WEP Warwick Evans Publishing

International Journal of Natural Resources and Environmental Studies

ISSN: 3006-2012 (Print), ISSN: 3006-0834 (Online) | Volume 8, Number 1, Year 2026
DOI: https://doi.org/10.62051/ijnres.v8n1.04
Journal homepage: https://wepub.org/index.php/IJNRES/index

Prediction of Coal Seam Floor Water Inrush Based on DBO-XGBoost Under Small-Sample Data

Xiaodong Li*, Shixin Huang

College of Resources and Environment, Henan Polytechnic University University, Henan Province, China

*Corresponding Author: Xiaodong Li

ABSTRACT

Coal seam floor water inrush is a major geological hazard restricting the safe production of coal mines, directly threatening personnel life and engineering property safety. Accurate prediction of its occurrence risk is of great engineering significance. To address the problems of insufficient generalization ability, easy missed judgments and misjudgments of traditional prediction models caused by scarce water inrush samples and unbalanced data categories in actual mining, this study proposes a prediction model (DBO-XGBoost) integrating the improved SMOTE algorithm and Dung Beetle Optimizer (DBO) with eXtreme Gradient Boosting (XGBoost). A total of 50 sets of water inrush case data from Ordovician limestone nationwide were collected, and 6 core characteristic indicators including water pressure and aquiclude thickness were selected. The improved adaptive SMOTE algorithm was used to balance the data categories, and the 8:2 training-test set split ratio was determined through ten-fold five-cross validation. The global optimization ability of DBO simulating the natural behavior of dung beetles was utilized to optimize the key hyperparameters of XGBoost, fully excavating the nonlinear coupling relationships among features. Comparative verification with 7 models such as XGBoost and PSO-XGBoost showed that the accuracy, precision, recall, and F1-score of the proposed model reached 0.88, 0.89, 0.88, and 0.87 respectively, with an AUC value of 0.928. Compared with the traditional XGBoost, the true positive rate increased by 13.3% and the false negative rate decreased by 40%. The model was applied to the first mining area of Dongda Coal Mine to realize the visual evaluation of water inrush risk. It exhibits excellent accuracy and stability under small sample and complex geological scenarios, providing reliable technical support for the prevention and control of coal mine water inrush disasters.

KEYWORDS

Coal seam floor water inrush; Dung Beetle Optimizer; eXtreme Gradient Boosting; Water inrush prediction

1. INTRODUCTION

As a pillar industry of basic energy and industrial raw materials in China, coal plays a crucial role in the progress of the national economy and society, and its importance cannot be ignored [1]. According to the latest statistical data released by the National Bureau of Statistics, as of December 2024, the national raw coal output reached 4.76 billion tons, an increase of about 1.3% compared with the previous year. This growth trend indicates that although China is actively promoting clean energy and sustainable development, coal remains the core basic energy to ensure national energy security [2-3]. In China's coal mining, many mining areas contain Taiyuan Group and Ordovician limestone. More than 60% of coal mines are threatened by the confined aquifer of Ordovician limestone at the floor to varying degrees. Especially in deep coal mining, due to the complex geological environment

and water inrush mechanism [4], the prediction of coal seam floor water inrush has become increasingly difficult [5-6]. The resulting water hazards have seriously restricted the intensive production efficiency of mines [7] and threatened the personal safety of miners. Therefore, it is crucial to timely and accurately predict floor water inrush, so as to take corresponding protective measures to reduce water inrush risks and ensure the personal safety of every miner.

In the field of water inrush prediction and risk assessment, scholars at home and abroad have carried out multi-dimensional research, forming various prediction methods from empirical formulas to intelligent algorithms. The early water inrush coefficient method based on geomechanics [8], although intuitive in principle, relies on manual experience and is difficult to quantify the coupling effect of multiple factors; numerical simulation methods [9-11] can restore the distribution of floor stress and seepage field, but have a long modeling cycle and high sensitivity to parameters, making it difficult to adapt to dynamic mining scenarios. In recent years, risk assessment models have shown a development trend of interdisciplinary innovation. Machine learning methods have been widely used in floor water inrush risk prediction [12-14], and have become the mainstream technical direction of water inrush prediction due to their nonlinear fitting ability. In addition, methods such as LSTM, SMOTE, and PCA are widely used in the construction of floor water inrush models to improve the optimal training and prediction effects of the models [15-17]. The RF-VIKOR-GIS framework proposed by Liu et al. [18] combined with the six-factor index system of Yangcheng Coal Mine significantly improved the spatial characterization ability of regional risk assessment through GIS grid analysis and heatmap visualization; Dong Lili et al. [19] proposed a water inrush prediction model based on LSTM neural network, which can capture the long-term dependence in time series data and has a good performance in floor water inrush prediction; Shi Longqing et al. [20] proposed a coal mine water inrush discrimination model based on PCA-PSO-ELM, which improved the accuracy of the model from the aspect of data processing; Yin Huiyong et al. [21] optimized the model using GA-BP neural network based on SSA optimization, expanded the search range and accuracy, and improved the applicability of the model. Although machine learning-based water inrush prediction models have been applied in practical engineering, current models still face many challenges: prediction bias caused by scarce water inrush sample data and unbalanced data categories; safety accidents caused by missed water inrush predictions of the model, etc.

To address the above technical problems, this study proposes a coal seam floor water inrush prediction model (DBO-XGBoost) integrating improved SMOTE data balancing technology and Dung Beetle Optimizer (DBO) with XGBoost. Firstly, the improved adaptive SMOTE is used to balance small sample data, and then the DBO algorithm is introduced. By simulating the five typical natural behaviors of dung beetles—"ball-rolling, dancing, foraging, stealing, and breeding", a collaborative optimization mechanism of "global exploration-local development-diversity maintenance" is constructed to realize the adaptive optimization of XGBoost hyperparameters, thereby achieving accurate prediction of water inrush. Finally, combined with ArcGIS spatial analysis technology, a regional water inrush hazard assessment map is constructed to realize the spatial visualization of coal seam floor water inrush hazard. This method provides theoretical support and technical path for the prevention and control of coal seam floor water inrush disasters under small sample data conditions.

2. DATA

2.1. Data Sources and Characteristics

By referring to relevant literature, the water inrush cases of Ordovician limestone nationwide were counted, and the original data of some mining areas in China were sorted out [22]. The original data were collected at different times and locations without classification and grouping, totaling 50 sets, including 20 sets of water inrush and 30 sets of non-water inrush. Combined with the actual situation

of the study area, this study selected 6 characteristic indicators: water pressure (x1), aquiclude thickness (x2), fault throw (x3), distance from fault to working face (x4), mining height (x5), and coal seam dip angle (x6). By establishing the nonlinear mapping relationship among these 6 characteristic indicators, the water inrush disaster under complex geological conditions is accurately characterized.

Using the known data information, a multivariate distribution histogram was drawn, as shown in Fig. 1, which shows the distribution of each indicator under water inrush and non-water inrush conditions (1 represents water inrush and 0 represents non-water inrush). It can be seen from the figure that the water pressure under both water inrush and non-water inrush conditions is mostly distributed between 1-2 MPa, and there are some cases of high water pressure but no water inrush, indicating that water inrush is affected by multiple factors, similar to the coal seam dip angle; compared with these two indicators, the mining height and fault throw under non-water inrush conditions are mostly distributed in the lower values on the left, and the aquiclude thickness and distance from fault to working face under water inrush conditions are mostly distributed in the lower parts on the left. In summary, each indicator has a certain relationship with whether water inrush occurs, but a single indicator cannot reflect this relationship, and only the combined action of multiple indicators can characterize the water inrush conditions under different geological conditions. To show the correlation between different indicators, a correlation heatmap of different indicators was drawn, as shown in Fig. 2 Except that some parameters (water pressure and aquiclude thickness, fault throw and distance from fault to working face) are moderately correlated, the correlation between most other parameters is weak, showing a nonlinear coupling characteristic.

