scientific reports

Check for updates

OPEN

Research on the predicted height of water-conducting fracture zones based on the BO-RFR model and SHAP analysis

Mei Qiu[1], Yansen Wen[1], Chao Teng[2] & Meng Shi[3,4]✉

Significant differences exist in the development height of water-conducting fracture zones (WCFZ) and their relationships with influencing factors between Carboniferous-Permian coal seams in central-eastern China and Jurassic coal seams in the western region. Utilizing 147 measured WCFZ height samples from Carboniferous-Permian seams and 111 from Jurassic seams, five influencing factors were selected: mining height, mining depth, working face slope length, hard rock ratio coefficient, and mining method. A Bayesian-optimized random forest regression model (BO-RFR) was then constructed to predict WCFZ heights, with the contributions of individual factors assessed through SHAP (SHapley Additive exPlanations) values. SHAP analysis indicated that, for Carboniferous-Permian seams, the factors influence WCFZ height in descending order as mining height, hard rock ratio coefficient, mining depth, working face slope length, and mining method. Conversely, for Jurassic seams, the ranking is mining height, mining depth, hard rock ratio coefficient, mining method, and working face slope length. This study not only validates the superior predictive capability of the BO-RFR model for WCFZ height but also systematically elucidates the differing mechanisms by which these factors impact WCFZ development across coal seams of two distinct depositional eras. The findings provide targeted theoretical support and practical decision-making guidance for the safety assessment of water-bearing coal mining in central-eastern and western mining regions of China.

Keywords WCFZ height, BO-RFR model, Influencing factors, Jurassic coal seams, Carboniferous-Permian coal seams

Coal has long been the primary energy source in China. However, coal mining inevitably disrupts the original stress equilibrium of the overlying rock strata, leading to the development of water-conducting fracture zones (WCFZ) that extend through these strata. Once these fracture zones connect with overlying water-rich aquifers, they may trigger catastrophic mine water inrushes, posing severe risks to mine safety and regional ecological stability[1]. Therefore, accurate prediction of WCFZ height is essential for ensuring the safe extraction of coal beneath aquifers and maintaining the safety of mining operations[2].

To predict the height of WCFZ, researchers both domestically and internationally have employed various methods, including empirical formulae, numerical simulations, physical modeling, and field measurements[3–5]. However, each of these approaches presents inherent limitations: field measurements provide the highest accuracy but involve substantial operational costs and are both time-consuming and labor-intensive[6]. The empirical formula method, although straightforward and easy to apply, accounts solely for mining height influence, resulting in notable prediction errors[7]. Numerical simulations depend critically on the selection of geological parameters, which are often challenging to determine with precision[8]. Physical similarity modeling demands accurate material proportioning and faces limitations in replicating complex geological conditions[9]. In recent years, machine learning techniques have offered novel solutions for predicting the height of WCFZ, demonstrating significant potential[10–12]. Zhu et al. (2022) developed an ICS-ELM model for WCFZ height prediction, achieving promising results[13]. Wang et al. (2024) addressed the common issue of traditional backpropagation neural networks (BPNNs) becoming trapped in local optima by proposing an improved BPNN model based on differential evolution and grey wolf optimizer (DEGWO)[10]. Xu et al. (2024) optimized extreme

[1]College of Earth Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, Shandong, China. [2]Cheng'an Engineering Technology Group CO., Ltd, 266555 Qingdao, China. [3]Shandong No: 3 Exploration Institute of Geology and Mineral Resources, Yantai 264004, China. [4]Observation and Research Station of South Yellow Sea Earth Multi-sphere, MNR, Yantai, China. ✉email: 717112565@qq.com

gradient boosting (XGBoost) by employing genetic algorithms, particle swarm optimization, and sparrow search algorithms, resulting in a combined optimization model for WCFZ height prediction[14]. However, most existing studies focus on constructing generalized models or predictions for single mining sites, with limited interpretability and insufficient exploration of the underlying geomechanical mechanisms.

It is particularly noteworthy that, with the westward shift of coal mining activities in China, the primary coal-bearing strata have transitioned from the Carboniferous-Permian in the central and eastern regions to the Jurassic in the western regions. These two coal-bearing sequences formed under distinctly different sedimentary environments: the former predominantly under marine–terrestrial transitional conditions, and the latter under terrestrial conditions. This results in significant differences in coal seam occurrence, overlying strata lithological composition, and mechanical properties[15]. Currently, the vast majority of studies either overlook these fundamental geological differences, conflate the two coal-bearing systems, or focus exclusively on one. To date, there is a lack of systematic research comparing the development height of WCFZ and the governing factors between the Carboniferous-Permian and Jurassic coal seams. This knowledge gap impedes the provision of targeted mining optimization and hazard prevention strategies tailored to the distinct geological contexts of different mining areas.

