Water Supply

Water Supply Vol 22 No 10, 7416 doi: 10.2166/ws.2022.323

Study on key technology of identification of mine water inrush source by PSO-LightGBM

Yuan Ji[a,b], Donglin Dong[a,b,*], Aoshuang Mei[a,b] and Zhonglin Wei[a,b]

[a] China University of Mining & Technology, Beijing 100083, China
[b] National Engineering Research Center of Coal Mine Water Hazard Controlling, Beijing 100083, China
*Corresponding author. E-mail: ddl_cumtb@126.com

DD, 0000-0002-0882-1781

ABSTRACT

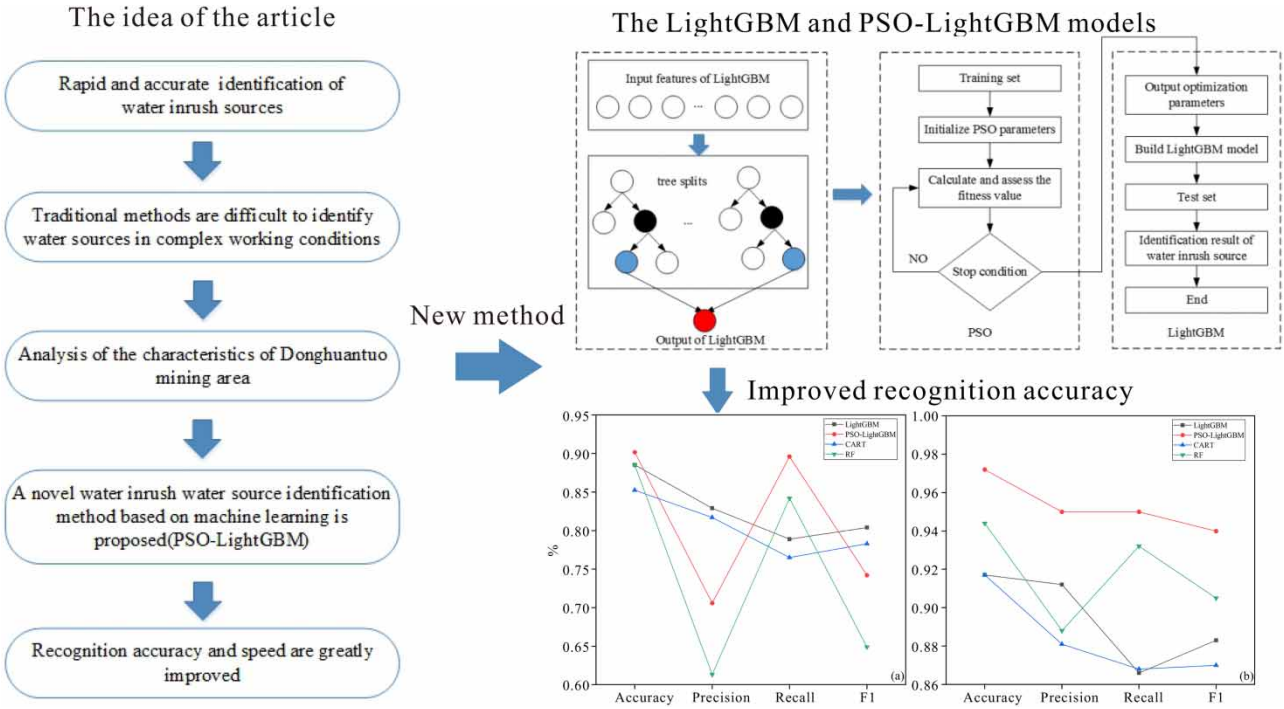
Mine water inrush is a major type of disaster in coal mine production in China. It causes heavy casualties and serious economic losses and threatens coal mine safety. To quickly and accurately identify mine water inrush source, according to the hydrochemical characteristics of different aquifers in the Donghuantuo mining area, this paper systematically analyzes the hydraulic connection of the aquifers in main coal mining areas before and after mining activities. Four types of hydrochemical data were collected: No. 5 coal seam roof water, No. 8 coal seam roof water, No. 122 coal seam floor water, and No. 1214 coal seam aquifer water in the Donghuantuo mining area. In addition, based on the hydrochemical data, the parameter selection of LightGBM was optimized by Particle Swarm Optimization (PSO) and constructed the PSO-LightGBM water inrush source identification model. The recognition accuracy of PSO-LightGBM model was compared with LightGBM model, classification regression tree (CART) model, and random forest (RF) model. The results showed that coal mining activities would have a significant impact on the water quality characteristics of the roof sandstone fissure water of No. 5 coal mine. Mining activities had a certain impact on the accuracy of the identification model. In addition, compared with the four recognition models, PSO-LightGBM model had the highest recognition accuracy of 97.22%. It showed that the model had high accuracy, stability, generalization ability, and important reference value for the identification of mine water inrush source.

Key words: hydrochemical characteristics, mine inrush, PSO-LightGBM, water source identification

HIGHLIGHTS

- The mine water environment changed significantly before and after the mining in the study area.
- Changes in water quality will affect the identification of water inrush sources to a certain extent.
- Establishment of PSO-LightGBM mine water source identification model.
- Comparison and analysis of PSO-LightGBM, LightGBM, RF and CART.
- The article analyzes the main reasons for the misjudgment of the model.

GRAPHICAL ABSTRACT



1. INTRODUCTION

The vigorous development of the coal industry is an important factor for the healthy and stable development of the economy (LaMoreaux *et al.* 2014). In recent years, with the continuous expansion of coal mining depth and width, the harm of mine water inrush has become more and more serious. Therefore, it is urgent to prevent and control mine water inrush. Quickly identifying the source of mine water inrush is a powerful tool to solve the problem. (Li 2018; Dong *et al.* 2020).

During the formation of coal mine water, the water contained in the lithosphere, hydrosphere, atmosphere, biosphere, and stratum is subject to various physical and chemical actions. Therefore, hydrochemical types are important information sources to characterize water-filled aquifers. The traditional method to judge the water source of coal mine is to classify according to the ion concentration of seven elements in groundwater directly or in combination with the identification model (Wang *et al.* 2020; Yang *et al.* 2021). In addition, some scholars analyzed the mine water source based on the change data of mine water level and temperature (Wu *et al.* 2019), and some scholars identify the water source based on the variation of mine water isotopes or trace elements (Singh *et al.* 2018; Guan *et al.* 2019). Some scholars determined the source of mine water according to fluorescence spectrum law of mine water (Yang *et al.* 2018; Hu *et al.* 2019).

Nowadays, the combination of traditional craftsmanship and computer technology is becoming more and more mature. Therefore, the identification methods of mine water inrush sources based on different water characteristics are constantly updated. Traditional methods of water temperature and water level discrimination or direct analysis of the hydrochemical data are being replaced by semi-quantitative analysis methods. The most widely used methods are combined with mathematical theory analysis, such as the cluster analysis method (Panagopoulos *et al.* 2016; Zhang *et al.* 2019), fuzzy mathematics method (Tiantian *et al.* 2019), grey theory method (Ju & Hu 2021), Bayesian discriminant method (Bogardi *et al.* 1982; Wu *et al.* 2016), Fisher feature extraction algorithm (Wang *et al.* 2021), GIS theoretical method (Donglin *et al.* 2012), SVM algorithm (Ma *et al.* 2018), and neural network algorithm (Chen *et al.* 2022; Yan *et al.* 2021). The above identification models are generally fast and effective, and machine learning methods have better applicability and advantages in identifying water inrush sources.

