ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Lost Circulation Prediction Method Based on an Improved Fruit Fly Algorithm for Support Vector Machine Optimization

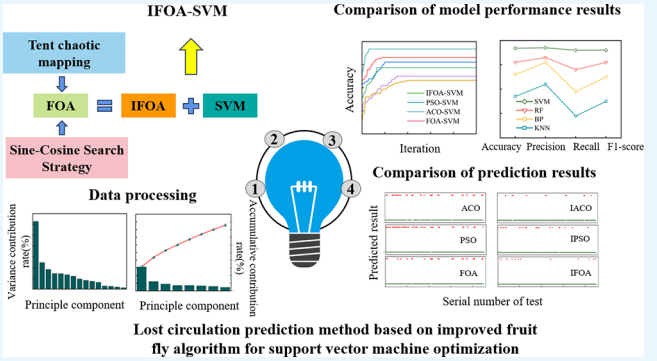
Song Deng,* Chunyu Pei, Xiaopeng Yan, Hongda Hao, Meng Cui, Fei Zhao, Chuchu Cai, and Yadong Shi

Cite This: *ACS Omega* 2023, 8, 32838−32847

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Lost circulation events during drilling operations are known for their abruptness and are difficult to control. Traditional diagnostic methods rely on qualitative indicators, such as mud pit volume changes or anomalous logging curve patterns. However, these methods are subjective and rely heavily on empirical knowledge, resulting in delayed or inaccurate predictions. To address this problem, there is an urgent need to develop efficient methods for a timely and accurate lost circulation prediction. In this study, a novel approach is proposed by combining principal component analysis (PCA) and empirical analysis to reduce the dimensionality of the model data. This dimensionality reduction helps to streamline the analysis process and improve prediction accuracy. The predictive model also incorporates an improved fruit fly optimization algorithm (IFOA) in conjunction with support vector machine (SVM) techniques. The actual instances of lost circulation serve as the evaluation criteria for this integrated method. To overcome the challenges associated with irregular population distribution within randomly generated individuals, a tent map strategy is introduced to ensure a more balanced and representative sample. In addition, the model addresses issues such as premature convergence and slow optimization rates by employing a sine−cosine search strategy. This strategy helps to achieve optimal results and speeds up the prediction process. The improved prediction model demonstrates exceptional performance, achieving accuracy, precision, recall, and F1 scores of 96.8, 97, 96, and 96%, respectively. These results indicate that the IFOA-SVM approach achieves the highest accuracy with a reduced number of iterations, proving to be an efficient and fast method for predicting the lost circulation events. Implementation of this methodology in drilling operations can lead to improved efficiency, reliability, and overall performance.



1. INTRODUCTION

In the process of drilling and completion, complex downhole conditions and accidents caused by inaccurate prediction have been threatening the whole process of drilling, which not only has a serious impact on drilling quality and drilling rate but also causes large economic and safety losses.[1,2] In the early days, the prediction of lost circulation accidents could only rely on human judgment by the relevant technicians through the real-time data collected by sensors. However, this requires the technicians to have a rich experience. Moreover, human judgment has a large subjective factor. In recent years, with the continuous development of intelligent algorithms, Chinese and foreign scholars have put forward many intelligent methods for predicting lost circulation accidents. These methods are more scientific and can predict lost circulation accidents more accurately.[3,4]

In the field of drilling engineering, the application of SVM in the lost circulation prediction model has become increasingly prominent. Both domestic and foreign researchers have utilized SVM to predict and prevent lost circulation accidents during the drilling process. For instance, Xie et al.[5] developed a neural network model and an SVM model to analyze and predict the risk of lost circulation based on 11 relevant indicators derived from drilling data. The results demonstrated the effectiveness and practicality of both models in predicting lost circulation incidents. Manshad et al.[6] employed SVM and radial basis function models to predict drilling fluid seepage in the Maroun oilfield. The findings highlighted the high accuracy of both the SVM model and the radial basis function model in predicting the drilling fluid seepage. Liu et al.[7] conducted an analysis of the characteristics of lost circulation and its influencing factors.

Received: June 6, 2023
Accepted: August 10, 2023
Published: August 31, 2023

They used SVM regression to predict lost circulation, improving the accuracy of their predictions.

Ahmed et al.[8] leveraged real-time drilling parameters, readily accessible as input parameters for their model. They constructed an SVM model as well as a radial basis function model to predict the extent of lost circulation in the formation. Wang et al.[9] employed the improved sparrow search algorithm to optimize the penalty parameter C and kernel parameter g of the support vector machine (ISSA-SVM) for predicting lost circulation accidents. This approach resulted in significant enhancements in both prediction accuracy and computing time.

It has been observed that while the SVM offers numerous advantages, it also presents certain challenges, such as the need for proper selection of kernel functions, kernel parameters, and penalty factors, as these factors significantly influence the classification results. Currently, scholars both domestically and internationally have proposed various methods for optimizing SVM parameters, including the particle swarm algorithm,[10] gradient descent method,[11] ant colony algorithm,[12] and genetic algorithm.[13] However, these methods have several limitations when applied to SVM parameter optimization. For instance, they may encounter issues such as getting stuck in local optima, which can hinder the attainment of the global optimum. Additionally, the time required to determine the optimal parameters may not meet expectations.

The fruit fly optimization algorithm (FOA) stands out from other algorithms in solving optimization problems due to its advantages of having fewer parameters and being easy to tune. This algorithm is frequently employed to search for optimal parameter combinations for the SVM in order to enhance their performance.

For instance, Shen et al.[14] utilized the FOA to search for optimal parameters for an SVM model and subsequently applied the optimized SVM to analyze a set of medical statistics. Yu et al.[15] took it a step further by using an adaptive step-size improved FOA to further optimize the parameters of an SVM model. They then applied the optimized model to predict the dynamic response of a magnetorheological elastomer-based vibration isolation device. Li et al.[16] developed a dynamic periodic strong control vector regression model for ship motion time horizon, incorporating a chaotic mesh representation. They optimized their model using an adaptive chaotic fruit fly algorithm.

However, the FOA has certain limitations that need to be addressed. The performance of a fruit fly's initial position significantly impacts the trajectory of its search. While fixing the step size may seem like a simplification, it can impose significant restrictions on the algorithm's ability to find the optimal solution. A step size that is too large may help the global search in the early stages, but it will hinder the precision and refinement of the search later on. Conversely, a step size that is too small impedes the global search, leading to a slow search process and a higher likelihood of getting stuck in local optima.