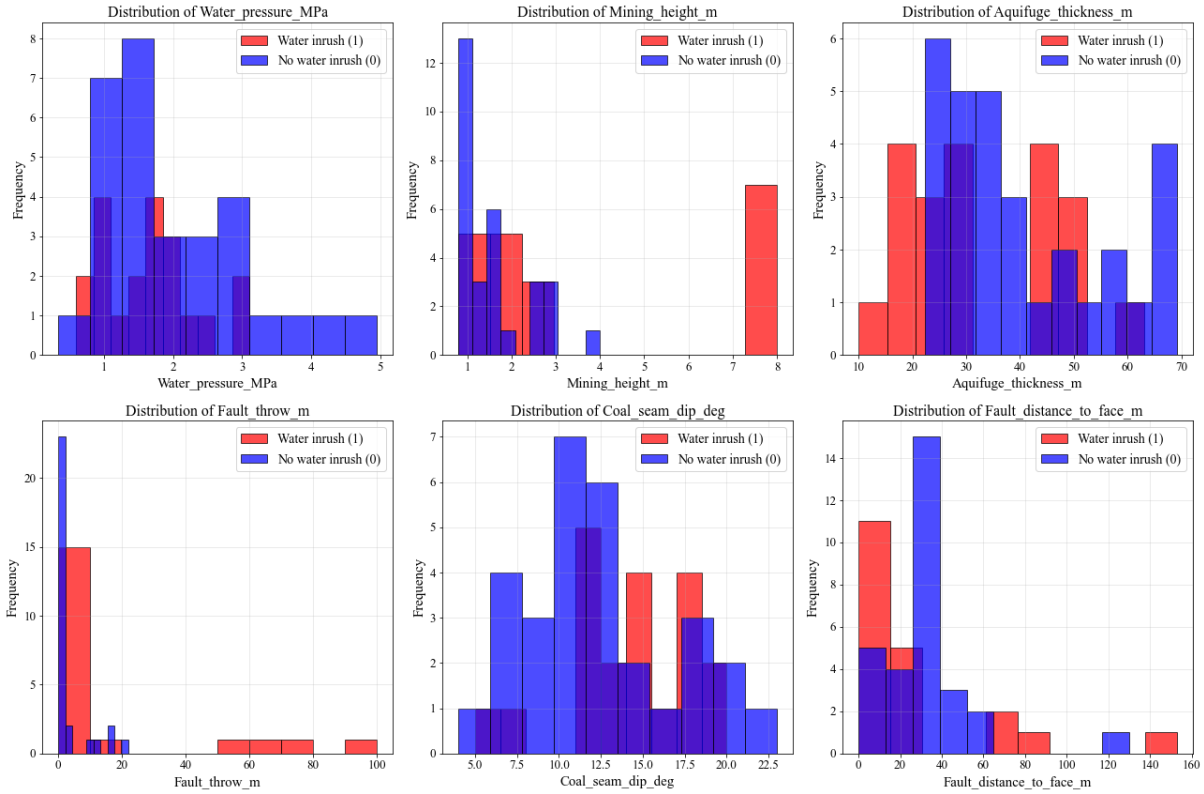

Fig. 1 Multivariate distribution histogram

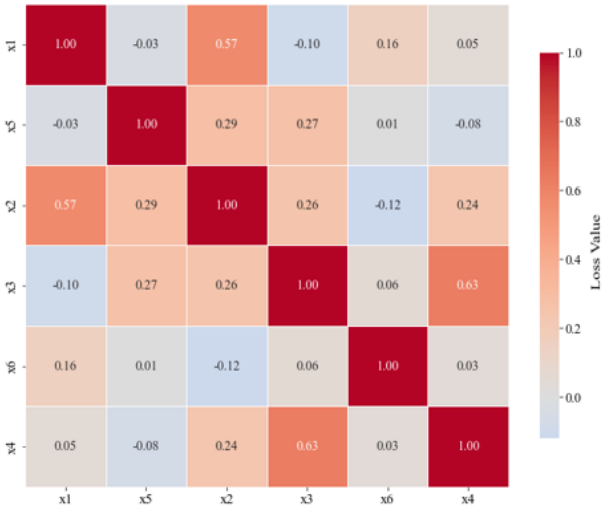

Fig. 2 Correlation heatmap of input variables

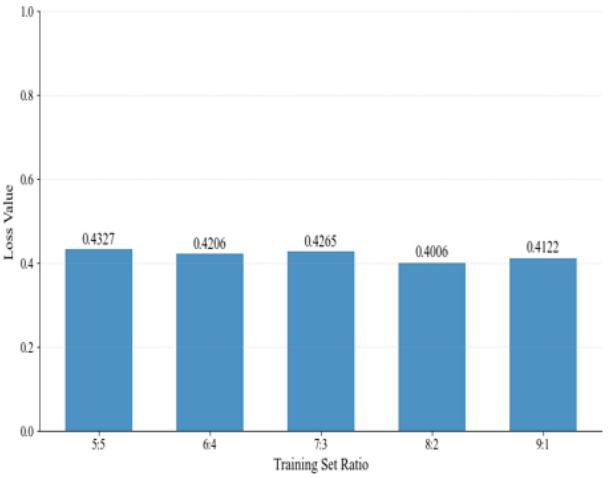

Fig. 3 Cross-entropy loss values under different training set ratios

2.2. Data Splitting

In model training, reasonable allocation of training data is crucial to model performance and generalization ability [23]. In this study, through ten-fold five-cross validation, different ratio configurations such as 5:5, 6:4, 7:3, 8:2, and 9:1 were tested, and the cross-entropy loss was monitored, as shown in Fig. 3 The results show that when the training set accounts for 80%, the model loss reaches the minimum value of 0.4, while the loss values of other ratios are around 0.42. Therefore, this study adopts an 8:2 ratio, selecting 40 sets of data as the training set and 10 sets as the test set.

2.3. SMOTE Oversampling of the Training Set

SMOTE is an intelligent oversampling technology used to solve the problem of class imbalance. Unlike the traditional simple replication of minority class samples, SMOTE balances the data set by generating synthetic samples. Aiming at the problem of extreme class imbalance in small sample data sets of coal mine water inrush, this study adopts a stratified dynamic k-value method to improve the traditional SMOTE. The traditional SMOTE generates synthetic samples through k-nearest neighbor linear interpolation, but its fixed k-value is prone to introducing noise in extreme imbalance scenarios [24]. The improved method performs stratified processing on the number of minority class samples, and can select an appropriate k-value by identifying the number of minority class samples, which to a certain extent retains the distribution characteristics of the original data and avoids certain noise risks.

3. RESEARCH METHODS

3.1. Principle of the XGBoost Algorithm

XGBoost is an optimized algorithm of the gradient boosting framework proposed by Chen [25] et al. It forms a strong prediction model by serially iteratively training CART regression trees and minimizing the loss function with a gradient descent strategy. Compared with the traditional GBDT, it introduces second-order derivative optimization, multi-dimensional regularization, and parallel design, which significantly improves the fitting accuracy, convergence efficiency, and generalization ability. It can accurately capture the core contradiction of small samples and nonlinear geological feature coupling in coal seam floor water inrush prediction. It can not only excavate the complex mapping relationship between 6 indicators such as water pressure and aquiclude thickness and water inrush risk, but also suppress overfitting in small samples through regularization.