In the application of machine learning algorithm, Baudron *et al.* (2013) established a model for identifying water sources using Random Forest algorithm Saghebian *et al.* (2014) established a water sources classification model for a certain region in Iran using Decision Tree method and Gan *et al.* (2021) used LightGBM model to predict the downstream water level of the river.

To sum up, the research on the identification of mine water inrush sources has accumulated an extremely rich theoretical basis and application results. However, with the development of information technology and the comprehensive application of multidisciplinary means, many new methods and theories have emerged. Among them, machine-learning algorithm has been applied in the prediction of water quality and water level, but there is still research space in the application of mine water source identification. Therefore, based on the qualitative analysis of hydrochemical data in the mining area, firstly, we made a comparative analysis of the ion concentration characteristics and hydrochemical types of mine water in the Donghuantuo area. Then, we constructed a LightGBM algorithm model based on particle swarm optimization (PSO) to realize the identification of mine water sources. Finally, the PSO-LightGBM model was compared with the LightGBM model, Classification and Regression Tree (CART) model, and Random Forest (RF) model to determine the more suitable model for mine water inrush sources identification. The research results have well verified the application value of machine learning algorithms in mine water inrush water source identification, which can provide theoretical support for the realization of rapid water inrush water source identification, and provide technical support for mine water hazard prevention and control in similar geological coalfields.

2. STUDY AREA

2.1. Geological and hydrogeological conditions

Donghuantuo Mine is located in Fengrun District, Tangshan City, Hebei Province. The northeast of the study area is long and narrow, belonging to alluvial plain terrain. The terrain of the study area is flat and covered by Quaternary alluvium, which is in angular unconformity contact with bedrocks of various ages (Figure 1).

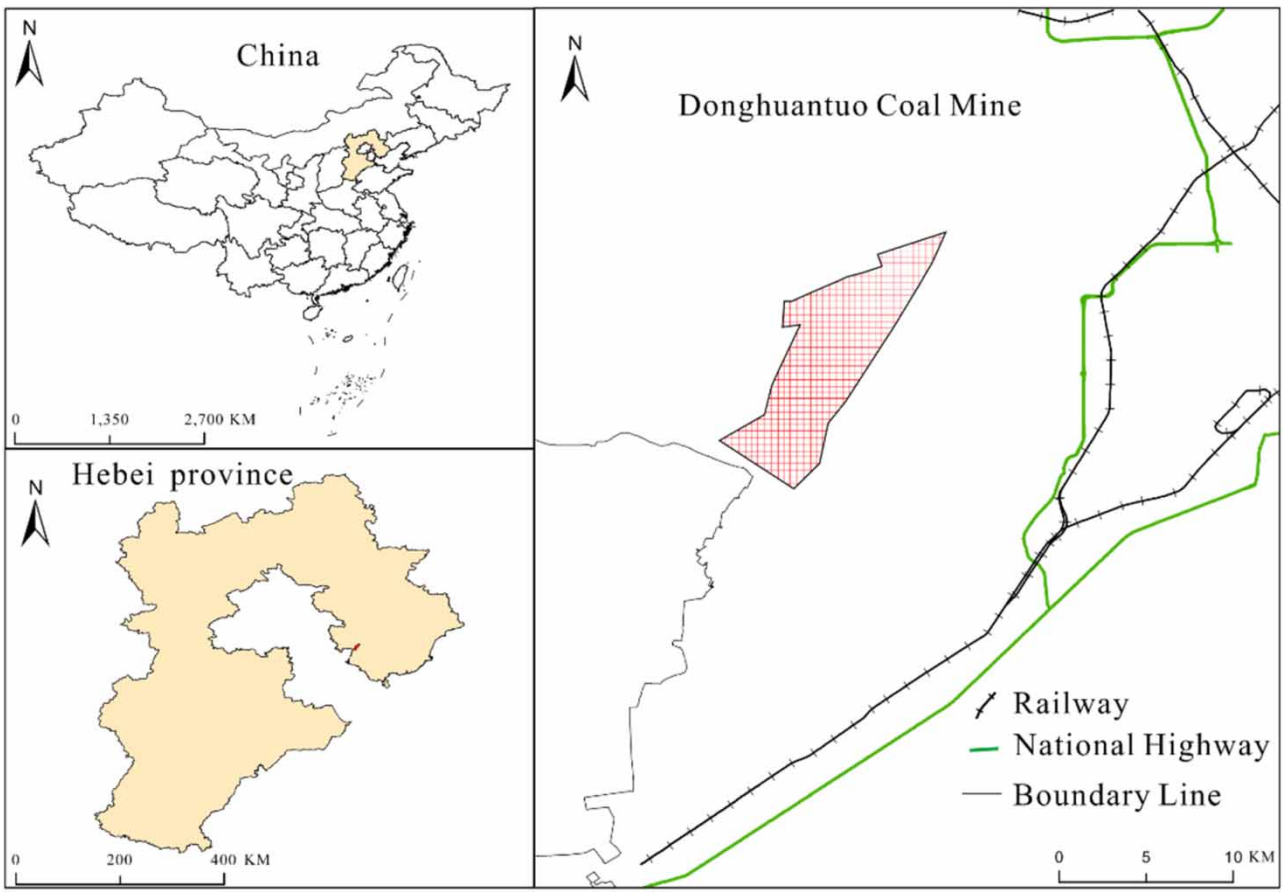


Figure1 | Geographic location of the Donghuantuo mine in Hebei Province, China.

Mine water in the study area mainly occurs in rock and soil pores, fissures, and karst fissures, which is a multi-aquifer structure (Figure 2). According to the special occurrence characteristics of the aquifer in the study area, there are three kinds of confined hydraulic systems in Donghuantuo coal: (1) quaternary pore – fissure aquifer, quaternary is all loose sediments, with stable water-repellent performance, and the water quality type is HCO_3^-- $Ca^{2+}Mg^{2+}Na^+$; (2) carboniferous and Permian sandstone fissure water, it includes the No.5 coal seam strong aquifer group, No.5 – No.12 coal seam weak water-bearing group Groups, No.1214 coal seam strong aquifers, and the water quality type is HCO_3^-- $Ca^{2+}Na^+$ or HCO_3^-- $Ca^{2+}Mg^{2+}$; (3) Ordovician limestone water, the top fissures, and karst caves are mostly filled with sand, gravel, and clay. The type of water quality is HCO_3^-- $Ca^{2+}Mg^{2+}$.