To overcome these challenges, this paper introduces a lost circulation prediction model that integrates a tent map with a positive cosine Drosophila algorithm optimized for SVM. By leveraging this novel approach, an accurate prediction of lost circulation accidents can be achieved while addressing the limitations associated with the FOA.

2. DATA PREPARATION

2.1. Data Processing. Due to the complex geological structure and the deviation of real-time logging data obtained by the sensor, as well as the difference in the original characteristics of the data, the numerical value presents a great difference, so the data need to be reprocessed. The data preprocessing process is as follows:

(1) Data collection: Collect raw logging data from logging tools or sensors;
(2) Data cleaning: Data cleaning and denoising can be achieved through various methods, such as spline interpolation, moving average, and local linear regression. These techniques help to eliminate errors and noise present in the data, ensuring a cleaner and more reliable data set for further analysis and modeling.
(3) Output value processing: The non-numerical data contained in the output value is numerically processed. For example, "0" indicates no lost circulation at a specific depth, and "1" indicates lost circulation.
(4) Data normalization: Eliminate the adverse effects of different orders of magnitude and dimensions in the analysis of combined multidata features.

In this paper, the standard max−min mapping transformation method is employed to normalize the model. The min-max normalization method involves determining the maximum and minimum values within the reorganized features. These extremum values are then used as the range to linearly scale the data within the features to the [0,1] interval. The deflation formula for this transformation is shown below:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x_{old} represents the original data in the data, x_{max} and x_{min} denote the maximum and minimum values in the features, respectively, and x_{new} denotes the normalized value in the original data after deflation.

The main flow of data processing is shown in Figure 1.

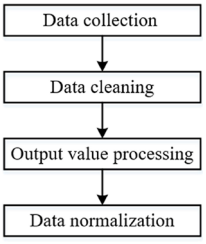


Figure 1. Data processing flowchart.

Using the above methodology, a data set consisting of 600 complete and valid instances, each containing 15 data features, was obtained. For a detailed breakdown of the data features in each category within the research area, refer to Table 1.

2.2. Lost Circulation Experience—PCA Method. *2.2.1. Artificial Empirical Characterization Methods.* Several factors come into play when encountering lost circulation while drilling in a formation. This work focuses specifically on the analysis of logging data to determine the occurrence of lost circulation. The analysis is based on the basic principles of several logging methods.[17−19]

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847

Table 1. Some of the Statistical Details about Logging Data in This Paper

no	parameter	mean	min	median	max
1	depth	4370.77	3832	4406	5203
2	DEN	2.33	1.41	2.38	2.59
3	CNL	13.93	0.04	0.33	60.97
⋮	⋮	⋮	⋮	⋮	⋮
14	GR	101.36	47.16	78.34	198.56
15	CAL	15.02	2.92	14.16	21.34

(1) Gamma logging: Lost circulation occurs when drilling fluids infiltrate the formation. In cases in which the drilling fluids are radioactive, this infiltration results in an elevated gamma curve (GR). Therefore, GR has been selected as the characteristic parameter to analyze and identify lost circulation events.

(2) Spontaneous potential logging: The presence of a leaky formation is characterized by its favorable porosity and permeability, which in turn leads to an anomalous spontaneous potential curve (SP). As the drilling fluids vary, the anomalies associated with leaky formation also become inconsistent. Therefore, SP was selected as a feature to capture and analyze these inconsistencies.

(3) Porosity logging: Lost circulation results in an increase in porosity, causing the compensated neutron log (CNL) to exhibit anomalous behavior. Therefore, CNL is selected as the feature to capture these changes.

(4) Resistivity logging: According to the detection depth, resistivity can be categorized as deep resistivity (Rd) and shallow resistivity (Rs). Rd provides information about the original formation, while Rs focuses on nearby formations and intrusive layers. The impact of seepage is more significant on SR. In the absence of seepage, the resistivity curves overlap, but in the presence of seepage, they diverge, forming a "double track" pattern. The disparity between the curves is more evident with severe seepage; thus, Rd and Rs are selected as characteristic parameters.

(5) Caliper well logging: After a drilling fluid leak occurs in the formation, it can cause damage to the well wall, potentially resulting in well collapse. During this process, there will be changes in the caliper well, and as a result, CAL is selected as a feature.

2.2.2. Principal Component Analysis. Excessive data can increase the complexity of the modeling algorithm and the computational requirements. Therefore, PCA is chosen to reduce data size, simplify the model, and improve algorithmic efficiency.[20,21] In this study, only the principal components with cumulative contribution rates greater than 85% are retained. The results of PCA are summarized in Table 2.

To make it easier and more intuitive to see the distribution of the difference contribution rate data in the table given above, Figures 2 and 3 are used to describe it.

In Figure 3, the horizontal coordinate is each principal component, the vertical axis indicates the variance contribution rate, the histogram shows the variance contribution rate of each principal component, and the line graph shows the cumulative variance contribution rate of the principal components. Combining Figures 2 and 3, the cumulative variance contribution of the first 8 components is 85.685% > 85%, indicating their importance in capturing the information on the data. Reducing the original 15-dimensional features to 8 dimensions reduces the computational complexity.

Table 2. Principal Component Analysis Results

serial number	principal component	variance contribution rate (%)	cumulative contribution (%)
1	SP	31.4	31.4
2	CAL	12.319	43.719
3	well depth	9.124	52.843
4	Rs	7.241	60.084
5	AC	7.177	67.261
6	Rd	6.677	73.938
7	CNL	6.026	79.964
8	Rt	5.721	85.685
9	GR	4.061	89.746
10	DEN	3.643	93.389
11	NGR	2.123	95.512
12	compensated sonic	1.491	97.003
13	compensation neutron	1.321	98.324
14	fluid density	0.991	99.315
15	gas content	0.685	100

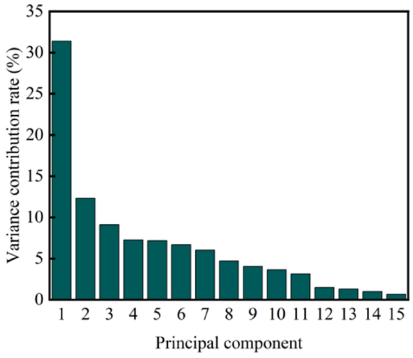


Figure 2. Variance contribution rate of 15 principal components.