The theoretical framework of XGBoost revolves around Boosting integration logic, objective function construction, and Taylor expansion optimization, with the core of improving model performance through iterative error correction.

3.1.1. Integrated Learning Logic

Taking the CART regression tree as the weak learner, the prediction value of the model for sample x_i in the t-th round is the sum of the prediction results of the first t trees:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{1}$$

where $\hat{y}_i^{(t-1)}$ is the prediction value of the (t−1)-th round, $f_t(x_i)$ is the output of the new tree (leaf node weight), and F is the function space of all possible CART trees (q(x) is the tree structure function that maps x to leaf node j, w_j is the node weight, and T is the number of leaves). Each new tree focuses on fitting the residual gradient of the previous model rather than directly fitting the difference, ensuring the accurate decline of the loss function.

3.1.2. Objective Function Construction

The objective function includes a loss function (measuring prediction deviation) and a regularization term (controlling tree complexity), which is suitable for the binary classification task of water inrush prediction:

$$L(t) = \Sigma_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Sigma_{k=1}^{t} \Omega(f_k) \tag{2}$$

Loss function: Cross-entropy loss (LogLoss) is adopted, and the prediction value is mapped to the water inrush probability through Sigmoid. The formula is:

$$l\left(y_i, \hat{y}_i^{(t)}\right) = -y_i log\sigma\left(\hat{y}_i^{(t)}\right) - (1 - y_i) log\ (1 - \sigma\left(\hat{y}_i^{(t)}\right)) \tag{3}$$

Regularization term: The balance between fitting ability and overfitting risk is achieved by constraining the complexity of the tree structure. The formula is:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

where γ is the leaf node penalty coefficient (controlling the number of leaf nodes), λ is the weight decay coefficient (suppressing excessive leaf node weights), T is the number of leaf nodes of the current tree, and w_j is the weight of the j-th leaf node.

3.1.3. Taylor Expansion Optimization

To efficiently solve the objective function, XGBoost approximates the loss function through second-order Taylor expansion. Substitute $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$ into the loss function, take $y_i^{(t-1)}$ as the expansion point, and ignore the constant term. The objective function can be simplified as:

$$L(t) \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{5}$$

where $g_i = \frac{\partial l(y_i, z)}{\partial z}\Big|_{z=y_i^{(t-1)}}$ is the first-order derivative of the loss function, and $h_i = \frac{\partial^2 l(y_i, z)}{\partial z^2}\Big|_{z=y_i^{(t-1)}}$ is the second-order derivative. This approximation converts the complex loss function into a quadratic function about $f_t(x_i)$, which greatly reduces the difficulty of solving.

3.1.4. Tree Structure and Weight Optimization

For a fixed tree structure q(x), substitute $f_t(x_i) = w_{q(x_i)}$ into the simplified objective function, and the optimal weight of the leaf node can be derived:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i+\lambda} \tag{6}$$

where I_j is the set of samples mapped to the j-th leaf node. At the same time, define the scoring function of the tree structure $L(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i+\lambda} + \gamma T$ to measure the quality of the tree structure. The model traverses features and split thresholds through a greedy strategy, selects the split method that maximizes the gain of the scoring function to build the tree, and automatically realizes tree pruning through the regularization term to further improve generalization ability.

3.2. Principle of the DBO Algorithm

The Dung Beetle Optimizer (DBO) is a new meta-heuristic algorithm proposed by Jiankai Xue et al. in 2023 [26]. Its core logic is derived from the bionic simulation of the ecological behavior of dung beetles in nature. By replicating the five typical behaviors formed by this species in the process of survival and reproduction—"ball-rolling, dancing, foraging, stealing, and breeding", an optimization framework with both global exploration ability and local development accuracy is constructed to provide an efficient solution for complex optimization problems. It is especially suitable for model hyperparameter tuning tasks under small sample and nonlinear scenarios. The core advantage of DBO lies in realizing the dynamic balance of exploration-development-diversity maintenance through multi-behavior collaborative simulation. The internal mapping relationship between its five bionic behaviors and the optimization algorithm is as follows:

3.2.1. Ball-rolling Behavior

In nature, dung beetles roll feces into balls and transport them in a specific direction to reserve food and breeding grounds for larvae. This process is guided by environmental signals such as the sun's position, and has obvious goal orientation and large-scale movement characteristics. In the algorithm, the ball-rolling behavior corresponds to the global search stage, aiming to expand the coverage of the solution space and avoid falling into local optima. At this time, each dung beetle individual takes the current global optimal solution as the sun's position, and combines the guidance of the average population position to achieve large-scale optimization through random direction adjustment. The position update formula is defined as:

$$X_i(t+1) = X_i(t) + \alpha \cdot sin\theta \cdot \left|X_i(t) - X_{gbest}(t)\right| + \beta \cdot cos\theta \cdot |X_i(t) - X_{avg}(t)| \tag{7}$$

where t is the current number of iterations; $X_i(t)$ is the current position of the i-th dung beetle individual; $X_{gbest}(t)$ is the global optimal position at iteration t; X_{avg} is the average population position; α and β are global search step size factors (usually α=2, β=2) used to control the exploration range; θ is a random angle (θ∈[0,2π]) that simulates the randomness of the direction of dung beetle ball-rolling to ensure the comprehensiveness of solution space exploration.

3.2.2. Dancing Behavior

Dung beetles will adjust their traveling direction through short-term dancing during ball-rolling to avoid obstacles or correct the transport path. This behavior is a dynamic calibration of the global exploration direction. In the algorithm, the dancing behavior is realized by introducing an adaptive direction factor. When the individual position update does not bring an improvement in fitness, the direction adjustment mechanism is triggered, and the θ value is dynamically corrected based on the relative distance between the current position and the global optimal solution. The formula is as follows:

$$\theta = \theta_0 + \gamma \cdot \frac{|X_i(t) - X_{gbest}(t)|}{max\left(\left|X_j(t) - X_{gbest}(t)\right|\right)} \cdot \pi \tag{8}$$

where θ_0 is the initial random angle; γ is the direction adjustment coefficient (γ=0.5); $max\left(\left|X_j(t) - X_{gbest}(t)\right|\right)$ is the maximum distance between individuals in the population and the global optimal solution. Normalization ensures the rationality of direction adjustment. This mechanism effectively improves the accuracy of global exploration and reduces invalid search costs.

3.2.3. Foraging Behavior

Some dung beetles abandon ball-rolling and instead randomly search for scattered feces resources. This behavior is random without a fixed direction and is an important supplement to maintaining population survival. In the algorithm, the foraging behavior corresponds to the diversity maintenance strategy. By randomly resetting the positions of some individuals, premature convergence of the population is avoided, which is especially suitable for the problem of scarce solution space information in small sample optimization scenarios. Its position update adopts a random sampling method:

$$X_i(t+1) = X_{min} + randa(0,1) \cdot (X_{max} - X_{min}) \tag{9}$$

Where Xmin and Xmax are the lower and upper bounds of the optimization variables, respectively; rand(0, 1) is a random number uniformly distributed within the interval [0, 1]. The algorithm controls the trigger ratio of this behavior via a preset foraging probability q (usually set to 0.3), ensuring exploration efficiency while maintaining population diversity.