Figure 2 shows the stratigraphic system of Donghuantuo mine area, corresponding coal seams, stratigraphic column, lithology, and aquifers. After investigation, it was found that in various water inrush accidents in Donghuantuo mine, the floor water inrush occurred many times in 121 and 122 coal seams during the roadway excavation, and the water inrush source was from No. 1214 coal seam strong aquifer. Among them, on July 18th, 2002, floor water gushing occurred during the mining of No. 122 coal, and the maximum water volume was 1.94 m^3/min. In 2013 and 2015, water inrush disasters occurred

Strata System	Coal Seam	Stratigraphic Column	Lithology	Aquifers
Quaternary			Fine sand, coarse sand, clay	Pore-fissure aquifer
Permian			Dark gray sandstone, fine sandstone, coarse sandstone	
	No.5, 6 coal		Siltstone, clay rock	Crack water in sandstone roof of No.5 coal seam
	No.7-121 coal		Gray medium□fine sandstone, siltstone, claystone	Roof water of No. 8 coal seam
Carboniferous	No.121,122 coal		Siltstone, clay rock	Floor sandstone fissure wate of No122 coal seam
	No.1214 coal		Light grey sandstone	No.1214 coal aquifer
			Siltstone, clay rock	
Ordovician			Gravel, clay rock	Ordovician limestone water

Figure 2 | Hydrogeological histogram of different water sources in the Donghuantuo mine.

during the mining of No.8 coal. The water inrush sources were both in the roof aquifer of the No.5 coal seam, and the maximum water volume reached 1.5 and 2.6 m^3/min respectively. Therefore, the main research objects were the crack water in the sandstone roof of No. 5 coal seam, No.8 coal roof water, No.122 coal seam floor sandstone fissure water, and No. 1214 coal aquifer. The approximate location of the water sample is shown in blue dots in Figure 2.

2.2. Data collection

Since the 1990s, 195 groups of hydrochemical data have been collected in Donghuantuo mine area. It includes No.5 coal seam roof sandstone fissure water (129 groups), No.8 coal roof water (12 groups), No.122 coal seam floor sandstone fissure water (36 groups), and No.1214 coal aquifers (18 groups). In this study, the original data samples mainly included cations (Ca^{2+}, Mg^{2+}, Na^{+}), anions (HCO_3^-, SO_4^{2-}, Cl^-), pH value, and total hardness (TH).

3. METHODS

3.1. LightGBM algorithm model

The LightGBM model was a newer algorithm proposed by Microsoft Research Asia in 2017 based on the gradient boosting framework of the decision tree algorithm (Ke *et al.* 2017), which has the advantages of good training effect and low computational complexity. It has been gradually applied to different types of data analysis tasks such as classification, regression, and sorting. Assuming a supervised dataset $X = \{(x_i, y_i)\}_{i=1}^{n}$, the purpose of the LightGBM algorithm is to find an approximation $f(x)$ of a function $\hat{f}(x)$, such that the function can minimize the specified fitness function $L(y, f(x))$. The fitness function is used to judge how well the model fits the data. The optimization function can be expressed as:

$$\hat{f} = arg\min_{f} E_{y,X} L(y, f(x)) \tag{1}$$

At the same time, the LightGBM model integrates k regression trees to fit the final model, which can be expressed as:

$$f_k(x) = \sum_{i=1}^{k} f_t(X) \tag{2}$$

The regression tree can be expressed as $w_{q(x)}$, $q \in \{1, 2, \ldots, J\}$, where w is the vector of leaf node sample weights, q is the regression tree structure, J is the number of leaves in the tree. Then, when the ith tree is obtained, all the information of the previous $(t-1)$-th tree needs to be used. Therefore, the objective function of the algorithm iteration ith generation is as follows:

$$\Gamma_t = L(y_i, F_{t-1}(x_i)) + \sum_{t=1}^{k} \Omega(f_k(x)) \tag{3}$$

where $\Omega(f_k(x))$ is the regularization term to prevent the model from overfitting the training data. In the optimization of the objective function, the objective function after the second-order Taylor expansion is expressed as:

$$\Gamma_t = \sum_{i=1}^{n} (L(y_i, F_{t-1}(x_i)) g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) + \sum_{t=1}^{k} \Omega(f_k(x)) \tag{4}$$

where g_i and h_i are the first-order and second-order gradient statistics of the fitness function, respectively. Since the regression tree has been defined above, the complexity of a tree is:

$$\Omega(f_i(x)) = \gamma J + \frac{1}{2} \lambda \sum_{j=1}^{J} w_j^2 \tag{5}$$

where J is the number of leaf nodes, γ is the leaf node coefficient, and λ is the regularization coefficient. Assuming that

$I_j = \{i|q(x_i) = j\}$ is the sample set divided into leaf nodes, the objective function can be changed to:

$$\Gamma_t = \sum_{j=1}^{J}\left(\left(\sum_{i\epsilon I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\epsilon I_j} h_i + \lambda\right)w_j^2\right) + \gamma J \tag{6}$$

For a certain tree structure *q(x)*, the optimal weight of each leaf node is divided into:

$$w_j^* = -\frac{\sum_{i\epsilon I_j} g_i}{\sum_{i\epsilon I_j} h_i + \lambda} \tag{7}$$

When the structure *q(x)* of the tree is determined, the corresponding objective function is:

$$\Gamma_J^* = -\frac{1}{2}\sum_{j=1}^{J}\frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I_j} h_i + \lambda} + \gamma J \tag{8}$$

When the calculation method of the objective function is determined, it is necessary to calculate the profit of splitting the leaf nodes of the tree, so that each tree has the smallest objective function, and select the feature with the largest profit as the splitting feature, and continue to iterate until the specified conditions are met. Assuming that $I = I_L \cup I_R$ is the parent sample set, and I_L and I_R are the sample sets of the left and right branches, respectively, the income of each node split is:

$$G = \frac{1}{2}\left(\frac{\left(\sum_{i\epsilon I_L} g_i\right)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{\left(\sum_{i\epsilon I_R} g_i\right)^2}{\sum_{i\epsilon I_R} h_i + \lambda} - \frac{\left(\sum_{i\epsilon I} g_i\right)^2}{\sum_{i\epsilon I} h_i + \lambda}\right) \tag{9}$$

On this basis, LightGBM adopts a variety of methods to carry out in-depth optimization for the problems of the increasingly large training data volume and the increasing data feature dimensions. The histogram algorithm is used to discretize the floating-point eigenvalues in the sample into K integers, which can effectively reduce the computational cost and storage cost (Figure 3(a)). At the same time, the Leaf-wise leaf growth strategy (Figure 3(b) and 3(c)) is adopted, which can achieve better accuracy, significantly reduce algorithm complexity, and greatly reduce training time consumption, thereby improving training efficiency and prediction accuracy. Combining the above characteristics, the training process of the LightGBM model is shown in Figure 3(d).

3.2. Particle swarm optimization algorithm model

Particle Swarm Optimization (PSO) is an evolutionary computing technique, which is derived from the study on birds predation behavior, first proposed by Kennedy & Eberhart (1995). PSO is also an iterative-based optimization tool. It first initializes a set of random solutions in the system, takes each individual as a particle without weight and volume in the n-dimensional space, and then searches for the optimal value through iteration, so that the particles in the solution space can be searched according to the optimal particle. This algorithm has a rapid searching speed and good initial convergence, so it is widely used in many fields.