To address the aforementioned issues, this study aims to quantitatively compare and elucidate the development patterns of WCFZ in Carboniferous-Permian and Jurassic coal seams for the first time, using a data analysis framework characterized by high accuracy, reliability, and interpretability. To achieve this goal, we selected a Bayesian-optimized random forest regression model (BO-RFR) as the core analytical tool. Random Forest Regression (RFR), proposed by Breiman (2001), is a non-parametric regression method suitable for problems with limited prior knowledge and incomplete data[16]. It efficiently handles high-dimensional datasets and provides robust predictive accuracy alongside strong tolerance to outliers and noise[17]. Furthermore, RFR includes an internal mechanism for estimating feature importance. The dataset constructed in this study covers coal-bearing strata of various sedimentary ages, characterized by a relatively limited sample size and inherent heterogeneity. RFR, an ensemble algorithm based on decision trees, effectively mitigates overfitting in datasets of this scale through its intrinsic mechanisms of random sampling and feature subspace selection. It robustly captures the complex nonlinear relationships and interactions between influencing factors and the height of hydraulic fracture zones. Compared to deep learning models, which depend on large volumes of labeled data and complex architectures, RFR typically demonstrates superior generalization performance and computational efficiency on medium-sized, multi-source heterogeneous datasets. Importantly, the structural properties of the RFR model offer a natural and efficient interface for detailed feature contribution analyses, such as SHAP (Shapley Additive exPlanations), which is critical for achieving the core research objective of elucidating variation mechanisms. Furthermore, from an engineering application perspective, the optimized RFR model features hyperparameters with clear physical interpretations, is relatively straightforward to train and deploy, and provides stable prediction outcomes, thereby facilitating its practical implementation in mining field operations. RFR models have been widely applied across disciplines, including ecology[18], biology[19], atmospheric science[20], and surveying[21], demonstrating notable success. Bayesian Optimization (BO) is an efficient global hyperparameter tuning algorithm that automatically identifies the optimal parameter set to maximize model performance, thereby providing a reliable and robust foundation for subsequent mechanistic analysis[22,23]. By integrating BO with RFR, denoted as BO-RFR, the objective is to obtain a model that achieves optimal and stable predictive performance under the given data conditions, ensuring the reliability of results derived from this model.

Based on this approach, the study collected 147 measured data sets of WCFZ heights for Carboniferous-Permian coal seams and 111 data sets for Jurassic coal seams. Five key influencing factors were selected: mining height, mining depth, working face slope length, hard rock ratio coefficient, and mining method. This study aims to achieve two primary objectives: (1) development of differentiated predictive models: establish high-accuracy BO-RFR models separately for Carboniferous-Permian and Jurassic coal seams, providing customized WCFZ height prediction tools tailored to distinct geological conditions; (2) quantitative comparison of development mechanisms: employ SHapley Additive exPlanations (SHAP) interpretable machine learning techniques to analyze and quantify the magnitude and direction of influencing factors on WCFZ height in the two coal systems, thereby elucidating the geomechanical causes of their differences and ultimately offering a scientific basis for formulating differentiated water hazard prevention strategies.

Analysis of influencing factors on the WCFZ height

Underground coal mining disrupts the original in-situ stress equilibrium of the overlying strata, inducing a series of complex mechanical deformations such as bending, subsidence, and fracturing. Variations in displacement rates within the strata result in the formation of distinct fractures, including transverse fractures parallel to the bedding strike and longitudinal fractures oriented perpendicular or obliquely to the bedding strike. The interconnected network of these fractures in the roof strata above the coal seam constitutes the WCFZ. Depending on the extent of overburden deformation, the WCFZ is further subdivided into the caving zone and the fracture zone. The development of WCFZ during coal mining is influenced by a complex set of factors. Based on previous research and practical mining experience, the primary factors include mining height, mining depth, working face slope length, mining method, hard rock ratio coefficient, mining rate, and coal seam dip angle[24]. Considering both the significance of these factors and the availability of data, this study focuses on five key variables—mining height, working face slope length, mining depth, hard rock ratio coefficient, and mining method—to predict the height of the WCFZ. Figure 1 presents a conceptual diagram of the WCFZ height and its influencing factors. These factors constitute an interconnected system, whose combined effects govern the scale and morphology of overburden damage.