3.2.4. Stealing Behavior

Individual dung beetles may steal feces balls from other individuals, which is a competitive behavior in the population. In the algorithm, the stealing behavior is designed to enhance the interaction between individuals. When the fitness of an individual is lower than the average fitness of the population, it will "steal" the position information of high-quality individuals to update its own position, thereby improving the convergence speed of the population. The position update formula is:

$$X_i(t+1) = X_k(t) + \delta \cdot randa(0,1) \cdot |X_i(t) - X_k(t)| \tag{10}$$

where $X_k(t)$ is the position of a randomly selected high-quality individual; δ is the stealing step size factor (δ=0.8); randn(0,1) is a random number following a standard normal distribution. This mechanism accelerates the diffusion of high-quality solution information in the population, and eliminates inferior solutions through competitive pressure to improve the overall optimization efficiency.

3.2.5. Breeding Behavior

Dung beetles bury feces balls underground and lay eggs, and larvae develop and grow in the feces balls. This process focuses on the in-depth utilization of high-quality resources. In the algorithm, the breeding behavior corresponds to the local search stage. When the individual position is close to the global optimal solution, local refined mining is triggered to perform small-scale search around the current high-quality solution to improve the accuracy of the solution. The position update formula is:

$$X_i(t+1) = X_{gbest}(t) + \epsilon \cdot rand(0,1) \cdot X_i(t) \tag{11}$$

where ϵ is the local development step size factor (ε=0.5), and fine exploration of the area around the high-quality solution is achieved through small step size control. This behavior complements the ball-rolling behavior: the former ensures the breadth of exploration, and the latter guarantees the accuracy of development, jointly realizing efficient traversal of the solution space.

3.3. DBO-XGBoost Hybrid Model

In this model, XGBoost is responsible for establishing the complex mapping relationship between 6 geological features such as water pressure (x1) and aquiclude thickness (x2) and water inrush risk (y $\in\{0,1\}$). Its tree integration structure can effectively capture the nonlinear coupling law between

features by iteratively fitting the negative gradient residual, and suppress overfitting in small sample data through multi-dimensional regularization. However, the performance of XGBoost is highly dependent on the hyperparameter combination: the learning rate (η) controls the iteration step size—too small a value leads to slow model convergence, and too large a value is prone to underfitting; the number of trees (nestimators) determines the model complexity—too few trees cannot excavate the deep laws of data, and too many trees exacerbate overfitting; the regularization parameters (λ, γ) balance the fitting accuracy and generalization ability. Improper parameter configuration will directly lead to performance degradation of the model in small sample scenarios. Experiments have verified that these parameters are the core influencing factors of water inrush prediction performance.

DBO regards the key hyperparameter combination of XGBoost as an individual in the optimization solution space. By simulating the five natural behaviors of dung beetles—ball-rolling, dancing, foraging, stealing, and breeding—it dynamically adjusts the intensity of exploration and development to search for the hyperparameter set that optimizes the prediction performance of XGBoost. The DBO-XGBoost model is constructed according to the above research methods, and the specific collaborative solution flowchart is shown in the Fig. 4:

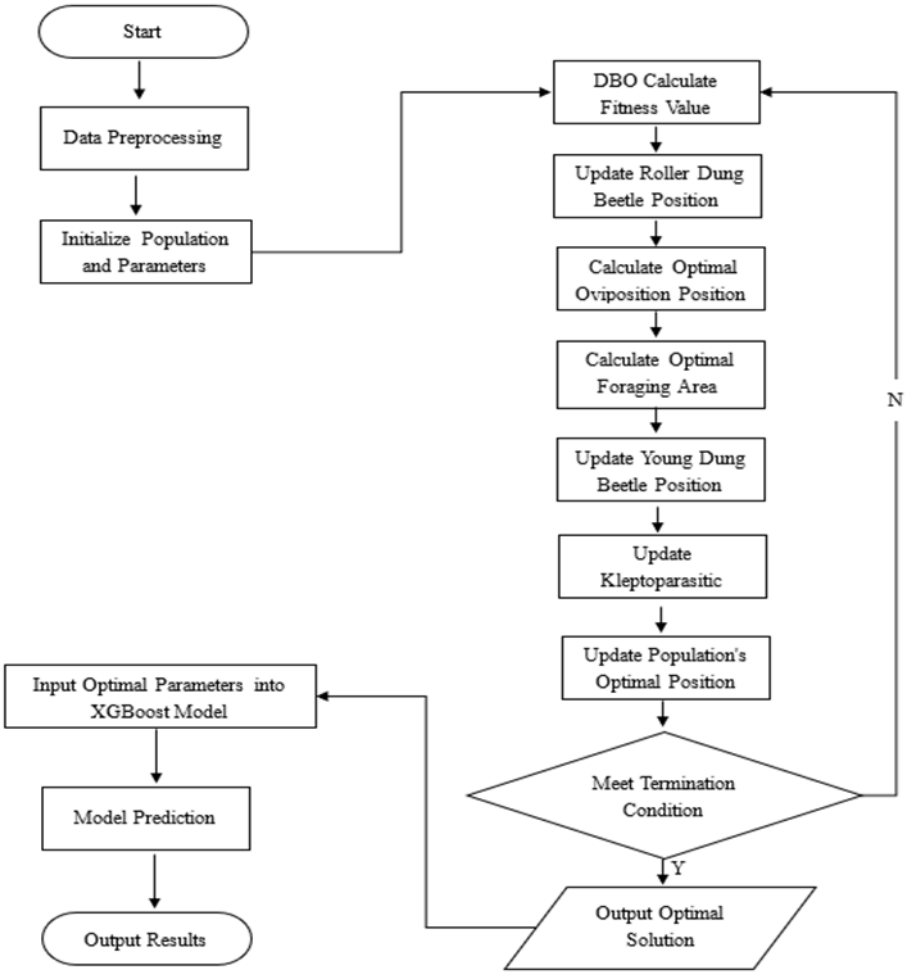

Fig. 4 Flowchart of DBO-XGBoost

4. ANALYSIS OF MODEL EVALUATION RESULTS

This study conducted two groups of comparative analyses of evaluation results: one is the comparative analysis of model evaluation results before and after SMOTE oversampling; the other is

the comparative analysis of the established DBO-XGBoost model with 7 models including PSO-XGBoost, XGBoost, DBO-SVM, SVM, DBO-GBDT, GBDT, DBO-RF, and RF. Confusion matrix, ROC curve, and performance radar chart were used to carry out evaluation result analysis from three aspects: classification results, discrimination ability, and multi-dimensional balance.

4.1. Analysis of Evaluation Results of SMOTE Oversampling

As shown in Fig. 5, the confusion matrix is used to quantitatively analyze the model classification performance[27]. The comparison between the confusion matrix of the DBO-XGBoost (SMOTE) model and the original model shows that the number of correctly classified samples of the DBO-XGBoost model processed by SMOTE is significantly higher. Among them, the number of true positives (TP=17) is higher than that of the original model (TP=16), and the number of false negatives (FN=3) is lower than that of the original model (FN=4). The performance radar chart in Fig. 6 shows that the accuracy, precision, recall, and F1-score of DBO-XGBoost processed by SMOTE are significantly higher than those of the original model, reaching 0.8800, 0.8937, 0.8800, and 0.8791 respectively. Fig. 7 is a comparison of ROC curves between the DBO-XGBoost model processed by SMOTE and the original model. It can be seen from the figure that under the same false positive rate, the true positive rate of the DBO-XGBoost model processed by SMOTE is mostly higher than that of the original model, and its AUC value is as high as 0.928, which is higher than 0.915 of the original model. Experiments show that after the improved SMOTE performs convex combination interpolation of the k-nearest neighbor topological structure (k=5) of minority class samples by Euclidean distance measurement, the ratio of positive to negative samples is optimized from 3:2 to 1:1, which effectively alleviates the model bias caused by data scarcity and class imbalance; training the DBO-XGBoost model with SMOTE-processed data for water inrush prediction improves the evaluation indicators such as accuracy, recall, and F1-score to a certain extent, and effectively increases the correct recognition rate of the model for water inrush samples and reduces the risk of missed water inrush predictions.