3.3. PSO-LightGBM algorithm optimization model

The parameters of the LightGBM model will directly affect the training speed and accuracy of the model. PSO has unique advantages in optimizing LightGBM parameters, which can effectively improve the effectiveness and accuracy of LightGBM. To optimize the training and recognition prediction of the model on this experimental data set, this paper uses the PSO to

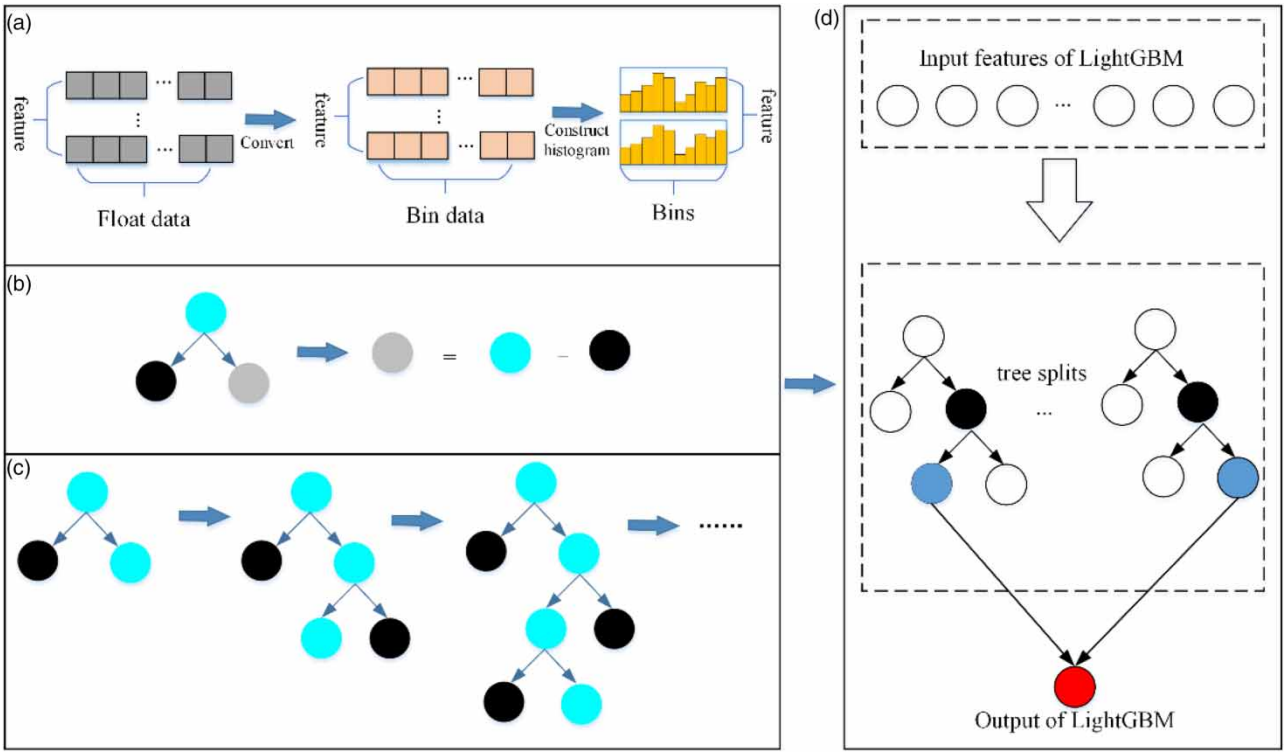


Figure 3 | The optimization strategy of the LightGBM model. (a) Histogram algorithm; (b) The principle of histogram acceleration; (c) Leaf-wise tree growth; (d) Model training process.

optimize the four parameters (num_leaves, min_data_in_leaf, learning_rate, max_depth) that have a great influence on the model. The technical route of the PSO-LightGBM model is shown in Figure 4. The basic steps of the optimization process are:

Step 1: Initialize particle swarm parameters, including the number of particles, learning rate, weighting coefficient, and the maximum number of iterations.

Step 2: Train the LightGBM model. The parameters that need to be optimized change as the particles position changes.

Step 3: Calculate and evaluate the fitness value. The fitness value is derived from the negative training accuracy score output by the LightGBM model, which is used to evaluate the performance of the particle swarm algorithm. The smaller the fitness value, the better the performance.

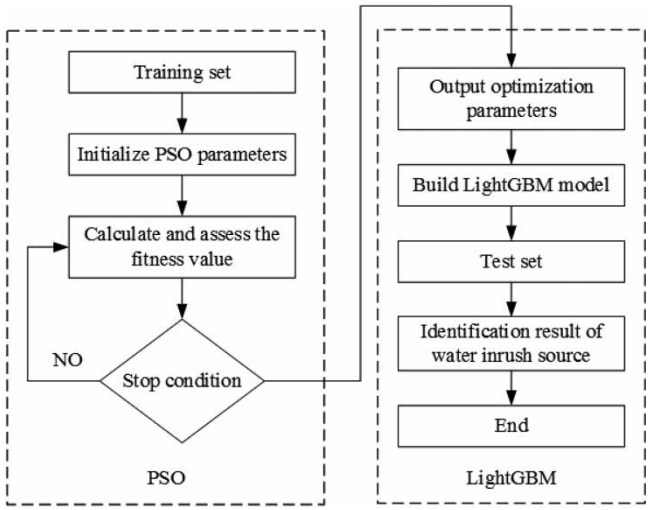


Figure 4 | The training process of the PSO-LightGBM model.

Step 4: Determine the stop state. When the number of iterations is reached, the iterative process is terminated to obtain the optimal parameters of the LightGBM model. Otherwise, the iterative calculation is performed.

Step 5: Validate the classification results of the model. The LightGBM model established by the optimization results is used to output the water inrush water source identification results.

4. RESULTS AND DISCUSSION

4.1. Hydrochemical characteristics

Analyze and discuss the collected hydrochemical data. The difference in ion concentration distribution in the aquifer and the difference in ion hydrochemical type in the same aquifer can be observed by the Piper trilinear diagram and Durov diagram (Figure 5).

Figure 5(a) and (b) shows the Piper's trilinear diagram and Durov diagram of No.5 coal roof water; (c) and (d) are Piper's trilinear diagram and Durov diagram of No.8 coal roof water; (e) and (f) are Piper's trilinear diagram and Durov diagram of No.122 coal floor water; (g) and (h) are Piper's trilinear diagram and Durov diagram of the No.1214 coal aquifer.

It can be seen from Figure 5(a) and 5(b) that the concentrations of Ca^{2+} and Mg^{2+} in the cations have decreased significantly after 2016, while the Na^{+} content has increased significantly, and the content of HCO_3^- is always predominant in the anion. The hydrochemical type changed from HCO_3^- -Ca^{2+} (71.8%) to HCO_3^--Na^{+} (100%) after 2016. The TDS increased significantly, indicating that the groundwater environment changed significantly before and after coal mining, and the runoff conditions became worse.