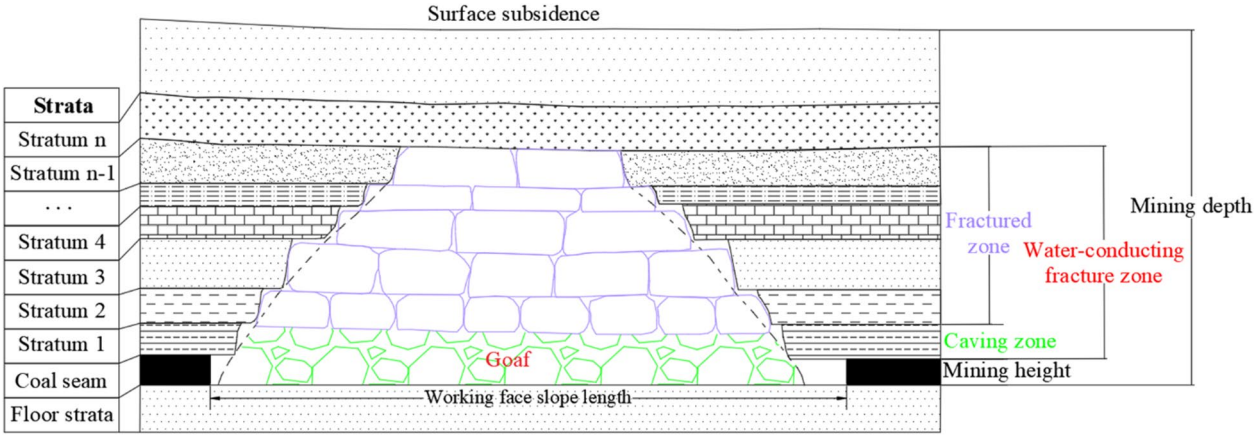

Fig. 1. Schematic diagram of the development of the WCFZ induced by coal seam mining.

Mining height (M)
Mining height is a key determinant of the development height of WCFZ. An increased mining height expands the plastic zone of the coal seam roof and enlarges the void space available for the caving rock strata, thereby promoting the upward extension of the WCFZ. Under specific conditions, the progression of roof fracture damage exhibits an approximate fractional functional relationship with mining height.

Mining depth (D)
According to the theory of mine pressure, the mining depth of a coal seam governs the magnitude of the in situ rock stress. With increasing mining depth, the in situ stress on the surrounding rock at the working face intensifies, leading to higher horizontal and vertical loads on the roof strata of the goaf and promoting the development of joint fractures. This results in fracturing and failure of the overlying rock layers, ultimately causing the height of the WCFZ to increase within a certain range as mining depth grows.

Working face slope length (L)
The length of the inclined working face reflects the extent of coal seam extraction and significantly affects the development of the WCFZ. Prior to the coal seam reaching full mining-induced disturbance, the height of the WCFZ increases gradually with the advancement of the working face. However, once the coal seam is fully disturbed by mining, the height of the WCFZ remains nearly constant despite continued advancement of the working face.

Hard rock ratio coefficient (b)
The hard rock ratio coefficient is defined as the ratio of the cumulative thickness of hard rock layers within the WCFZ to the height of the fracture zone. The specific calculation method is provided in Eq. (1). This coefficient not only reflects the uniaxial compressive strength characteristics of the overlying strata but also characterizes the structural composition of the rock layers. It effectively overcomes the limitations of existing standards that rely solely on uniaxial compressive strength, which cannot adequately differentiate roof rock types and their structural configurations.

$$b = \frac{\sum h}{H_W} \tag{1}$$

$$H_W = (15 \sim 20)M \tag{2}$$

Where b represents the hard rock ratio coefficient, Σh denotes the cumulative thickness of hard roof rock layers, and H_W is the height of the WCFZ. H_W is calculated using the Eq. (2)[25]:
Where M donates the mining height.