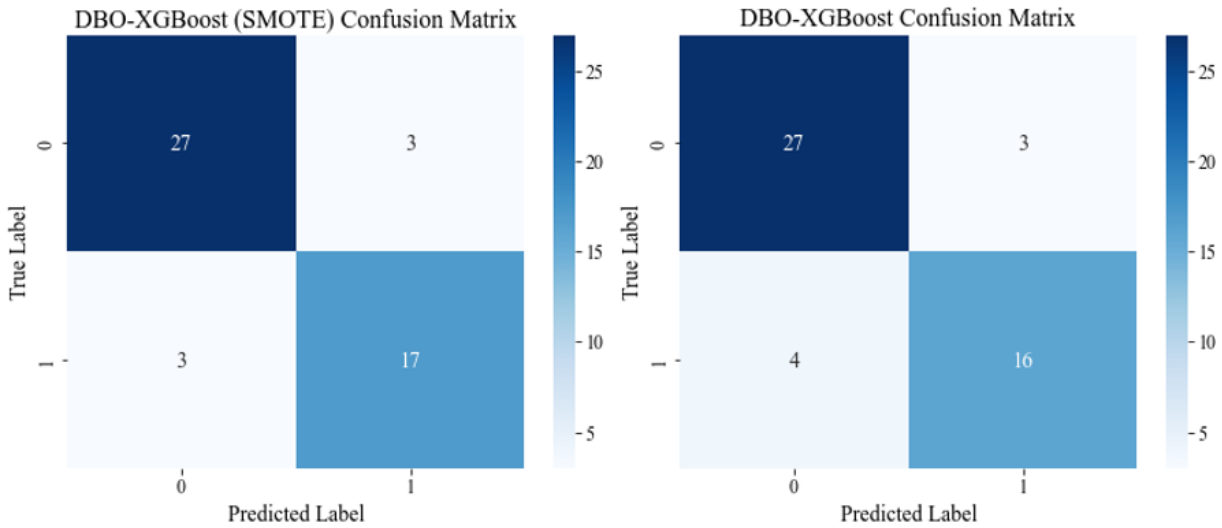

Fig. 5 Comparison of confusion matrices

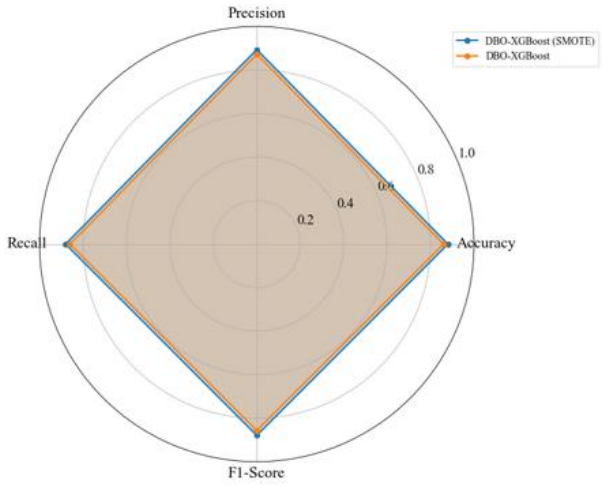

Fig. 6 Multi-indicator radar chart

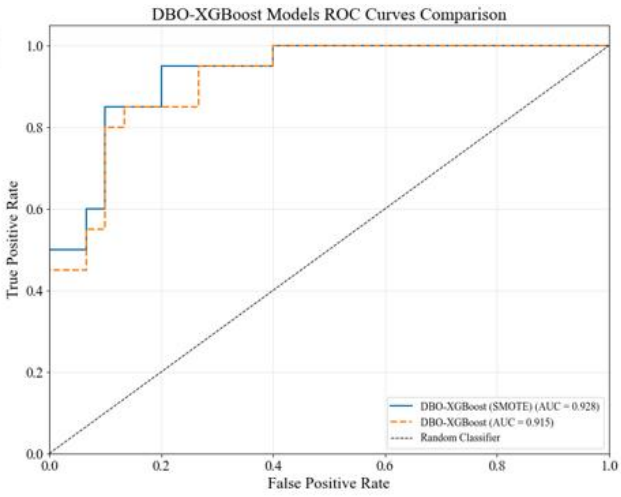

Fig. 7 ROC curve comparison chart

4.2. 4.2 Analysis of Evaluation Results of the DBO-XGBoost Model

As shown in Fig. 8, the confusion matrix is used to quantitatively analyze the classification performance of each model. The DBO-XGBoost model has 27 true negatives (TN) and only 3 false positives (FP) in the non-water inrush category; in the water inrush category, the true positives (TP) reach 17 and the false negatives (FN) are only 3. Compared with the traditional XGBoost (TN=26, TP=15, FN=5), the true positive rate of DBO-XGBoost is increased by 13.3% and the false negative rate is reduced by 40%; compared with the PSO-optimized XGBoost (TP=16, FN=4), its true positive rate is increased by 6.25% and the false negative rate is reduced by 25%, indicating that the optimization effect of DBO on XGBoost hyperparameters is more significant. In addition, by comparing the fusion effects of DBO with different base models (such as DBO-SVM, DBO-GBDT, DBO-RF), it can be seen that the DBO-XGBoost model is superior to other DBO-optimized models in terms of true negative and true positive indicators, and the fluctuations of false negative and false positive are smaller, which proves that the tree integration characteristics of XGBoost have a higher matching degree with the optimization mechanism of DBO. The combination of the two can give full play to their respective advantages and show more stable classification performance in water inrush prediction tasks with nonlinear and complex samples. In addition, in coal seam floor water inrush prediction, the control of false negatives (actual water inrush but predicted as non-water inrush) is directly related to engineering safety. The false negative of the DBO-XGBoost model is only 3, which is significantly lower than that of comparative models such as SVM (FN=7), GBDT (FN=7), and DBO-SVM (FN=5); at the same time, its true positive rate (water inrush recognition rate) reaches 17, which is higher than that of optimized models such as DBO-SVM (TP=15) and DBO-GBDT (TP=15), indicating that the DBO-XGBoost model can not only effectively reduce the risk of missed water inrush predictions but also accurately identify actual water inrush samples.