Through Piper's trilinear diagram and Durov diagram of No.8 coal roof water (Figure 5(c) and 5(d)), No.12-2 coal roof water (Figure 5(e) and 5(f)), and No.12-14 coal aquifer (Figure 5(g) and 5(h)) it can be seen that the Ca^{2+} was always dominant in cations, HCO_3^- was always dominant in anions, and the hydrochemical type has not changed before and after 2016, all of which was HCO_3^- Ca^{2+} water (100%). TDS was less than 420 mg/L before and after 2016, and there was no significant change. Compared with No.5 coal roof hydrogeological characteristics, mining activities have less impact on No.8 coal roof water, No.12-2 coal floor water, and No.12-14 coal aquifer.

All data samples mainly include cations (Ca^{2+}, Mg^{2+}, Na^{+}), anions (HCO_3^-, SO_4^{2-}, Cl^-), PH value, and total hardness (TH) as the original discriminant indicators using the principal component analysis (PCA) method for processing. Among them, the correlation coefficient matrix between various water source components is shown in Table 1.

It can be seen from Table 1 that the correlation coefficient value of Mg^{2+} and Ca^{2+} is 0.893, and the correlation coefficient value of HCO_3^- and Na^{+} is 0.967, indicating that there is a strong correlation between some variables. The correlation coefficient values of SO_4^{2-} and Na^{+}, HCO_3^- are 0.403 and 0.389, respectively, and the correlation coefficient values of Cl^- and Mg^{2+} is 0.213, indicating that some variables had moderate correlations. According to different PCA dimensions, the cumulative contribution rate of extracted principal components is shown in Table 2.

As shown in Table 2, the variance eigenvalues of Ca^{2+}, Mg^{2+}, and Na^{+} are all greater than 1, indicating that they are of great significance for distinguishing different types of water sources, and their cumulative contribution rates are 51.532, 72.592, and 86.242% respectively. This showed that the first three principal components already carry most of the information of the original data and can accurately distinguish different types of water sources.

According to the distribution law of the information content of each principal component in the gravel diagram in Figure 6, it can be intuitively seen that the scatter points of Ca^{2+}, Mg^{2+}, and Na^{+} were located on the steep slope, while the characteristic values of the last five scatter points were all less than 1.

To sum up, the groundwater environment has changed significantly before and after coal mining. Therefore, to discuss whether the change of hydrochemical data will affect the water source identification results, we had put forward two cases for comparative analysis. The data of No.8 coal roof water (marked as B), No.122 coal floor water (marked as C), No.1214 coal aquifer (marked as D) remain unchanged: (1) water source identification based on all data of No. 5 coal (marked as A1); (2) water source identification based on the data after the change of No. 5 coal (marked as A2) after 2016.

4.2. Model identification results

The accuracy rate (A), the precision rate (P), the recall rate (R), and the F1 value were selected for the LightGBM model, PSO-LightGBM model, Classification and Regression Tree (CART) model, Random Forest (RF) model to compare and analyze their classification performance.

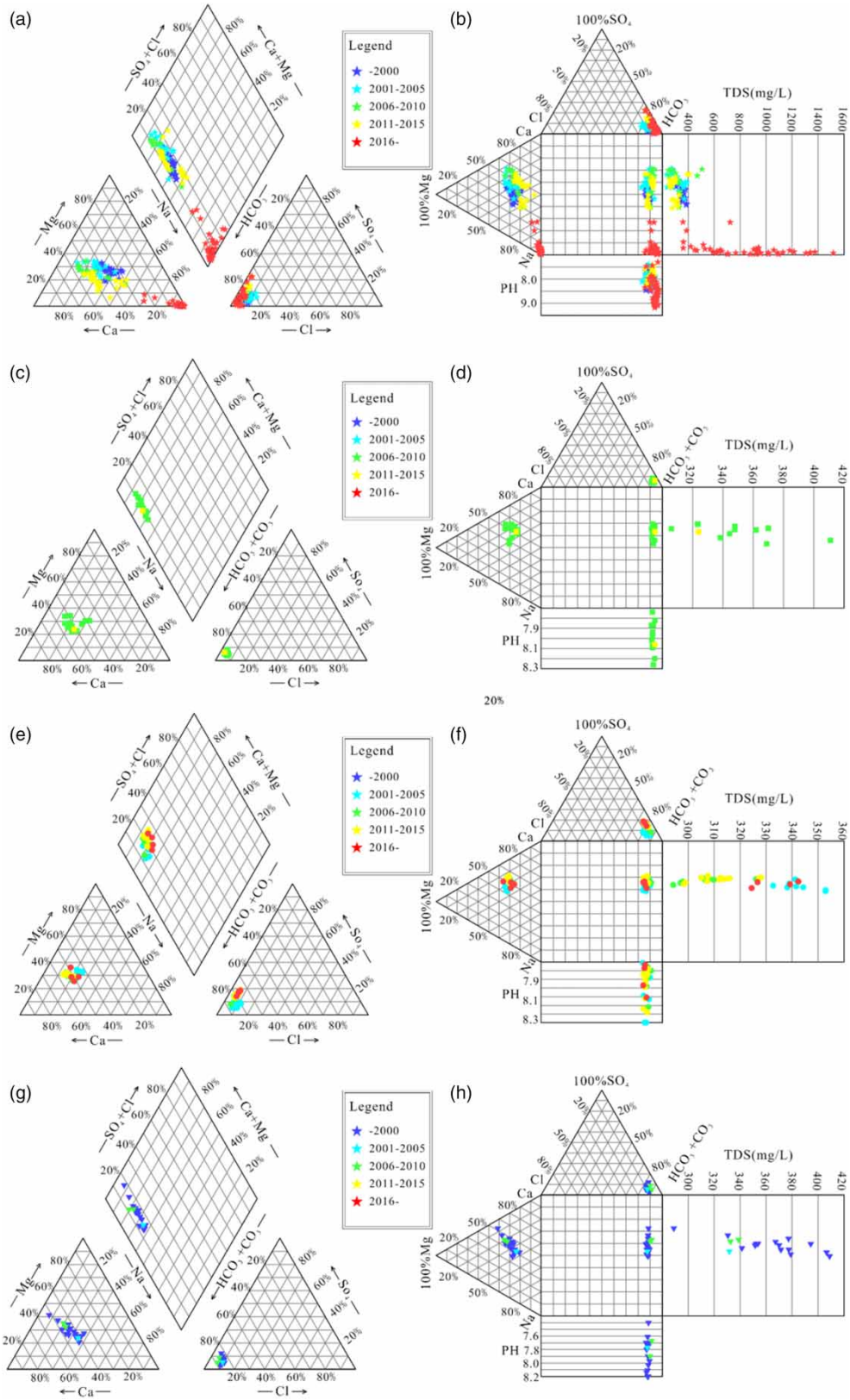


Figure 5 | Piper trilinear diagrams and Durov diagrams of water samples in the study area. (a), (b) Piper's trilinear diagram and Durov diagram of No.5 coal roof water; (c), (d) Piper's trilinear diagram and Durov diagram of No.8 coal roof water; (e), (f) Piper's trilinear diagram and Durov diagram of No.122 coal floor water; (g), (h) Piper's trilinear diagram and Durov diagram of the No.1214 coal aquifer.