Mining method (w)
With technological advancements, the degree of mechanization in coal mining has increased substantially, leading to the replacement of traditional blasting mining by more efficient extraction methods such as fully mechanized longwall mining and fully mechanized caving mining. Different mining techniques induce varying degrees of disturbance to the overlying roof strata, resulting in differential impacts on the development height of the WCFZ. More aggressive methods, such as blasting mining, generally facilitate greater development of this fracture zone. As the mining method represents a qualitative factor affecting the height of the WCFZ, this study assigns numerical labels for analytical convenience: fully mechanized longwall mining is designated as 0, fully mechanized caving mining as 1, blasting mining as 2, and layered mining as 3.

Sample data and methods

Sample data

To investigate the prediction of the WCFZ height during the mining of Carboniferous-Permian and Jurassic coal seams, and to minimize sample bias, representative observational data were collected from typical domestic mining regions. This dataset comprises 147 samples from Carboniferous-Permian coal seams[1] and 111 samples from Jurassic coal seams[12,25,26]. The sample data are primarily derived from two coal-bearing regions in China: the Carboniferous-Permian coal mines in the east-central area and the Jurassic coal mines in the western area, as shown in Fig. 2. The Carboniferous-Permian coal seams, distributed mainly across Shanxi, Henan, Shandong, and Anhui provinces, are deeply buried, have relatively thin seam thicknesses, and an average dip angle of approximately 10°. The principal mining methods employed are longwall mining and fully mechanized caving. The overlying strata in this region consist of well-cemented rock masses, predominantly formed by siliceous and calcareous cementation, resulting in high uniaxial compressive strength. In contrast, Jurassic coal mines are located within the Ordos Basin, spanning Inner Mongolia, Ningxia, and Shaanxi provinces. Jurassic coal seams in this area are shallower, thicker, and characterized by smaller dip angles. Fully mechanized caving is the primary mining technique. The overburden here exhibits weaker cementation, mainly argillaceous, producing rock masses with comparatively lower uniaxial compressive strength. Consequently, the distinct geological settings of the Carboniferous-Permian and Jurassic coal seams result in significantly different development heights of mining-induced WCFZ within the respective stratigraphic sequences.

To more clearly illustrate the development height of the WCFZ in the Carboniferous-Permian and Jurassic coal seams and the factors influencing them, violin plots of the observational data were generated, as shown in Fig. 3. A violin plot integrates features of both box plots and kernel density plots, effectively displaying the data range, mean, median, quartiles, and probability density distribution. Taking the violin plot of the WCFZ height in Fig. 3 as an example, the left section illustrates the distribution for the Carboniferous-Permian coal seams, with a mean height of approximately 47.3 m, a 25th percentile of 35.2 m, and a 75th percentile of 57.5 m. The right section represents the data distribution for the Jurassic coal seams, where the mean WCFZ height is about 118 m, with the 25th and 75th percentiles at 83.4 m and 141 m, respectively. Both the mean values and the probability density of the WCFZ height for the Jurassic coal seams exceed those of the Carboniferous-Permian seams significantly. The ratio of the WCFZ height to the mining height, referred to as the fracture-to-mining ratio (k), quantifies the vertical extent of fracture development per unit of coal seam extraction. This metric is a crucial descriptive parameter for evaluating the sensitivity and efficiency of overburden failure. In this study, the fracture-to-mining ratio is introduced to provide essential data context and a comparative benchmark for subsequent mechanistic interpretations; however, it is not utilized as an input feature in the predictive model. As illustrated in Fig. 3, the fracture-to-mining ratio for Jurassic coal seams is significantly higher than that of Carboniferous-Permian seams, providing empirical evidence that the overburden in Jurassic strata is geotechnically weaker and subjected to more severe failure. Moreover, the mining height and working face slope length for the Jurassic coal seams are generally higher compared to those for the Carboniferous-Permian seams, while the mining depth and hard rock ratio coefficient are comparatively lower.