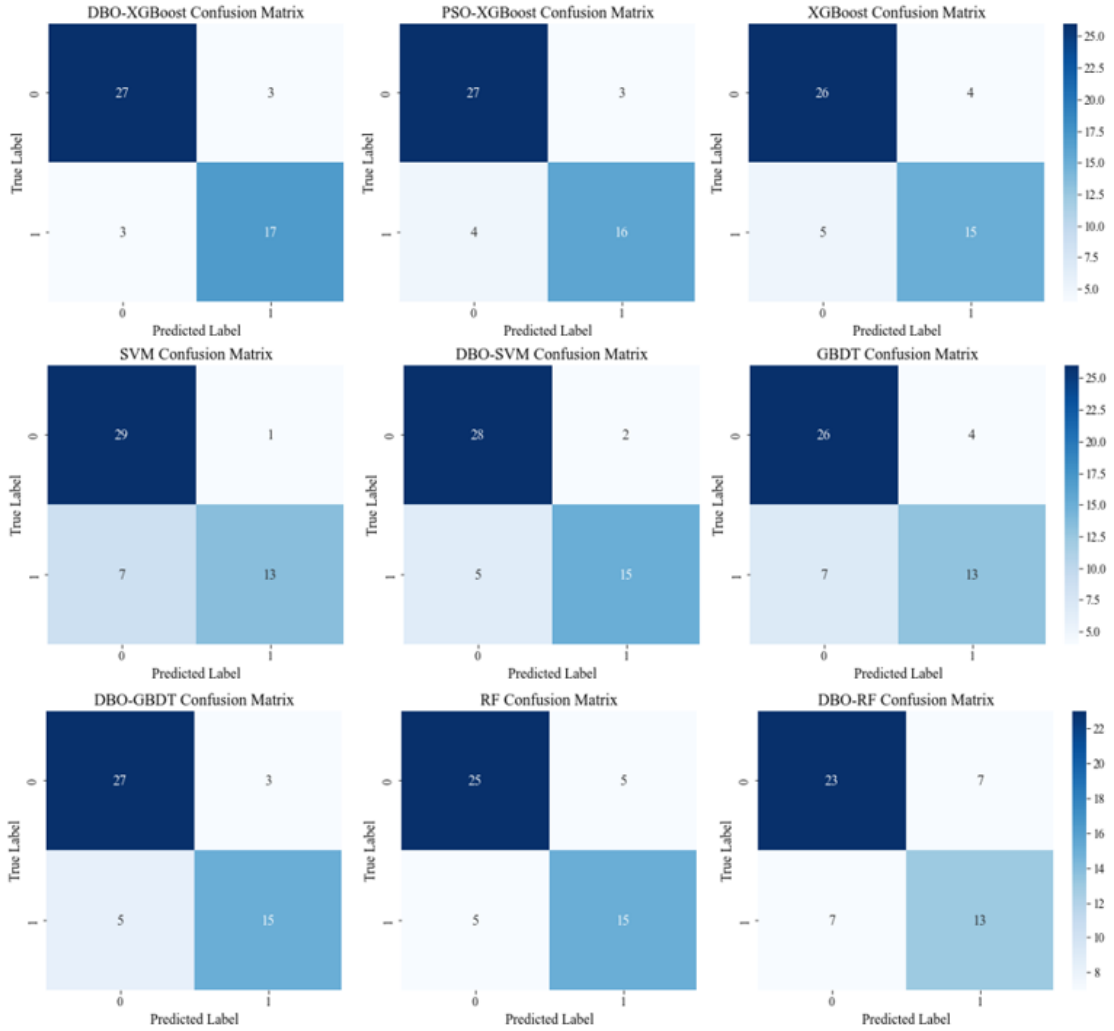

Fig. 8 Comparison of confusion matrices of multiple models

To further verify the class discrimination ability of the DBO-XGBoost model and other comparative models for the two types of samples, as shown in Fig. 9, the ROC curves of each model are drawn, with the false positive rate (misjudgment rate of non-water inrush samples) as the horizontal axis and the true positive rate (recognition rate of water inrush samples) as the vertical axis. The curve position and the Area Under the Curve (AUC) are the core indicators to measure the model's class discrimination ability. The ROC curve comparison shows that in the XGBoost model group, the ROC curve of DBO-XGBoost is generally above the curves of XGBoost and PSO-XGBoost, and its AUC (0.928) is higher than 0.912 of the latter two, meaning that under the same misjudgment rate of non-water inrush samples, the DBO-XGBoost has a better recognition rate of water inrush samples; in the GBDT and Random Forest (RF) model groups, the AUC (0.928) of DBO-XGBoost is higher than that of GBDT (0.872), DBO-GBDT (0.911), and DBO-RF (0.873) respectively, with a relative increase of 1.8%~5.7%, indicating that the optimization effect of DBO on XGBoost hyperparameters is significantly better than its optimization on base models such as GBDT and Random Forest; in the SVM model group, although the AUC (0.932) of DBO-SVM is slightly higher, the ROC curve of DBO-XGBoost has smaller fluctuations and stronger stability of classification discrimination. Experiments show that DBO-XGBoost not only achieves accurate classification at the individual sample level but also has better discrimination ability between water inrush and non-water inrush categories, which can simultaneously reduce the engineering risks of missed water inrush predictions and misjudged non-water inrush predictions, further confirming its performance advantage in coal seam floor water inrush prediction tasks.

The performance radar chart further intuitively presents the multi-indicator collaborative performance of each model. The comparison of the multi-dimensional performance of the models through the performance radar chart in Fig. 10 shows that the performance polygon corresponding to the DBO-XGBoost model is in the outermost area of all models, indicating that it has achieved collaborative optimization in accuracy (0.8800), precision (0.8937), recall (0.8800), and F1-score (0.8791). Compared with the precision (0.8728) and recall (0.8600) of the PSO-XGBoost model, the precision of the DBO-XGBoost model is increased by 2.4% and the recall by 2.3%; compared with the accuracy (0.8200) and F1-score (0.8259) of the original XGBoost model, its accuracy is increased by 7.3% and the F1-score by 6.4%, indicating that the optimization effect of DBO on XGBoost hyperparameters is significantly better than that of the traditional intelligent optimization algorithm PSO. Comparing the fusion effects of DBO with different base models (such as DBO-SVM, DBO-GBDT, DBO-RF), the performance polygon coverage of the DBO-XGBoost model is significantly larger than that of other DBO-optimized models: the F1-score of the DBO-SVM model is only 0.8555, and the accuracy of the DBO-GBDT model is 0.8400, both lower than that of DBO-XGBoost, which proves that the tree integration learning characteristics of XGBoost have a higher matching degree with the optimization mechanism of DBO, and can give full play to the nonlinear fitting ability of the base model and the hyperparameter optimization advantage of the optimization algorithm, showing more stable comprehensive performance in water inrush prediction tasks driven by complex geological data.