Table 1 | Linear correlation coefficient matrix between various water source components

	Ca^{2+}	Mg^{2+}	Na^{+}	HCO_3^-	SO_4^{2-}	Cl^-	PH	TH
Ca^{2+}	1.000							
Mg^{2+}	0.893	1.000						
Na^{+}	−0.855	−0.790	1.000					
HCO_3^-	−0.728	−0.649	0.967	1.000				
SO_4^{2-}	−0.175	−0.185	0.403	0.389	1.000			
Cl^-	−0.043	0.213	0.038	0.081	−0.144	1.000		
PH	−0.595	−0.498	0.490	0.402	−0.148	0.113	1.000	
TH	0.536	0.319	−0.394	−0.344	0.081	−0.628	−0.349	1.000

Table 2 | The cumulative contribution rate of extracted principal components

	Initial eigenvalues			Extract the load sum of squares		
Element	Total	Percent variance	Cumulation %	Total	Percent variance	cumulation %
Ca^{2+}	4.123	51.532	51.532	3.864	48.301	48.301
Mg^{2+}	1.685	21.061	72.592	1.694	21.169	69.470
Na^{+}	1.092	13.649	86.242	1.342	16.772	86.242
HCO_3^-	0.486	6.074	92.316			
SO_4^{2-}	0.377	4.708	97.024			
Cl^-	0.171	2.138	99.163			
PH	0.062	0.781	99.943			
TH	0.005	0.057	100.000			

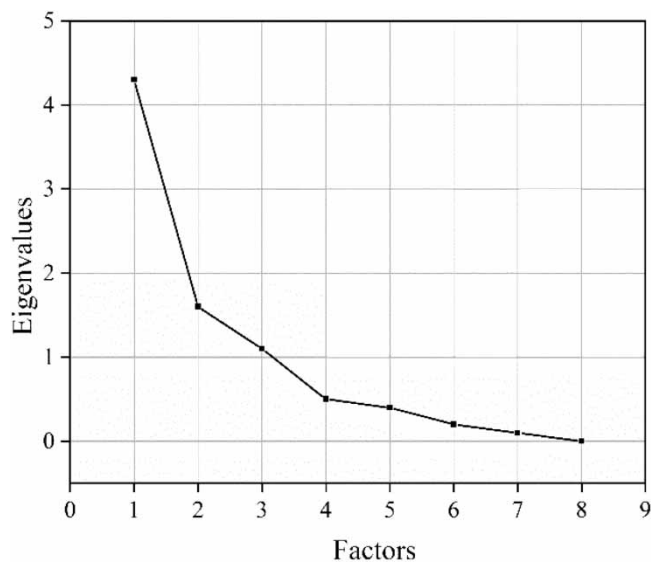


Figure 6 | The gravel diagram of each principal component information.

The multi-class problem is divided into multiple two-class problems for evaluation. There are four cases where the predicted results of the classifier are combined with the actual results on the dataset, as shown in Table 3.

Among them, T represents True, F represents False, P represents Positive, and N represents Negative. TP means predicts correctly, FP means predicts incorrectly, FN means predicts incorrectly, and TN means predicts correctly.

Accuracy (A) is the percentage of correctly classified samples to the total number of samples, with higher values indicating better identification accuracy. The expression is:

$$\mathrm{A} = \frac{TP + \mathrm{TN}}{TP + \mathrm{TN} + FP + \mathrm{FN}} \tag{10}$$

The precision rate (P) is the proportion of all predicted positive samples that are positive, and it is an evaluation index for the prediction result. The higher the P value, the more correctly predicted samples. The expression is:

$$P = \frac{TP}{TP + FP} \tag{11}$$

Recall rate (R) is the proportion of the actual positive samples identified as positive samples, and it is an evaluation index for the original sample data. The higher the R value, the more accurate the identification model. The expression is:

$$R = \frac{TP}{TP + FN} \tag{12}$$

Since P and R are difficult to achieve simultaneously large, they are all able to reflect the accuracy index. Therefore, we propose a score ($F1$), which combines the two index features of precision and recall, and takes a balance, the expression is:

$$F1 = \frac{2P{*}R}{P + R} \tag{13}$$

According to all the data of No.5 coal, the results of water source identification are shown in Figure 7(a), including 90 training sets and 39 test sets. In addition, based on the data after the change of No.5 coal mine in 2016, the identification results of water source are shown in Figure 7(b), including 31 training sets and 14 test sets. Meanwhile, No.8 coal roof water has eight training sets and four test sets; No.122 coal floor water has 23 training sets and 13 test sets; No.1214 coal aquifer has 13 training sets and five test sets. Further, the test sets data are shown in the Supplementary data.

Figure 7 shows the classification accuracy, precision, recall, and F1 value of LightGBM, PSO-LightGBM, CART, and RF models. It can be seen that in the prediction and classification models of water source types, PSO-LightGBM is slightly better than RF and better than LightGBM, and all three optimized classification models are better than CART.

The comparison and analysis of the results between the real and predicted values of the PSO-LightGBM model are shown in Figure 8. As shown in Figure 8, (a) is the comparison diagram of the real and predicted values of the PSO-LightGBM model based on all the data training sets of No.5 coal; (b) is a comparison diagram of PSO-LightGBM identification model based on the test sets. When we use all No.5 coal roof sandstone fissure water monitoring data as the identification model, the test results mainly have certain error effects on the No.8 coal roof water and the water source of the No.1214 coal aquifer.

The comparison and analysis of the results between the real and predicted values of the PSO-LightGBM model are shown in Figure 9. As shown in Figure 9, (a) is the comparison diagram of the real and predicted values of the PSO-LightGBM model