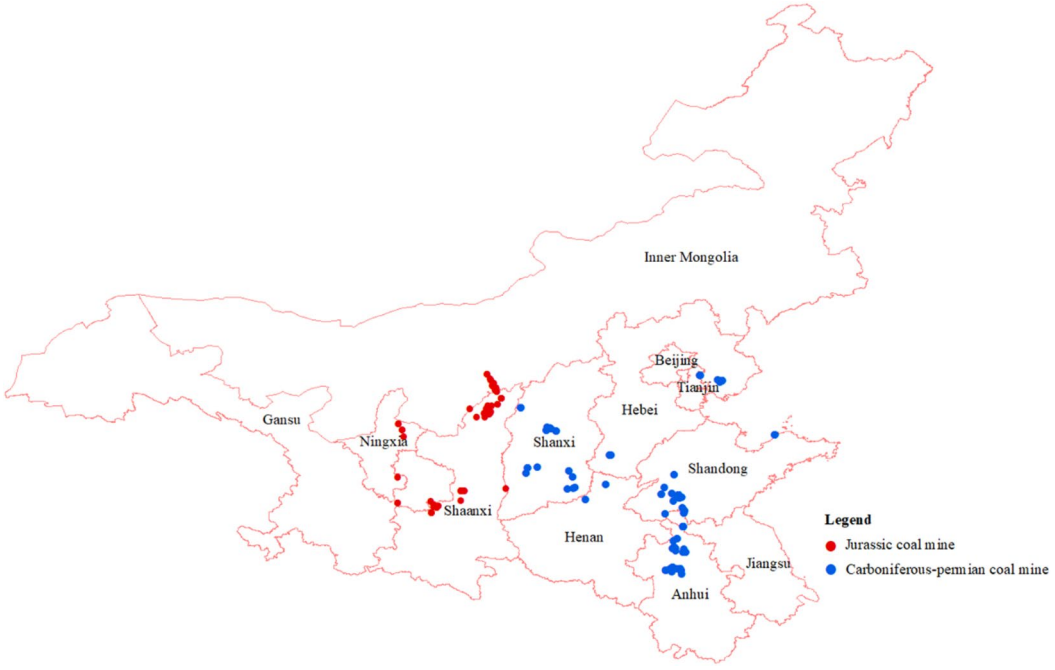

Fig. 2. The spatial distribution map of 147 samples from Carboniferous-Permian coal seams (data source:[1]) and 111 samples from Jurassic coal seams (data source:[12,26,27]). This map was generated by Y.W. using ArcGIS software version 10.7 (https://www.esri.com/en-us/arcgis/products).

