.

Jiang, Y., Li, X., & Huang, C. (2013). Automatic calibration a hydrological model using a master-slave swarms shuffling evolution algorithm based on self-adaptive particle swarm optimization. *Expert Systems with Applications, 40*(2), 752–757.

Lee, H. J., Kim, H., Kim, Y. T., Won, K., Czosnyka, M., & Kim, D. J. (2021). Prediction of life-threatening intracranial hypertension during the acute phase of traumatic brain injury using machine learning. *IEEE Journal of Biomedical and Health Informatics, 25* (10), 3967–3976.

Li, L. L., Xiong, J. L., Tseng, M. L., Yan, Z., & Lim, M. K. (2022). Using multi-objective sparrow search algorithm to establish active distribution network dynamic reconfiguration integrated optimization. *Expert Systems with Applications, 193*, Article 116445.

Li, L., Zhang, Z., Xiong, Y., Hu, Z., Liu, S., Tu, B., & Yao, Y. (2022). Prediction of hospital mortality in mechanically ventilated patients with congestive heart failure using machine learning approaches. *International Journal of Cardiology, 358*, 59–64.

Liu, G., Shu, C., Liang, Z., Peng, B., & Cheng, L. (2021). A modified sparrow search algorithm with application in 3d route planning for UAV. *Sensors, 21*(4), 1224.

Liu, J., Zhang, X., Wang, X., Xu, L., Li, J., & Fang, X. (2016). Single-molecule force spectroscopy study of the effect of cigarette carcinogens on thrombomodulin-thrombin interaction. *Science Bulletin, 61*(15), 1187–1194.

Liu, X., Guo, J., Wang, H., & Zhang, F. (2022). Prediction of stock market index based on ISSA-BP neural network. *Expert Systems with Applications, 204*, Article 117604.

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine, 121*, Article 103761.

Ma, Z. F., Tian, H. P., Liu, Z. C., & Zhang, Z. W. (2020). A new incomplete pattern belief classification method with multiple estimations based on KNN. *Applied Soft Computing, 90*, Article 106175.

Mirjalili, S. (2015). Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. *Knowledge-Based Systems, 89*, 228–249.

Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software, 95*, 51–67.

Paynter, N. P., LaMonte, M. J., Manson, J. E., Martin, L. W., Phillips, L. S., Ridker, P. M., Robinson, J. G., & Cook, N. R. (2014). Comparison of lifestyle-based and traditional cardiovascular disease prediction in a multiethnic cohort of nonsmoking women. *Circulation, 130*(17), 1466–1473.

Qian, X., Li, Y., Zhang, X., Guo, H., He, J., Wang, X., Yan, Y., Ma, J., Ma, R., & Guo, S. (2022). A cardiovascular disease prediction model based on routine physical examination indicators using machine learning methods: A cohort study. *Frontiers in Cardiovascular Medicine, 9*, Article 854287.

Rao, C. J., Gao, M. Y., Wen, J. H., & Goh, M. (2022a). Multi-attribute group decision making method with dual comprehensive clouds under information environment of dual uncertain Z-numbers. *Information Sciences, 602*, 106–127.

Rao, C. J., Liu, M., Goh, M., & Wen, J. H. (2020). 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers. *Applied Soft Computing, 95*, Article 106570.

Rao, C. J., Liu, Y., & Goh, M. (2022b). Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model. in press *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747 -022- 00854-y.

Rao, C. J., Wang, C., Hu, Z., Xiao, X. P., & Goh, M. (2022c). Grey uncertain linguistic multi-attribute group decision making method based on GCC-HCD. in press *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2022.3166526.

Rao, C. J., Zhang, Y., Wen, J. H., Xiao, X. P., & Goh, M. (2023). Energy demand forecasting in China: A support vector regression-compositional data second exponential smoothing model. *Energy, 263*, Article 125955.

Rosenson, R. S., & Marcovina, S. M. (2018). Refining lipoprotein(a) associated cardiovascular risk in women. *Journal of the American College of Cardiology, 72*(3), 297–299.

Saba, P. S., Parodi, G., & Ganau, A. (2021). From risk factors to clinical disease: New opportunities and challenges for cardiovascular risk prediction. *Journal of the American College of Cardiology, 77*(11), 1436–1438.

Saeed, A., Nambi, V., Sun, W., Virani, S. S., Taffet, G. E., Deswal, A., … Ballantyne, C. M. (2018). Short-term global cardiovascular disease risk prediction in older adults. *Journal of the American College of Cardiology, 71*(22), 2527–2536.

Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using k-kearest neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of Advanced Research, 4*(4), 331–344.

Segar, M. W., Jaeger, B. C., Patel, K. V., Nambi, V., Ndumele, C. E., Correa, A., … Pandey, A. (2021). Development and validation of machine learning-based race-specific models to predict 10-year risk of heart failure a multicohort analysis. *Circulation, 143*(24), 2370–2383.

Sm, A., Smm, B., & Al, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 46–61.

Tay, D., Poh, C. L., Van Reeth, E., & Kitney, R. I. (2015). The effect of sample age and prediction resolution on myocardial infarction risk prediction. *IEEE Journal of Biomedical and Health Informatics, 19*(3), 1178–1185.

Tharwat, A., & Schenck, W. (2021). A conceptual and practical comparison of PSO-style optimization algorithms. *Expert Systems with Applications, 167*, Article 114430.

Tian, M. Z. (2015). Several hot issues in the research of statistical reconstruction in the era of big data. *Statistical Research, 32*(5), 3–12.

Tian, Z., & Chen, H. (2021). A novel decomposition-ensemble prediction model for ultra-short-term wind speed. *Energy Conversion and Management, 248*, Article 114775.

Tokodi, M., Schwertner, W. R., Kovács, A., Tősér, Z., Staub, L., Sárkány, A., … Kosztin, A. (2020). Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: The SEMMELWEIS-CRT score. *European Heart Journal, 41*(18), 1747–1756.

Wang, J., Rao, C. J., Goh, M., & Xiao, X. P. (2023). Risk assessment of coronary heart disease based on cloud-random forest. *Artificial Intelligence Review, 56*, 203–232.

Wang, Y. K., Pan, Z. B., & Dong, J. (2022). A new two-layer nearest neighbor selection method for KNN classifier. *Knowledge-Based Systems, 235*, Article 107604.

Xue, J., & Shen, B. (2020). A novel swarm intelligence optimization approach: Sparrow search algorithm. *Systems Science & Control Engineering an Open Access Journal, 8*(1), 22–34.

Yach, D., Hawkes, C., Gould, C. L., & Hofman, K. J. (2004). The global burden of chronic Diseases-Overcoming impediments to prevention and control. *JAMA, 291*(21), 2616–2622.

Yan, H. M., Jiang, Y. T., Zheng, J., Peng, C. L., & Li, Q. H. (2006). A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications, 30*(2), 272–281.

Yan, X., & Jia, M. (2018). A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing. *Neurocomputing, 313*, 47–64.

Zhang, M., Chen, W. L., Zhang, Y., Liu, F., Yu, D. S., Zhang, C. Y., & Gao, L. (2021). Fault diagnosis of oil-immersed power transformer based on difference-mutation brain storm optimized Catboost model. *IEEE Access, 9*, 168767–168782.

Zhang, Y., Zhao, Z., & Zheng, J. (2020). CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *Journal of Hydrology, 588*, Article 125087.

Zhao, Y., Wood, E. P., Mirin, N., Cook, S. H., & Chunara, R. (2021). Social determinants in machine learning cardiovascular disease prediction models: A systematic review. *American Journal of Preventive Medicine, 61*(4), 596–605.