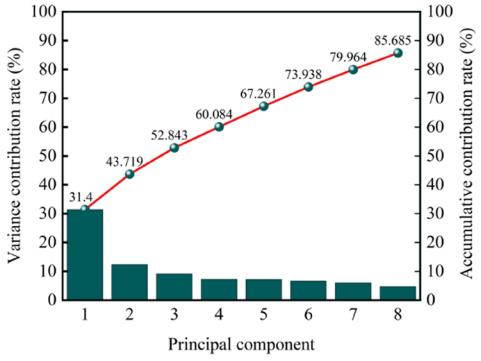


Figure 3. Variance contribution rate of principal components.

3. LOSS PREDICTION MODEL

3.1. Support Vector Machine. Assuming a linear system, input and output data are $\{x_i,y_i\}$, $(i = 1,2,\cdots,n)$, $x_i \in R^n$ is n system input vector, and $y_i \in R$ is the output vector. The basic idea of the SVM method is to map the input samples from the original dimension n to the high-dimension feature space F through the nonlinear transformation $\phi(\cdot)$, and construct an optimal linear function:[22,23]

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847

$$f(x) = \omega^T \varphi(x) + b \tag{2}$$

The variables ω and b are the normal vector and intercept of the hyperplane, respectively.

The standard SVM uses ε as an insensitive loss function and risk minimum estimation to establish the objective optimization form:

$$\min \frac{1}{2}\omega^T\omega + c\sum_{i=1}^{N}(\xi + \xi_i^*)\varepsilon$$

$$s{\cdot}t\begin{cases} y_i - \omega^T\varphi(x_i) - b \le \varepsilon + \xi_i \\ \omega^T\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i \ge 0,\ \xi_i^* \ge 0,\ i = 1, 2, \cdots, n \end{cases} \tag{3}$$

where c is the equilibrium factor in the formula and is assumed to be 1; ξ_i, ξ_i^* is the penalty factor, which is the extent to which the sample point exceeds the fitting accuracy ε. If the sample point is within the accuracy range, $\xi_i = \xi_i^* = 0$, $i = 1,2,\cdots,n$.

According to the objective function and its constraint requirements in eq 3, the Lagrange equation can be established:

$$\begin{aligned} L(\omega, b, \xi_i, \xi_i^*) = {} & \frac{1}{2}\omega^T\omega + c\sum_{i=1}^{N}(\xi + \xi_i^*) \\ & - \sum_{i=1}^{n}\alpha_i[\varepsilon + \xi_i + y_i - \omega^T\varphi(x_i) - b] \\ & - \sum_{i=1}^{n}\alpha_i^*[\varepsilon + \xi_i^* + y_i - \omega^T\varphi(x_i) - b] \\ & - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{4}$$

where $\alpha_i, \alpha_i^* \ge 0$, $\eta_i, \eta_i^* \ge 0$, $i = 1, 2, \cdots, n$.

The partial derivative of each variable in eq 4 is calculated, and the dual optimization problem can be obtained by permutation where x_i matching $\alpha_i - \alpha_i^* \ne 0$ is the support vector. The variable ω that represents the complexity of the function is a linear combination of the mapping function $\phi(\cdot)$. Therefore, the computational complexity of system identification by SVM is independent of spatial dimension and depends on the number of samples.[24,25] The kernel function is used instead of nonlinear mapping:

$$\Psi(x_i, x_j) = \varphi(x_i^T)\varphi(x_j) \tag{5}$$

Optimization eq 4 can be converted to solving the dual problem:

$$\begin{aligned} \max J = {} & -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\Psi(x_i, x_j) \\ & - \varepsilon\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) \end{aligned}$$

$$s{\cdot}t{\cdot}\begin{cases} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*), & \alpha_i^* \notin [0, c] \\ \alpha_i, & \alpha_i^* \in [0, c] \end{cases} \tag{6}$$

By solving the quadratic programming problem of eq 6, we can obtain the standard result of system identification of SVM.

Some parameters in machine learning that need to be set manually and can be tuned to improve the performance and effect of learning are called hyperparameters. The super parameters in SVM are penalty factor C and Gaussian kernel parameter gamma. C represents the tolerance capacity of the model to errors, and gamma reflects the distribution of data mapped to the high-dimensional feature space. The larger C and gamma are, the more likely they are to overfit, and the smaller they are, the more likely they are to underfit. The selection of these two super parameters directly affects the prediction performance of the SVM model. In order to improve the prediction accuracy of the model, it is necessary to optimize the two parameters by an intelligent algorithm.

3.2. Improved Fruit Fly Optimization Algorithm.

3.2.1. Fruit Fly Optimization Algorithm. FOA is a novel swarm intelligence optimization algorithm inspired by the principles of bionics. In this algorithm, when a fruit fly in the population discovers a food source, other individuals use a visual search to determine the location of the optimal individual. At the same time, the remaining flies move closer to their optimal individual. The search process continued until the FOA reached its maximum number of iterations. Inspired by the foraging behavior of fruit flies, we derived this algorithm with a set of effective numerical solution methods. By implementing these methods, the FOA not only reduces the time required for data exploration, but also significantly improves the practicality and applicability of numerical solutions.[26,27]

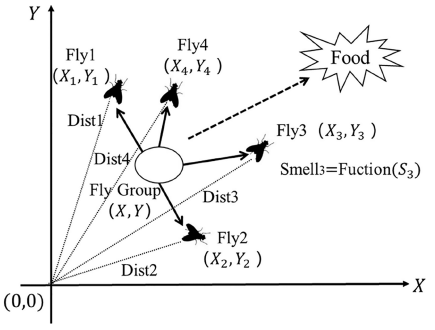


Figure 4. Foraging process of Drosophila.