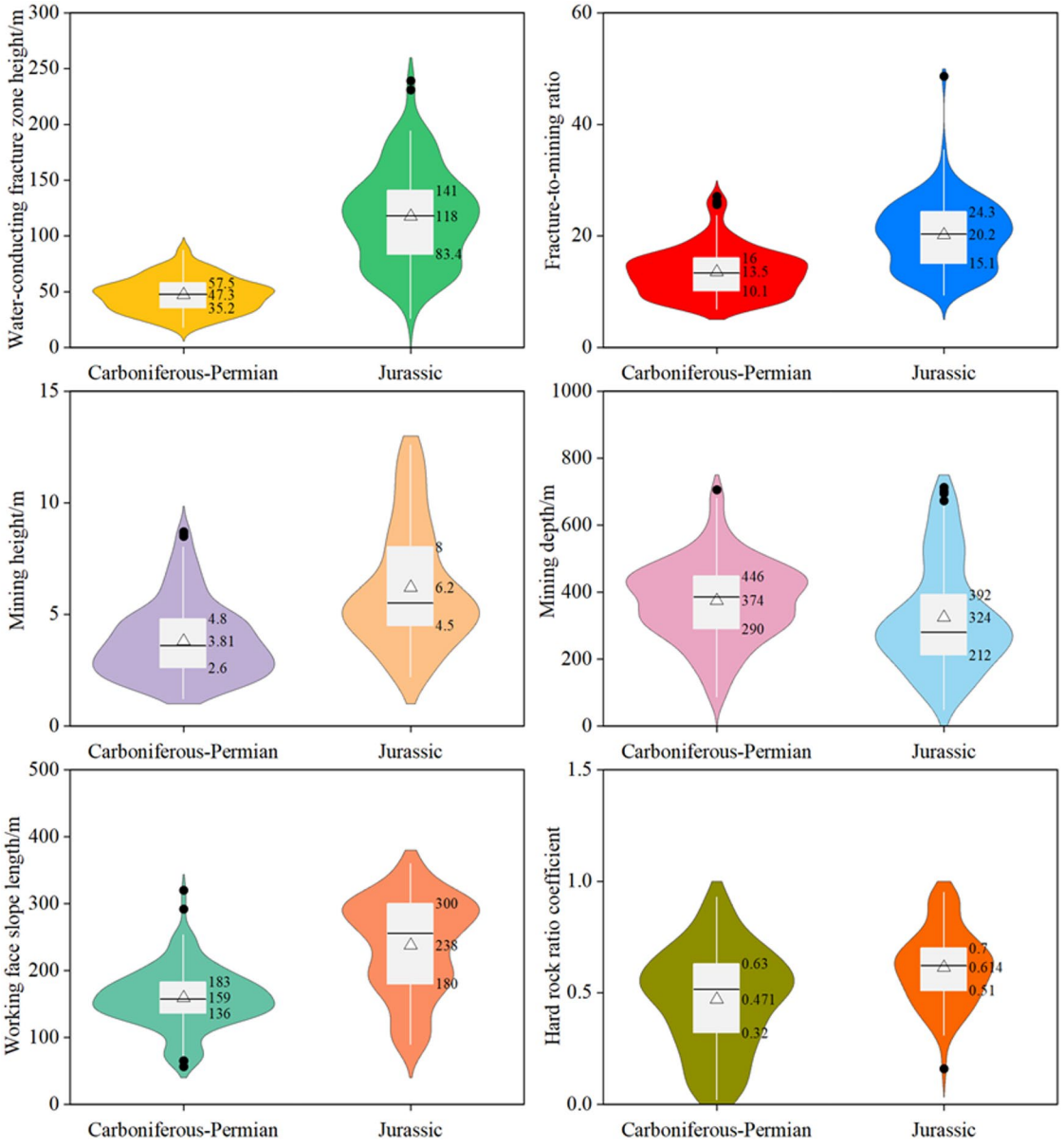

Fig. 3. The violin plot of the observational data.

The marked differences in data distribution observed in the violin plots arise from the distinct sedimentary environments and geomechanical properties of the two coal-bearing strata. The Carboniferous-Permian coal seams are characterized by deep burial, limited mining height, hard overburden, and a complex lithological composition, resulting in comparatively lower and less variable WCFZ heights. In contrast, the Jurassic coal seams generally exhibit shallow burial, greater mining height, weak overburden, and relatively homogeneous stratigraphy, which lead to significantly higher WCFZ heights and fracture-to-mining ratios, along with a broader range of data variation. The data distribution for the Carboniferous-Permian seams is typically more compact and skewed toward lower values, reflecting more constrained and predictable conditions. Conversely, the Jurassic coal seams show broader data distributions biased toward higher values, especially in terms of WCFZ height and fracture-to-mining ratios, suggesting that under weak geological conditions, mining-induced stress responses are more variable and mining-related damage is more severe.

Methods

The research methodology applied in this study is summarized in Fig. 4 and comprises the following steps: (i) A systematic collection of measured WCFZ height samples was conducted for Carboniferous-Permian and Jurassic coal seams from representative mining districts in China. Five key influencing factors—mining height, mining depth, working face slope length, hard rock ratio coefficient, and mining method—were selected to construct a feature-target dataset, which was then divided into training and testing sets according to a predetermined ratio. (ii) For each coal seam dataset, a predictive model based on RFR was developed. BO was applied to automatically