Table 3 | Parameter definitions of indices

	Actual result	
Prediction result	**1**	**0**
1	TP	FN
0	FP	TN

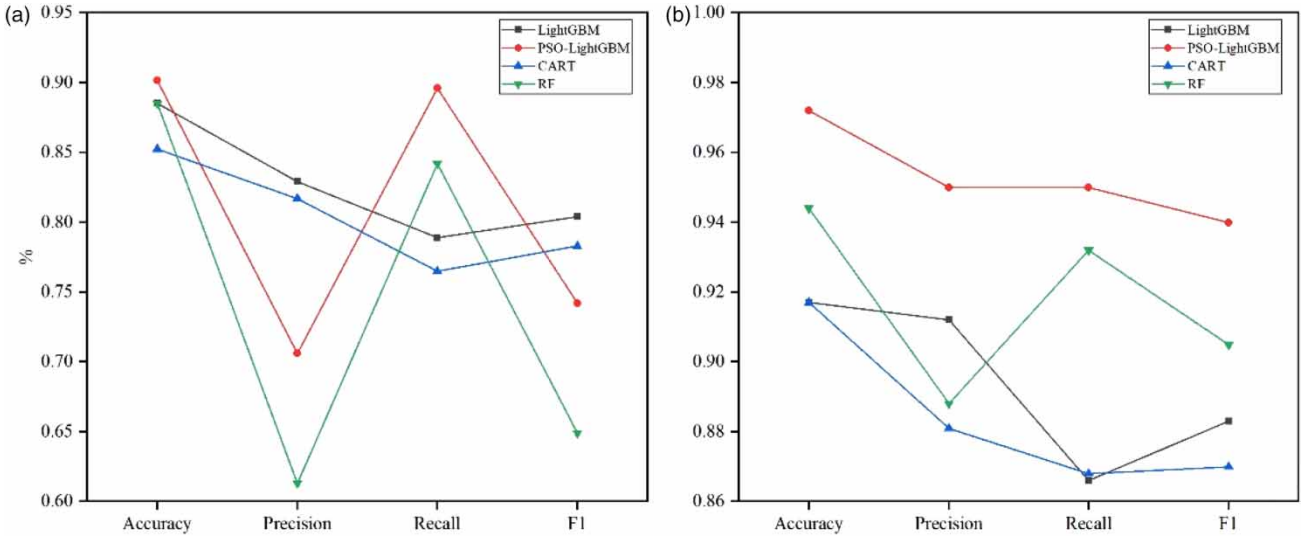


Figure 7 | Comparison of prediction performance of the four models. (a) All data for No. 5 coal; (b) All data for No. 5 coal after the changes in 2016.

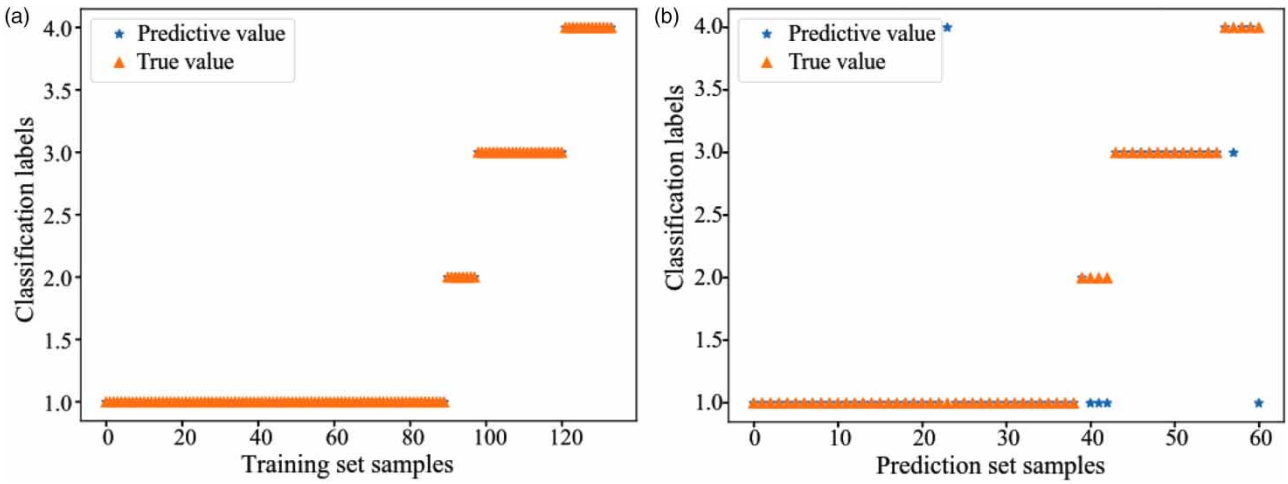


Figure 8 | Comparison of PSO-LightGBM identification model based on the overall data of No.5 coal. (a) The training set data; (b) The test set data.

based on the data training sets after the change of No. 5 coal; (b) is a comparison diagram of PSO-LightGBM identification model based on test sets.

When using the monitoring water sample data after coal mining as the identification model, it can be intuitively seen that the identification accuracy rate will be improved to a certain extent.

To sum up, from the comparison between the classification value and the real value and the analysis of the error results of each model, it can be seen that the classification value of PSO-LightGBM is closer to the real value, and the classification performance is better. It can be concluded that the data changes after mining have a certain influence on the results of the recognition model, and the optimized recognition model has higher accuracy.

5. CONCLUSIONS

The 196 hydrochemical data monitored in the Donghuantuo mining area in the past 30 years were deeply analyzed by traditional hydrochemical analysis methods, and a PSO-lightGBM water inrush source identification model was established based on the above data. The main research conclusions are as follows:

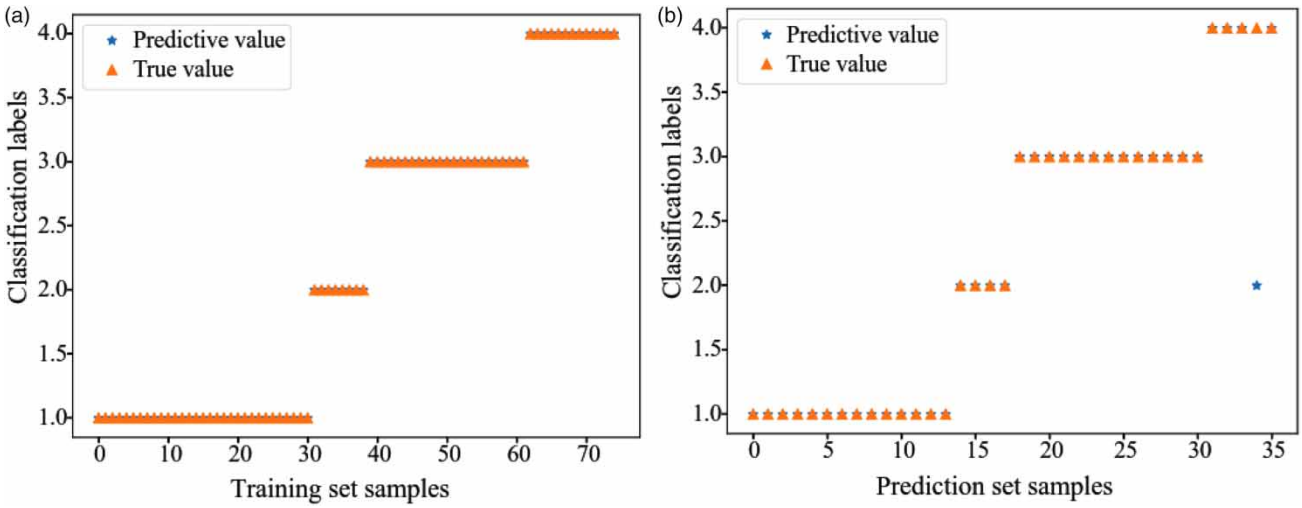


Figure 9 | Comparison of PSO-LightGBM identification model based on the data of No.5 coal after mining. (a) The training set data; (b) The test set data.

(1) We optimized the LightGBM model with particle swarm optimization (PSO) and established the PSO-LightGBM mine water source identification model. The model has the characteristics of simple operation and high identification accuracy. Furthermore, the identification accuracy of four identification models, including PSO-LightGBM, LightGBM, RF, and CART, were compared and analyzed. The identification accuracy of PSO-LightGBM is the highest, reaching 97.22%, and the recognition accuracy of the CART model is relatively low.