Figure 4 depicts the foraging behavior of the fruit fly population. Based on the behavior characteristics of the fruit fly, the FOA can be divided into the following steps:

(1) Random initial fruit fly population position.

$$\text{Init}X_{\_\text{axis}} \tag{7}$$

$$\text{Init}Y_{\_\text{axis}} \tag{8}$$

(2) The random direction and distance of fruit fly individuals using olfactory methods to search for food.

$$X_i = X_{\_\text{axis}} + \text{Random Value} \tag{9}$$

$$Y_i = Y_{\_\text{axis}} + \text{Random Value} \tag{10}$$

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847

(3) Because the food location cannot be known, the distance between the point and the origin (Dist_i) is first estimated, and then the taste concentration determination value (S_i) is calculated, which is the reciprocal of the distance.

$$\mathrm{Dist}_i = \sqrt{X_i^2 + Y_i^2} \tag{11}$$

$$S_i = \frac{1}{\mathrm{Dist}_i} \tag{12}$$

(4) The odor concentration determination value (S_i) is substituted into the odor concentration determination function (or *fitness function*) to obtain the odor concentration (Smell_i) of the individual position of the fruit fly.

$$\mathrm{Smell}_i = \mathrm{Function}(S_i) \tag{13}$$

(5) Find the fruit fly with the highest flavor concentration in this population (find the maximum)

$$[\mathrm{bestSmell\ bestIndex}] = \max(\mathrm{Smell}) \tag{14}$$

(6) Retain the best taste concentration value and x- and y-coordinates when the fruit fly population uses vision to fly to the location.

$$\mathrm{Smellbest} = \mathrm{bestSmell} \tag{15}$$

$$X_{\_\mathrm{axis}} = X(\mathrm{bestIndex}) \tag{16}$$

$$Y_{\_\mathrm{axis}} = Y(\mathrm{bestIndex}) \tag{17}$$

(7) Go to iterative optimization, repeat steps(2)−(5), and determine whether the flavor concentration is better than the previous iteration. If so, execute step (6).

3.2.2. Improved Fruit Fly Optimization Algorithm. 3.2.2.1. Tent Map. The map is a common phenomenon in nonlinear systems characterized by randomness, traversal, and regularity. Utilizing this property to construct the initial position of Drosophila can not only maintain the diversity of the population but also make the distribution of the Drosophila population more uniform.[28,29] Tent Map has a uniform traversal effect and faster convergence, and its expression is

$$x_{k+1} = \begin{cases} 2x_k, & 0 \le x_k \le 0.5 \\ 2(1 - x_k), & 0.5 \le x_k \le 1 \end{cases} \tag{18}$$

However, the Tent chaotic sequence suffers from the presence of small and unstable cycle points, which limits its effectiveness. To address this drawback and prevent the Tent chaotic sequence from getting trapped in these unstable cycles, this study draws inspiration from the work of literature.[30] By introducing a random variable, an improved expression for chaos is proposed as follows:

$$x_{k+1} = \begin{cases} 2\left(x_k + \mathrm{rand}(0,\ 1) \times \dfrac{1}{\mathrm{NT}}\right), & 0 \le x_k \le 0.5 \\ 2\left(1 - \left(x_k + \mathrm{rand}(0,\ 1) \times \dfrac{1}{\mathrm{NT}}\right)\right), & 0.5 \le x_k \le 1 \end{cases} \tag{19}$$

$$x_{i,j} = x_{\min,j} + x_{k,j} \times (x_{\max,j} - x_{\min,j}) \tag{20}$$

where NT represents the number of particles in the sequence, and rand(0,1) is a random number ranging from 0 to 1, which produces a distinct value for each iteration, ensuring the unique characteristics of the Tent map. The process of generating the initial population using Tent chaotic mapping involves the following specific steps:

Step 1: Randomly generate the initial value (0,1), noting k = 0;

Step 2: Iterate using eq 19 with j and k incremented by 1 to generate the sequence of $x_{k,j}$, iterate with eq 20 with i incremented by 1 to generate the sequence of $x_{i,j}$, and similarly to generate the sequence of $y_{k,j}$, which is then the matrix initialized by the population.

3.2.2.2. Sine Cosine Algorithm. In FOA, the Drosophila updates positions around the best Drosophila using a step size and calculates the flavor concentration decision value for candidate solutions. The choice of strategy impacts the search process and results. The original fruit fly position update is influenced by the optimal fruit fly's position and step size, leading to potential regional extreme traps and long search times.[31−33] In this paper, we propose using the sine−cosine strategy instead of randomly generated directions and distances to improve the convergence speed and solve the problem of getting only regional extremes. Using the sine−cosine search, the updated formula for the position of an individual is shown in eq 21.

$$X_{ij}^{g+1} = \begin{cases} X_{ij}^{g} + r_1 \times \sin(r_2) \times \left| r_3 X_{\_\mathrm{axis}}^{g} - X_{ij}^{g} \right|, & r_4 \le 0.5 \\ X_{ij}^{g} + r_1 \times \cos(r_2) \times \left| r_3 X_{\_\mathrm{axis}}^{g} - X_{ij}^{g} \right|, & r_4 > 0.5 \end{cases}$$

$$Y_{ij}^{g+1} = \begin{cases} Y_{ij}^{g} + r_1 \times \sin(r_2) \times \left| r_3 Y_{\_\mathrm{axis}}^{g} - Y_{ij}^{g} \right|, & r_4 \le 0.5 \\ Y_{ij}^{g} + r_1 \times \cos(r_2) \times \left| r_3 Y_{\_\mathrm{axis}}^{g} - Y_{ij}^{g} \right|, & r_4 > 0.5 \end{cases} \tag{21}$$

where g is the number of iterations, X_{ij}^{g}, Y_{ij}^{g} denotes the position of the ith Drosophila in the jth dimension at the gth iteration, and $X_{\_\mathrm{axis}}^{g}$, $Y_{\_\mathrm{axis}}^{g}$ is the contemporary optimal position.

To balance the global search and local refinement development, the sine−cosine search strategy is improved. The assignment of variable r has a great impact on the overall search performance. By nonlinearly adjusting r, the global search is enhanced first, followed by the local refinement search. The transition from global to local is accelerated by nonlinear decreasing. The tuned expression is shown below:

$$r_1 = a \times \mathrm{e}^{-2\left(\frac{g}{\mathrm{max}gen}\right)^4} \tag{22}$$

where maxgen is the maximum number of iterations.