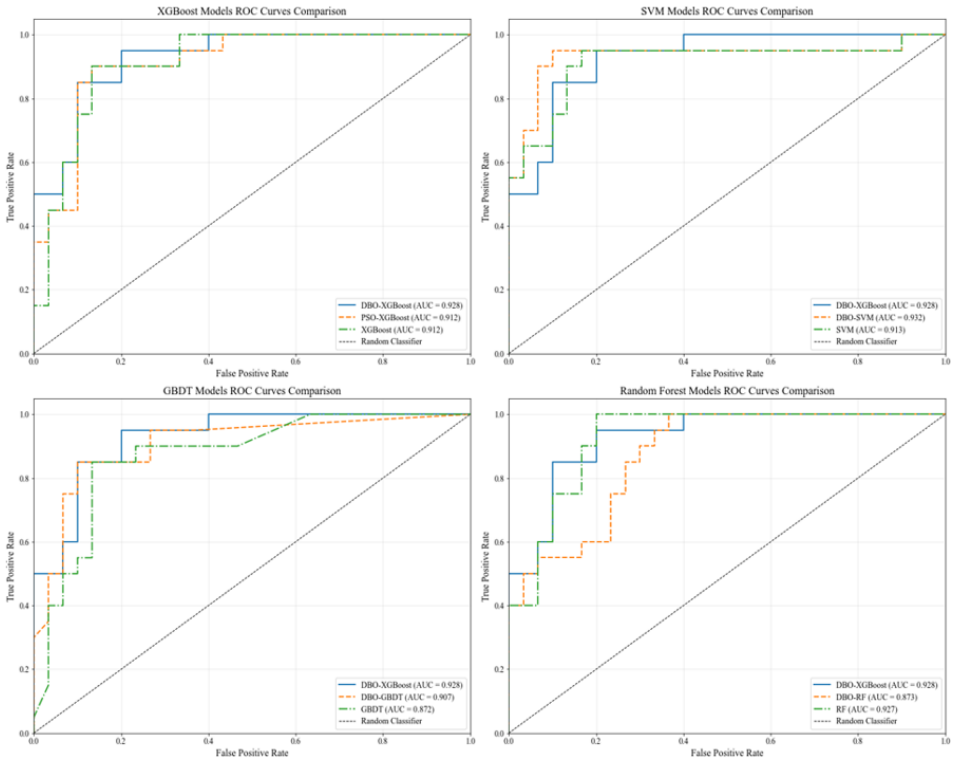

Fig. 9 Comparison of ROC curves of multiple models

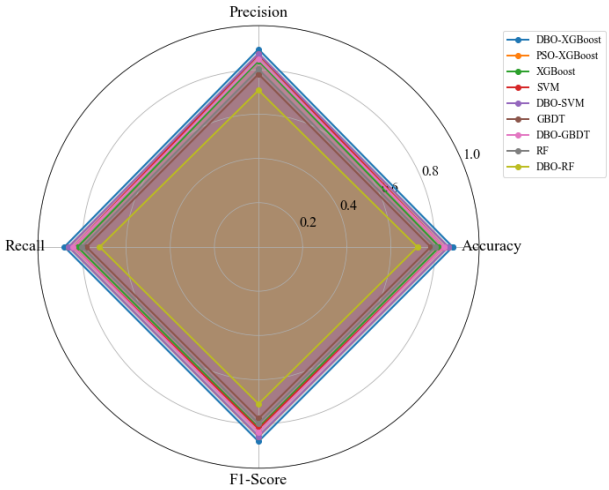

Fig. 10 Performance radar chart of multiple models

5. ENGINEERING APPLICATION OF THE DBO-XGBOOST MODEL

5.1. Overview of the Study Area

This study takes the first mining area of Dongda Coal Mine in Jincheng City, Shanxi Province as the research area for engineering application. Dongda Coal Mine is located in the southeastern part of Shanxi Province, about 36km northwest of Jincheng Urban Area, at the southern end of the Qinshui Composite Syncline Depression between the Taihang Mountain Composite Anticline Uplift and the Huoshan North-South Anticline Uplift. The first mining area mainly includes the east and west wings of the first panel and the west wing of the second panel, with minable coal seams including No. 3 and No. 15 coal seams. This study mainly focuses on the No. 3 coal seam. The average burial depth of the No. 3 coal seam is 650m, the average coal thickness is 5m, and the coal seam dip angle is 6°. The Ordovician limestone karst water level elevation in the mining area is +604~+616.69, and the overall water richness is weak. However, the No. 3 coal seam in the mining area faces the problem of mining under pressure, with potential water inrush hazards, and structural factors such as faults and folds are relatively developed, which bring great potential safety hazards to the safe production of the coal mine and have certain research value.

5.2. Water Inrush Probability Prediction and Visual Risk Assessment

The DBO-XGBoost model is used to calculate the water inrush probability. Taking the water pressure, aquiclude thickness, fault throw, distance from fault to working face, mining height, and coal seam dip angle at each drilling hole in the research area as input factors, the predicted value of water inrush probability is output through the DBO-XGBoost model. Combined with ArcGIS and using the natural breaks classification method, the research area is divided into high-risk areas (0.473-0.63), medium-risk areas (0.338-0.472), and low-risk areas (0.199-0.337), and the risk assessment map is finally output. As shown in Fig. 11, the water inrush probability values in the research area show continuous gradient and spatial differentiation characteristics. The high water inrush risk areas are concentrated in the northern part of the research area. This area has a small aquiclude thickness, relatively high water pressure (7.45-8.11Mpa), and surrounding fault zones, which are prone to the development of water-conducting fracture zones, resulting in a high water inrush probability. The medium-risk areas are mainly distributed in the central and southeastern parts of the research area. The central area is affected by high water pressure (7.10-7.88Mpa), and the water inrush probability is also relatively high; the southeastern part has a relatively large fault zone, including the Hangjiashan Fault and many

small faults. The throw of the Hangjiashan Fault reaches 25-90m, which is potentially dangerous, but the water pressure in this area (6.25-7.12Mpa) is relatively small, so the water inrush probability is relatively high. The low-risk areas are mainly distributed in the southwestern part. This area has fewer faults and a longer distance, and the water pressure (5.96-6.77Mpa) is also relatively low, so the water inrush probability is relatively small. The results show that the water inrush probability risk assessment map obtained by the DBO-XGBoost model is highly coupled with the actual geological conditions of the research area. The spatial correspondence between hydrogeological data and risk zoning provides visual evidence for the multi-parameter collaborative analysis framework proposed in the study, and provides a scientific basis for mine water hazard prevention and control.

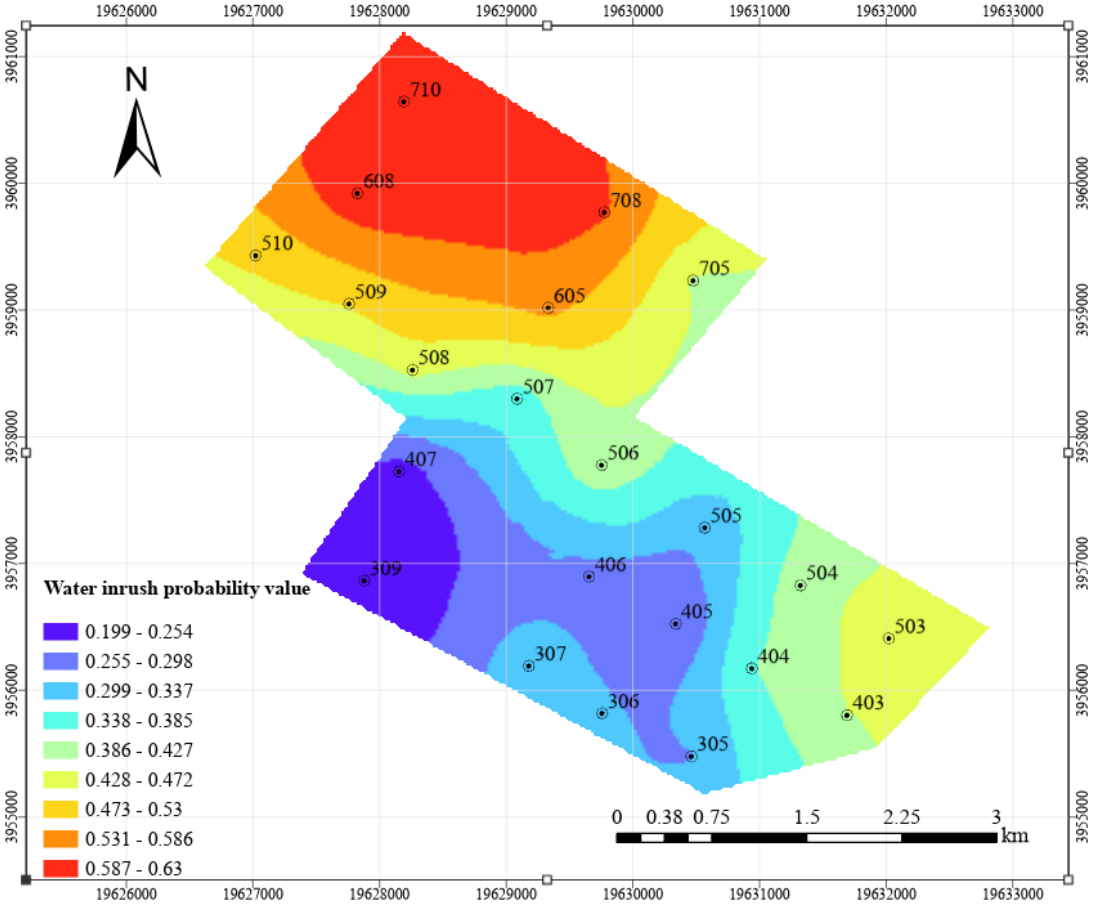

Fig. 11 DBO-XGBoost water inrush risk assessment map

6. CONCLUSIONS

By integrating cutting-edge machine learning technologies with geological engineering professional theories, this study constructs an integrated model of Dung Beetle Optimizer and eXtreme Gradient Boosting (DBO-XGBoost), solving the problem of water inrush from the Ordovician aquifer at the floor from a new perspective. The conclusions obtained in this study are as follows:

(1) Innovation in data processing methods: Aiming at the small sample data and unbalanced coal mine geological data, an adaptive SMOTE method with stratified dynamic k-value is proposed. Different k-values are used according to the number of different minority samples to effectively alleviate the problem of data class imbalance, providing a new method for machine learning modeling of complex geological data.