(2) The mine water environment changed significantly before and after the mining of No.5 coal seam in the study area. After 2016, its hydrochemical type changed from $HCO_3^- Ca^{2+}$ type water (71.8%) to $HCO_3^- Na^+$ type (100%). Moreover, there has been a significant increase in TDS, and this change will affect the identification of water inrush sources to a certain extent.

(3) Through the analysis of the identification factors, it is found that the main reason for the misjudgment of the model is that the water quality between adjacent aquifers is relatively similar, or the established model identification interval is not accurate enough. Especially after coal mining occurs, the mine water environment usually changes significantly, and it is necessary to analyze the changes of water samples in time, thereby improving the reliability of water sample data, and ultimately strengthen the accuracy of water inrush water source identification results.

ACKNOWLEDGEMENTS

This study was financially supported by the National Natural Science Foundation (41972255), the National Natural Science Foundation (U171020056), and the Ministry of Science and Technology of China (2017YFC0804104). In addition, Y. Ji is supported by the China Scholarship Council.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Baudron, P., Alonso-Sarría, F., García-Aróstegui, J. L., Cánovas-García, F., Martínez-Vicente, D. & Moreno-Brotóns, J. 2013 Identifying the origin of groundwater samples in a multi-layer aquifer system with random forest classification. *J. Hydrol.* **499**, 303–315.

Bogardi, I., Duckstein, L. & Szidarovszky, F. 1982 Bayesian analysis of underground flooding. *Water Resour. Res.* **18** (4), 1110–1116.

Chen, Y., Tang, L. & Zhu, S. 2022 Comprehensive study on identification of water inrush sources from deep mining roadway. *Environ. Sci. Pollut. Res.* **29**, 19608–19623.

Dong, S., Zheng, L., Tang, S. & Shi, P. 2020 A scientometric analysis of trends in coal mine water inrush prevention and control for the period 2000–2019. *Mine Water Environ.* **39**, 3–12.

Donglin, D., Wenjie, S. & Sha, X. 2012 Water-inrush assessment using a GIS-based Bayesian network for the 12-2 coal seam of the kailuan donghuantuo coal mine in China. *Mine Water Environ.* **31** (2), 138–146.

Gan, M., Pan, S., Chen, Y., Cheng, C., Pan, H. & Zhu, X. 2021 Application of the machine learning LightGBM model to the prediction of the water levels of the lower Columbia river. *J. Mar. Sci. Eng.* **9** (5), 496.

Guan, Z., Jia, Z., Zhao, Z. & You, Q. 2019 Identification of inrush water recharge sources using hydrochemistry and stable isotopes: a case study of Mindong No. 1 coal mine in north-east Inner Mongolia, China. *J. Earth Syst. Sci.* **128** (7), 1–12.

Hu, F., Zhou, M., Yan, P., Li, D., Lai, W., Zhu, S. & Wang, Y. 2019 Selection of characteristic wavelengths using SPA for laser induced fluorescence spectroscopy of mine water inrush. *Spectrochim. Acta A* **219**, 367–374.

Ju, Q. & Hu, Y. 2021 Source identification of mine water inrush based on principal component analysis and grey situation decision. *Environ. Earth Sci.* **80** (4), 1–14.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. 2017 LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA, December 2017, pp. 3149–3157.

Kennedy, J. & Eberhart, R. 1995 Particle swarm optimization. In: *Proceedings of ICNN'95-International Conference on Neural Networks*. IEEE, Vol. 4, pp. 1942–1948.

LaMoreaux, J. W., Wu, Q. & Wanfang, Z. 2014 New development in theory and practice in mine water control in China. *Carbonate Evaporite* **29** (2), 141–145.

Li, P. 2018 Mine water problems and solutions in China. *Mine Water Environ.* **37** (2), 217–221.

Ma, D., Duan, H., Cai, X., Li, Z., Li, Q. & Zhang, Q. 2018 A global optimization-based method for the prediction of water inrush hazard from mining floor. *Water* **10** (11), 1618.

Panagopoulos, G. P., Angelopoulou, D., Tzirtzilakis, E. E. & Giannoulopoulos, P. 2016 The contribution of cluster and discriminant analysis to the classification of complex aquifer systems. *Environ. Monit. Assess.* **188** (10), 1–13.

Saghebian, S. M., Sattari, M. T., Mirabbasi, R. & Pal, M. 2014 Ground water quality classification by decision tree method in Ardebil region, Iran. *Arabian J. Geosci.* **7** (11), 4767–4777.

Singh, R., Venkatesh, A. S., Syed, T. H., Surinaidu, L., Pasupuleti, S., Rai, S. P. & Kumar, M. 2018 Stable isotope systematics and geochemical signatures constraining groundwater hydraulics in the mining environment of the Korba Coalfield, Central India. *Environ. Earth Sci.* **77** (15), 1–17.

Tiantian, W., Dewu, J., Jian, Y., Ji, L. & Qiangmin, W. 2019 Assessing mine water quality using a hierarchy fuzzy variable sets method: a case study in the Guojiawan mining area, Shaanxi Province, China. *Environ. Earth Sci.* **78** (8), 1–13.

Wang, Y., Shi, L., Wang, M. & Liu, T. 2020 Hydrochemical analysis and discrimination of mine water source of the Jiaojia gold mine area, China. *Environ. Earth Sci.* **79** (6), 1–14.

Wang, Y., Shi, L. & Wang, M. 2021 A new evaluation model for discrimination of mine water quality based on EWM, fuzzy, fisher, and DS evidence theory – a case study in the Jiaojia gold mine area, China. *Arabian J. Geosci.* **14** (19), 1–13.

Wu, J., Xu, S., Zhou, R. & Qin, Y. 2016 Scenario analysis of mine water inrush hazard using Bayesian networks. *Saf. Sci.* **89**, 231–239.

Wu, Q., Mu, W., Xing, Y., Qian, C., Shen, J., Wang, Y. & Zhao, D. 2019 Source discrimination of mine water inrush using multiple methods: a case study from the Beiyangzhuang Mine, Northern China. *Bull. Eng. Geol. Environ.* **78** (1), 469–482.

Yan, P., Shang, S., Zhang, C., Yin, N., Zhang, X., Yang, G., Zhang, Z. & Sun, Q. 2021 Research on the processing of coal mine water source data by optimizing BP neural network algorithm with sparrow search algorithm. *IEEE Access* **9**, 108718–108730.

Yang, Y., Yue, J., Li, J. & Yang, Z. 2018 Mine water inrush sources online discrimination model using fluorescence spectrum and CNN. *IEEE Access* **6**, 47828–47835.

Yang, J., Dong, S., Wang, H., Li, G., Wang, T. & Wang, Q. 2021 Mine water source discrimination based on hydrogeochemical characteristics in the northern Ordos Basin, China. *Mine Water Environ.* **40** (2), 433–441.

Zhang, H., Xing, H., Yao, D., Liu, L., Xue, D. & Guo, F. 2019 The multiple logistic regression recognition model for mine water inrush source based on cluster analysis. *Environ. Earth Sci.* **78** (20), 1–15.

First received 26 February 2022; accepted in revised form 2 September 2022. Available online 9 September 2022