3.3. Model Evaluation Method. The evaluation of the generalization performance of the lost circulation classifier not only requires effective and feasible experimental methods but also needs to measure the generalization ability of the model, namely, performance measurement.[34−36] Lost circulation prediction models need to be measured using different indicators. When the performance of the classifier is evaluated, different performance indexes often lead to different judgment results. In an attempt to comprehensively assess the effectiveness of the model's performance, this study used evaluation methods such as accuracy, precision, recall, F1-

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847

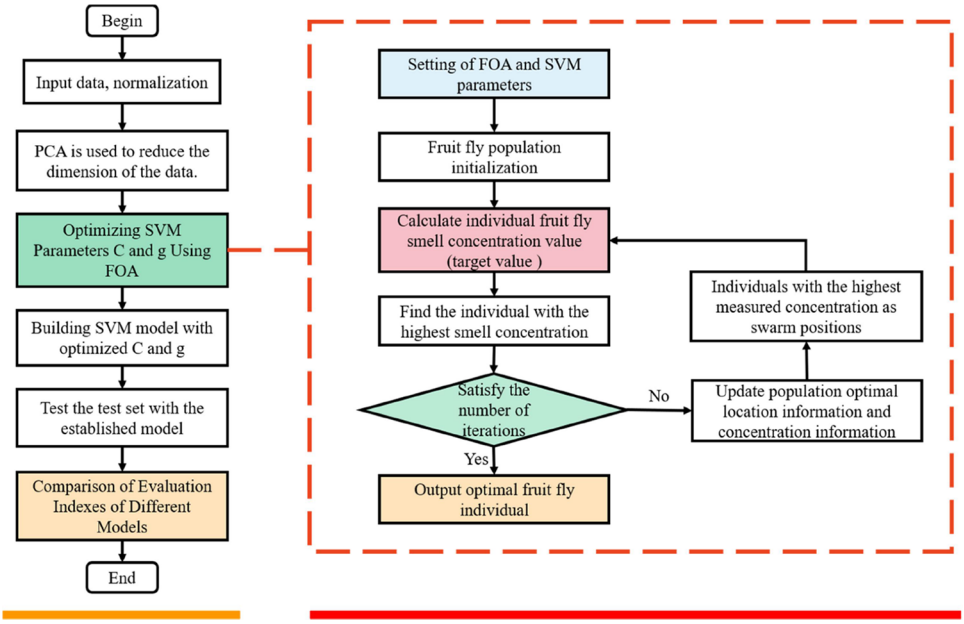


Figure 5. IFOA optimization SVM flowchart.

score, confusion matrix, and area under the subject's working characteristic curve. Based on the actual results of the lost circulation and the predicted results predicted by the algorithm, the samples are classified into true positive (TP), false positive (FP), false negative (TN), and false negative (FN). The accuracy is calculated as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{23}$$

The precision is calculated as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{24}$$

The recall is calculated as follows:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{25}$$

The F1 score is calculated as follows:

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{26}$$

The confusion matrix organizes the true classification of each sample by row and the predicted classification of each sample by column. The confusion matrix can reflect the proportion of all the predicted results of the classification model to the real results to observe the performance of the model in different classification categories.

The receiver operating characteristics (ROC) curve is utilized to assess the generalization performance with the area under the curve (AUC) serving as the evaluation metric. The ROC curve is derived from the true positive rate (TPR) and false positive rate (FPR) at different identification levels. TPR and FPR are calculated as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{27}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{28}$$

The ROC curve can show the change in a classifier's ability to identify samples at different thresholds. The AUC calculated by the ROC curve can show the advantages and disadvantages of the classification performance. The evaluation result is generally between 0.5 and 1, and the larger the value is, the better the classification performance is.

3.4. Forecasting Process. In this paper, we apply IFOA to optimize the parameters of SVM. Specifically, we utilize the fitness function of the actual problem as the concentration function for the fruit flies. Following the principle of fitness optimization, the IFOA iteratively optimizes the parameters until the algorithm's completion, ultimately selecting the optimal model parameters. The main process is depicted in Figure 5.

4. RESULTS AND DISCUSSION

4.1. Experimental Environment. The paper utilized a computer operating system of Windows 10, 64-bit, and the development tool used was Python 3.9.7. The hardware environment of the computer consisted of an Inter(R) Xeon Platinum 8373C CPU with a clock speed of 2.60 GHz, and a total of 256 Gigabytes of RAM.

4.2. Application Example. The lost circulation experimental data were collected from 13 wells in the southern margin zone of the Junggar Basin. A total of 600 data sets were obtained from the logging curve data and comprehensive well completion logs. Of these, 300 data sets represented drilling data under normal conditions, while the remaining 300 data sets represented drilling data during lost circulation events. A combination of PCA and empirical analysis was used to generate the input parameters. The 300 sets of drilling data for normal conditions and 300 sets of drilling data during lost circulation were divided into a training set and a test set according to a 7:3 ratio. This division facilitated the classification prediction of lost circulation accidents.

To optimize the SVM models, classification prediction models were established using various algorithms: IFOA, FOA, improved ant colony optimization (IACO), ant colony optimization (ACO), particle swarm optimization (PSO),

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847

and improved particle swarm optimization (IPSO). Where IACO draws on ref 37 methodology, IPSO draws on ref 38 methodology. Each algorithm had a population size of 20 and a maximum iteration count of 100. The value range for penalty factor C and gamma was set to [0, 1000]. The SVM model achieved the best performance when the penalty factor C was set to 102 and the gamma to 0.02. Figure 6 displays the fitness curves of the six algorithms in optimizing the SVM for the lost circulation classification prediction.

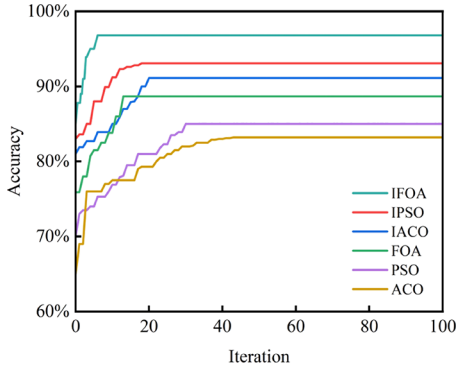


Figure 6. Finding the optimal parameter fitness curve.

Figure 6 clearly demonstrates that the IFOA achieves the highest fitness for classification accuracy with only six iterations. In contrast, the IPSO, IACO, FOA, PSO, and ACO algorithms require 18, 22, 13, 30, and 43 iterations, respectively, to reach their optimal fitness values. None of these algorithms reach the same high level of fitness as the IFOA. These results indicate that the IFOA outperforms other optimization algorithms in quickly and accurately identifying penalty parameter C and kernel parameter g of SVM.

All six algorithms are employed to optimize penalty parameter C and kernel parameter g of the SVM in constructing the lost circulation classification prediction model. Comparative analysis of the six algorithms reveals that IFOA-SVM exhibits the best prediction performance, achieving an accuracy of 96.8%. This accuracy surpasses the prediction accuracies of IPSO-SVM, IACO-SVM, FOA-SVM, PSO-SVM, and ACO-SVM models by 3.8, 5.68, 8.12, 11.8, and 13.8%, respectively. These findings highlight that the IFOA combined with SVM not only achieves faster optimization of parameters but also yields higher classification prediction accuracy compared with other intelligent optimization algorithms. The optimization effect of the algorithm is significant. The accuracy of the three algorithms' prediction results on the test set is illustrated in Figure 7.