(2) Innovation in model architecture and performance optimization: Aiming at the modeling problem of multi-source nonlinear features in coal mine water inrush prediction under small sample data, the proposed DBO-XGBoost integrated model shows significant performance advantages. Empirical

results show that the comprehensive performance of the model is at the optimal level. Compared with benchmark models such as Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT), the AUC index of the model is improved by 0.1-5.6 percentage points, and it has stronger robustness and generalization ability. Its prediction accuracy (Accuracy=89%) and engineering applicability have reached the advanced level in the industry.

(3) Visual presentation of water inrush probability risk: For the water inrush probability value output by DBO-XGBoost, combined with ArcGIS and using the natural breaks classification method, a visual water inrush risk assessment map of the first mining area of Dongda Coal Mine is realized, providing a basis for water inrush prevention and control in Dongda Coal Mine.

REFERENCES

[1] XIE H P, WANG J H, WANG G F, et al. New concepts of coal revolution and prospects for coal science and technology development[J]. Journal of China Coal Society, 2018, 43(05): 1187-1197.

[2] WU B, WANG J X, QU B L, et al. Development, effectiveness, and deficiency of China's coal mine safety supervision system[J]. Resources Policy, 2023, 82: 103524.

[3] ZENG Y F, ZHU H C, WU Q, et al. Mechanism and prevention prospect of different types of coal seam floor water disaster in China[J]. Journal of China Coal Society, 2025, 50(02): 1073-1099.

[4] WANG J L. Common problems and solutions in coal mine water prevention and control work[J]. Energy and Energy Conservation, 2021, (03): 32-34. DOI:10.16643/j.cnki.14-1360/td.2021.03.015.

[5] SUN W B, XUE Y C, SHAO J L, et al. Application of grey correlation analysis in coal mine casualty accidents[J]. Coal Technology, 2019, 38(03): 178-181. DOI:10.13301/j.cnki.ct.2019.03.060.

[6] LIU X, YANG S, TAN Y, et al. An innovative test method for mechanical properties of sandstone under instantaneous unloading confining pressure[J]. International Journal of Mining Science and Technology, 2024, 34(12): 1677-1692.

[7] HU W Y, ZHAO C H. Evolution of water hazard control technology in China's coal mines[J]. Mine Water and the Environment, 2021, 40: 334-344.

[8] ZHANG G Y. Study on water inrush mechanism and regional restoration technology of ultra-thin aquiclude floor[D]. China Coal Research Institute, 2022.

[9] ZHOU M R, SONG H P, HU F, et al. Application of spectral clustering combined with LIF in identification of mine water inrush source types[J]. Spectroscopy and Spectral Analysis, 2021, 41(02): 435-440.

[10] WU C S, BAI Q H, WANG X Y. Prediction and risk assessment of mine water inrush based on fuzzy cluster analysis[J]. China Safety Science Journal, 1995, (S2): 79-83.

[11] HOU E K, XI H Q, WEN Q, et al. Prediction of water inflow in coal seam mining face under concealed burned area based on GMS[J]. Journal of Safety and Environment, 2022, 22(05): 2482-2492.

[12] ZHENG Q S, WANG C F, ZHU Z H. Research on the prediction of mine water inrush disasters based on multi-factor spatial game reconstruction[J]. Geomechanics and Geophysics for Geo-Energy and Geo-Resources, 2024, 10: 41.

[13] YIN H C, ZHANG G Z, WU Q, et al. Transfer learning with transformer-based models for mine water inrush prediction: A multivariate analysis using sparse and imbalanced monitoring data[J]. Mine Water and the Environment, 2024: 1-20.

[14] LI B, WU Q, YANG Y, et al. Characteristics of roof rock failure during coal seam mining and prediction techniques for mine water inflow in exposed karst areas[J]. Bulletin of Engineering Geology and the Environment, 2024, 83: 388.

[15] YIN H C, WU Q, YIN S X, et al. Predicting mine water inrush accidents based on water level anomalies of borehole groups using long short-term memory and isolation forest[J]. Journal of Hydrology, 2023, 616: 128813.

[16] ZHANG Y, TANG S F, SHI K, et al. An evaluation of the mine water inrush based on the deep learning of ISMOTE[J]. Natural Hazards, 2023, 117: 1475-1491.

[17] LI Q, LU C J, ZHAO H. Risk assessment of floor water inrush based on TOPSIS combined weighting model: A case study in a coal mine, China[J]. Earth Science Informatics, 2023, 16: 565-578.

[18] LIU W T, HAN M K. A practical method for floor water inrush prediction using a hybrid artificial intelligence model and GIS[J]. Mine Water and the Environment, 2023, 42: 220-229.

[19] DONG L L, FEI C, ZHANG X, et al. Coal mine water inrush prediction based on LSTM neural network[J]. Coal Geology & Exploration, 2019, 47(02): 137-143.

[20] SHI L Q, DONG C L, HENG P G, et al. PCA-PSO-ELM model for coal mine water inrush source discrimination[J]. China Sciencepaper, 2021, 16(09): 919-924.

[21] YIN H Y, ZHOU X L, LANG N, et al. Coal seam floor water inrush prediction model based on SSA-optimized GA-BP neural network and its application[J]. Coal Geology & Exploration, 2021, 49(06): 175-185.

[22] SHI L Q, TAN X P, WANG J, et al. Floor water inrush hazard assessment based on PCA-Fuzzy-PSO-SVC[J]. Journal of China Coal Society,2015,40(01): 167-171. DOI:10.13225/j.cnki.jccs.2014.0370.

[23] QIU Y G, ZHOU J. Short-term rockburst damage assessment in burst-prone mines: An explainable XGBoost hybrid model with SCSO algorithm[J]. Rock Mechanics and Rock Engineering, 2023, 56: 8745-8770.

[24] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[25] CHEN X Y, WANG X F, ZHANG D S, et al. Advanced modeling of seepage dynamics and control strategies in thick coal seams under high-confined aquifer conditions: A case study[J]. Alexandria Engineering Journal, 2025, 111: 415-431.

[26] XUE J K, SHEN B. Dung beetle optimizer: A new meta-heuristic algorithm for global optimization[J]. The Journal of Supercomputing, 2023, 79(7): 7305-7336.

[27] Qiu, Yingui, & Zhou, Jian Short-term rockburst damage assessment in burst-prone mines: An explainable XGBoost hybrid model with SCSO algorithm[J]. Rock Mechanics and Rock Engineering, 2023, 56: 8745-8770.