4.3. Model Performance Evaluation. To evaluate the prediction performance of the SVM model, three other machine learning models, namely, random forest (RF), backpropagation (BP) neural network, and K-nearest neighbors (KNN), were selected for comparison. The training parameters after the hyperparameter optimization of the RF model are set as follows: the number of trees is 110; the minimum number of sample splits is 2; and the minimum number of leaf node samples is 1.

For the BP model, after hyper-parameter optimization, the input layer consisted of the prediction index for lost circulation, the output layer determined the presence or absence of lost circulation, and there were two hidden layers. The learning rate of the network was set to 0.06. As for the KNN model, after hyper-parameter optimization, the number of neighbors was set to 3, and the distance metric used was Euclidean distance.

In this study, we conducted a comprehensive analysis of various evaluation metrics, including accuracy, precision, recall, F1-score, ROC, and confusion matrix. The comparison of these metrics for the four models is depicted in Figure 8. From

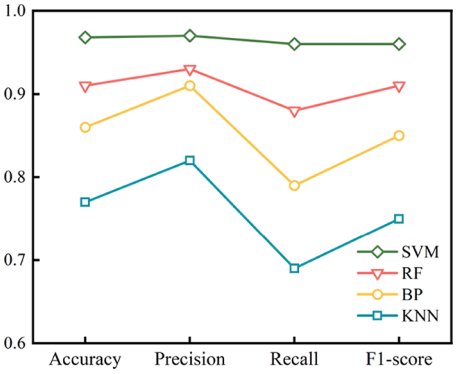


Figure 8. Comparison of the results of the four model evaluation indicators.

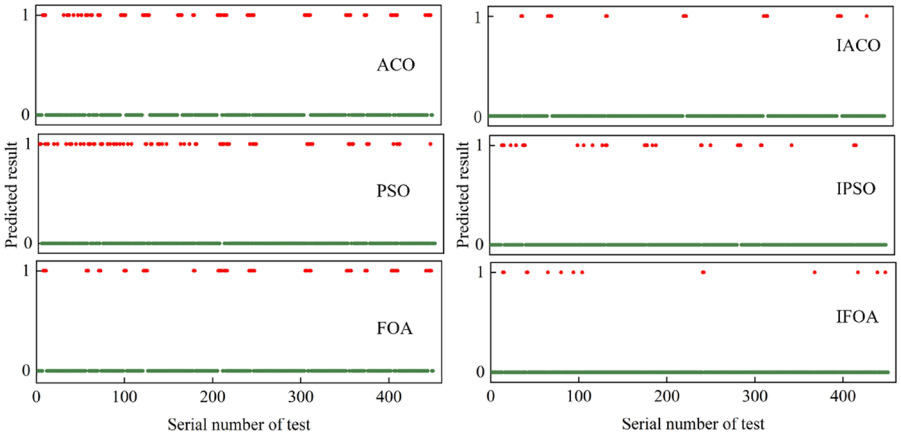


Figure 7. Comparison of the prediction results of the 6 algorithms, where 0 represents a correct prediction and 1 represents an incorrect prediction.

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847

the figure, it is evident that the SVM model outperforms the other models in all four metrics, demonstrating a superior prediction accuracy for the test set of lost circulation samples. Specifically, the IFOA-SVM model achieved an accuracy of 96.8%, precision of 97%, recall of 96%, and F1 score of 96%.

These findings highlight the effectiveness and reliability of the IFOA-SVM model in accurately predicting lost circulation incidents. The high performance across multiple evaluation metrics further solidifies the superiority of the SVM approach in this particular context.

Figure 9 presents a comparison of the confusion matrices for the four models. Each row represents the percentage of

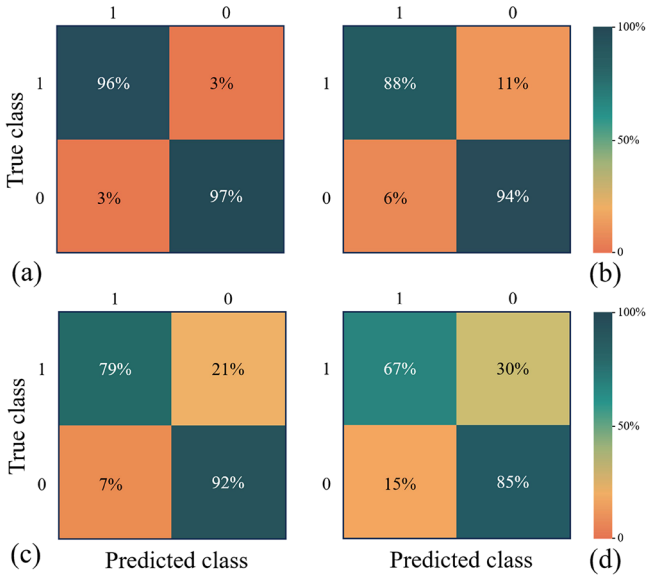


Figure 9. Confusion matrix of four models to identify lost circulation: (a) SVM; (b) RF; (c) BP; and (d) KNN.

samples classified as lost circulations or nonlost circulations by the respective model.

Figure 9 compares the confusion matrices of the four models. The SVM model shows the highest accuracy, correctly identifying 96% of lost circulations (289 samples) and 97% of nonlost circulations (291 samples). The RF model accurately predicts 88% of lost circulations (265 samples) and 94% of nonlost circulations (283 samples). The BP model achieves 79% identification for lost circulations (237 samples) and 92% identification for nonlost circulations (278 samples). The KNN model achieves 67% identification for lost circulations (201 samples) and 85% identification for nonlost circulations (255 samples).

In the confusion matrix, the IFOA-SVM model demonstrates a 94% identification accuracy for nonlost circulation data. For instances where seepage occurs, the model achieves an 88% recognition accuracy. These results indicate that the IFOA-SVM model performs well in accurately identifying segments without seepage. However, when it comes to data with lost circulation, there may be instances where the impact of lost circulation is minimal, leading to a slightly lower recognition accuracy compared with cases without lost circulation.

To assess the generalization performance of the proposed algorithm, the desired generalization performance is evaluated here by using ROC curves. The results of this evaluation are depicted in Figure 10.

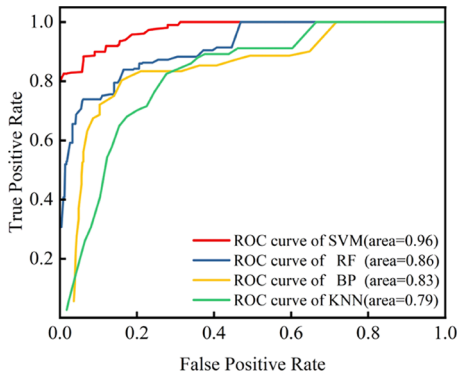


Figure 10. ROC curves of four models for identifying lost circulation.

Based on this figure, it is evident that the IFOA-SVM model achieves the highest AUC value on the ROC curve. This signifies that the IFOA-SVM model demonstrates a superior capability in recognizing lost circulations and exhibits a more effective classification performance. Taking into account all of the classification metrics, the IFOA-SVM model attains the optimal classification for lost circulations in this study.

5. CONCLUSIONS

We propose an IFOA and integrate it with SVM to establish a predictive model for lost circulation. By applying this approach, we draw the following conclusions:

(1) Through a comparative analysis of the six algorithms, it can be concluded that the IFOA-SVM model demonstrates the highest prediction effectiveness, achieving an accuracy of 96.8%. In comparison, the IPSO-SVM, IACO-SVM, FOA-SVM, PSO-SVM, and ACO-SVM models achieved prediction accuracies of 3.8, 5.68, 8.12, 11.8, and 13.8%, respectively.

(2) Evaluation metrics, including accuracy, precision, recall, F1 score, ROC, and confusion matrix, were comprehensively analyzed. The IFOA-SVM model achieved the highest accuracy, precision, recall, and F1-score (96.8, 97, 96, and 96%, respectively) among the four models. It also demonstrated superior performance in the confusion matrix and ROC curves. These results confirm the IFOA-SVM model's strong generalization ability on the test samples and its suitability for building a reliable lost circulation prediction model.

While this paper presents a lost circulation prediction model utilizing a relatively novel machine learning method, it is important to note that the model has yet to be fully implemented in field applications. Thus, future research will focus on the following areas:

(1) To enhance the prediction accuracy of lost circulation incidents, it is crucial to acquire additional geological parameters that are closely associated with the phenomenon. By the addition of the model with these relevant geologic parameters, the prediction accuracy can be significantly improved.

(2) Optimize the lost circulation prediction model, such as solving the effect of imbalance between lost circulation sample data and nonlost circulation sample data, to improve the accuracy of the model.

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847

AUTHOR INFORMATION

Corresponding Author

Song Deng − *School of Petroleum Engineering, Changzhou University, Changzhou 213100, China;* orcid.org/0000-0003-4322-8534; Email: dengsong@cczu.edu.cn

Authors

Chunyu Pei − *School of Petroleum Engineering, Changzhou University, Changzhou 213100, China;* orcid.org/0000-0001-6235-3635

Xiaopeng Yan − *School of Petroleum Engineering, Changzhou University, Changzhou 213100, China*

Hongda Hao − *School of Petroleum Engineering, Changzhou University, Changzhou 213100, China*

Meng Cui − *CNPC Engineering Technology Research and Development Company Limited, Beijing 102206, China*

Fei Zhao − *CNPC Engineering Technology Research and Development Company Limited, Beijing 102206, China*

Chuchu Cai − *School of Petroleum Engineering, Changzhou University, Changzhou 213100, China*

Yadong Shi − *School of Petroleum Engineering, Changzhou University, Changzhou 213100, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c03919

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the Program of Polar Drilling Environmental Protection and Waste Treatment Technology (2022YFC28064003) and Supported by CNPC-CZU Innovation Alliance. The authors would like to express their sincere gratitude to all participants for their valuable contributions.

REFERENCES

(1) Gan, C.; Cao, W.-H.; Wu, M.; Chen, X.; Hu, Y.-L.; Liu, K.-Z.; Wang, F.-W.; Zhang, S.-B. Prediction of drilling rate of penetration (ROP) using hybrid support vector regression: A case study on the Shennongjia area, Central China. *J. Pet. Sci. Eng.* **2019**, *181*, No. 106200.

(2) Deng, S.; He, J.; Kang, B.; Liu, W.; Ling, D.; Pei, C. Wellbore Flow Model and Process Optimization for Gas-Lift Leakage Drilling for Shallow Shale Formations. *ACS Omega* **2022**, *7* (9), 7806−7815.

(3) Lu, C.; Wu, M.; Chen, X.; Cao, W.; Gan, C.; She, J. Torsional vibration control of drill-string systems with time-varying measurement delays. *Inf. Sci.* **2018**, *467*, 528−548.

(4) Alkinani, H. H.; Al-Hameedi, A. T. T.; Dunn-Norman, S.; Flori, R. E.; Amer, A. S. Applications of Artificial Neural Networks in the Petroleum Industry: A Review. In *SPE Middle East Oil and Gas Show and Conference*, 2019.

(5) Xie, P.; Jiang, L. W.; Zhao, Y.; He, H. L. Research on real-time prediction method of well surge and leakage based on neural network. *Modern Comput.* **2018**, *611* (11), 23−28.

(6) Hou, X.; Yang, J.; Yin, Q.; Liu, H.; Chen, H.; Zheng, J.; Wang, J.; Cao, B.; Zhao, X.; Hao, M. Lost circulation prediction in south China sea using machine learning and big data technology. In *Offshore Technology Conference*; OTC, 2020; p D041S053R005.

(7) Liu, B.; Li, C.; Li, S.; Wang, G.; Liu, H. Lost circulation prediction based on SVR. *Drilling Prod. Technol.* **2019**, *42* (6), 17−20.

(8) Abbas, A. K.; Bashikh, A. A.; Abbas, H.; Mohammed, H. Q. Intelligent decisions to stop or mitigate lost circulation based on machine learning. *Energy* **2019**, *183*, 1104−1113.

(9) Wang, X.; Zhang, Q. Improved Sparrow Search Algorithm to Optimize Lost Circulation Prediction of Support Vector Machine. *Sci. Technol. Eng.* **2022**, *22* (34), 15115−15122.

(10) Fan, Y.; Zhang, C.; Xue, Y.; Wang, J.; Gu, F. Vibration.A bearing fault diagnosis using a support vector machine optimized by the self-regulating particle swarm. *Shock Vib.* **2020**, *2020*, No. 9096852.

(11) Mahdi, G. J. M. A Modified Support Vector Machine Classifiers Using Stochastic Gradient Descent with Application to Leukemia Cancer Type Dataset. *Baghdad Sci. J.* **2020**, *17* (4), 1255−1255.

(12) Li, Y.; Yang, P.; Wang, H. Short-term wind speed forecasting based on improved ant colony algorithm for LSSVM. *Cluster Comput.* **2019**, *22*, 11575−11581.

(13) Gokulnath, C. B.; Shantharajah, S. An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Comput.* **2019**, *22*, 14777−14787.

(14) Shen, L.; Chen, H.; Yu, Z.; Kang, W.; Zhang, B.; Li, H.; Yang, B.; Liu, D. Evolving support vector machines using fruit fly optimization for medical data classification. *Knowledge-Based Syst.* **2016**, *96*, 61−75.

(15) Yu, Y.; Yousefi, A. M.; Yi, K.; Li, J.; Wang, W.; Zhou, X. A new hybrid model for MR elastomer device and parameter identification based on improved FOA. *Smart Struct. Syst.* **2021**, *28* (5), 617−629.

(16) Li, M.-W.; Geng, J.; Han, D.-F.; Zheng, T.-J. Ship motion prediction using dynamic seasonal RvSVR with phase space reconstruction and the chaos adaptive efficient FOA. *Neurocomputing* **2016**, *174*, 661−680.

(17) Yin, Q.; Yang, J.; Tyagi, M.; Zhou, X.; Hou, X.; Cao, B. Field data analysis and risk assessment of gas kick during industrial deepwater drilling process based on supervised learning algorithm. *Process Saf. Environ. Protec.* **2021**, *146*, 312−328.

(18) Sabah, M.; Talebkeikhah, M.; Agin, F.; Talebkeikhah, F.; Hasheminasab, E. J. Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: A case study from Marun oil field. *J. Pet. Sci. Eng.* **2019**, *177*, 236−249.

(19) Nasiri, A.; Ghaffarkhah, A.; Moraveji, M. K.; Gharbanian, A.; Valizadeh, M. Experimental and field test analysis of different loss control materials for combating lost circulation in bentonite mud. *J. Nat. Gas Sci. Eng.* **2017**, *44*, 1−8.

(20) Kim, Y.; Johnson, M. S.; Knox, S. H.; Black, T. A.; Dalmagro, H. J.; Kang, M.; Kim, J.; Baldocchi, D. Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. *Global Change Biol.* **2020**, *26* (3), 1499−1518.

(21) Saccenti, E.; Camacho, J. Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometr. Intell. Lab. Syst.* **2015**, *149*, 99−116.

(22) Alkinani, H. H.; Al-Hameedi, A.; Dunn-Norman, S. Predicting the risk of lost circulation using Support Vector Machine model. In *ARMA US Rock Mechanics/Geomechanics Symposium*; ARMA, 2020; p ARMA-2020-1154.

(23) Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189−215.

(24) Faccini, D.; Maggioni, F.; Potra, F. A. Robust and distributionally robust optimization models for linear support vector machine. *Comput. Oper. Res.* **2022**, *147*, No. 105930.

(25) Chen, K.; Yang, X.; Song, X. An Intelligent Diagnosis Method for Lost Circulation Based on Engineering Logging Data. *China Pet. Machin.* **2022**, *11*, No. 003.

(26) Hu, G.; Xu, Z.; Wang, G.; Zeng, B.; Liu, Y.; Lei, Y. Forecasting energy consumption of long-distance oil products pipeline based on improved fruit fly optimization algorithm and support vector regression. *Energy* **2021**, *224*, No. 120153.

(27) Cong, Y.; Wang, J.; Li, X. Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm. *Procedia Eng.* **2016**, *137*, 59−68.
(28) Gokhale, S.; Kale, V. An application of a tent map initiated Chaotic Firefly algorithm for optimal overcurrent relay coordination. *Int. J. Electr. Power Energy Syst.* **2016**, *78*, 336−342.
(29) Huang, H.; Feng, X. A.; Zhou, S.; Jiang, J.; Chen, H.; Li, Y.; Li, C. A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features. *BMC Bioinf.* **2019**, *20*, 290.
(30) Fan, Y.; Wang, P.; Heidari, A. A.; Wang, M.; Zhao, X.; Chen, H.; Li, C. Rationalized fruit fly optimization with sine cosine algorithm: a comprehensive analysis. *Expert Syst. Appl.* **2020**, *157*, No. 113486.
(31) Li, M.; Lu, X.; Wang, X.; Lu, S.; Zhong, N. Biomedical classification application and parameters optimization of mixed kernel SVM based on the information entropy particle swarm optimization. *Comput. Assist. Surg.* **2016**, *21* (Supp 1), 132−141.
(32) Pan, Q.-K.; Sang, H.-Y.; Duan, J.-H.; Gao, L. An improved fruit fly optimization algorithm for continuous function optimization problems. *Knowledge-Based Syst.* **2014**, *62*, 69−83.
(33) Wang, W.; Zhang, M.; Liu, X. Improved fruit fly optimization algorithm optimized wavelet neural network for statistical data modeling for industrial polypropylene melt index prediction. *J. Chemom.* **2015**, *29* (9), 506−513.
(34) Yacouby, R.; Axman, D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 2020; pp 79−91.
(35) Haghighi, S.; Jasemi, M.; Hessabi, S.; Zolanvari, A. PyCM: Multiclass confusion matrix library in Python. *J. Open Source Softw.* **2018**, *3* (25), 729.
(36) Caelen, O. A Bayesian interpretation of the confusion matrix. *Ann. Math. Artif. Intell.* **2017**, *81* (3−4), 429−450.
(37) Zhao, J.; Tang, Y. F.; Jiang, G. P.; Xu, F. Y.; Ding, J. Mobile robot path planning based on improved ant colony algorithm. *J. Nanjing Univ. Posts Telecommun., Nat. Sci. Ed.* **2019**, *39* (06), 73−78.
(38) Hu, Q.; Qu, R.; Hu, Z.; Gong, S. C. Research on Wine Classification Based on PCA-IPSO-SVM. *J. Chongqing Univ. Sci. Technol., Nat. Sci. Ed.* **2023**, *25* (02), 103−109.

https://doi.org/10.1021/acsomega.3c03919
ACS Omega 2023, 8, 32838−32847