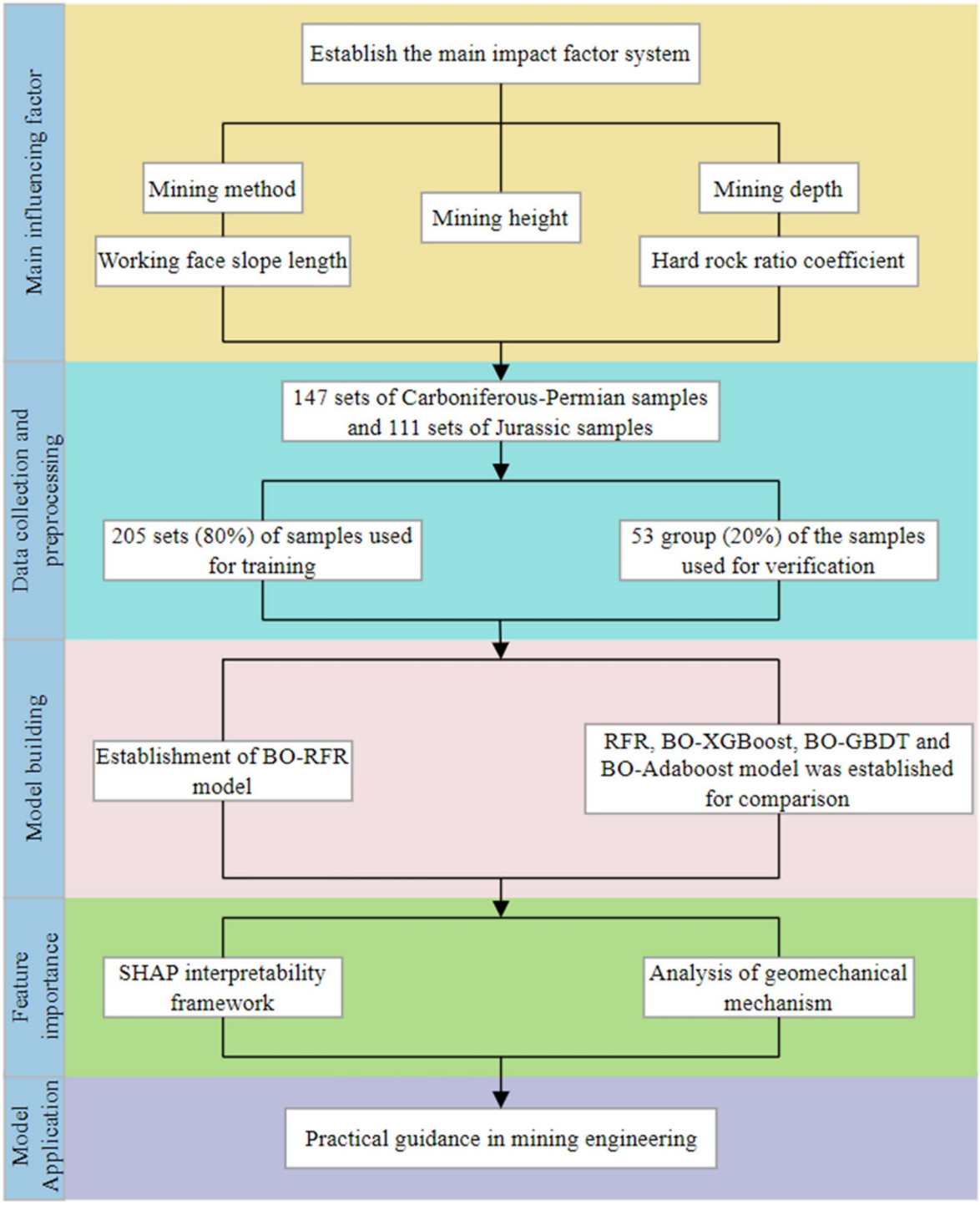

Fig. 4. Research method flow.

tune critical hyperparameters, yielding a high-performance BO-RFR model. (iii) The predictive accuracy of the BO-RFR model was assessed using an independent testing set. Comparative analyses were conducted against the non-optimized RFR model and other BO-optimized ensemble models, including XGBoost, GBDT, and AdaBoost, to demonstrate the superiority of the BO-RFR approach. (iv) The SHAP interpretability framework was applied to quantitatively assess the contribution magnitude and direction of each influential factor on the WCFZ height. This analysis elucidated the dominant controlling factors and their geomechanical mechanisms within different coal-bearing strata. (v) Based on the model predictions and interpretability results, complemented by traditional statistical methods such as Grey Relational Analysis and empirical formulas, a multi-faceted comparative discussion was conducted. The study ultimately proposed differentiated mitigation strategies with practical significance for mining engineering applications.

Random forest regression (RFR)

RFR is a machine learning algorithm based on the Bagging (Bootstrap Aggregating) ensemble strategy. Its core principle involves constructing a large number of diverse decision trees and integrating their predictions to achieve superior accuracy and generalization compared to individual models[27]. The randomness in RFR arises from two key processes: first, each decision tree is trained on a bootstrap sample—i.e., a subset of the original training data randomly drawn with replacement—resulting in each tree using approximately 63.2% of the original samples. The remaining approximately 36.8% of out-of-bag (OOB) samples serve as an unbiased estimate for model performance. Second, at each node split within a tree, the algorithm selects a random subset of features—typically the square root or one-third of the total feature set—rather than searching all features for the optimal split. This dual-randomization mechanism enhances diversity among trees and reduces the overall model variance, thereby effectively mitigating overfitting[28].

The construction of a single regression decision tree involves a recursive, greedy splitting process that partitions the feature space into a set of homogeneous subregions[29]. For any node t, impurity is quantified by the mean squared error (*MSE*), as expressed in Eq. (3):

$$I(t) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{3}$$

Where $I(t)$ denotes the impurity of node t, N represents the number of samples at the node, y_i is the target value of the i-th sample, and $\overline{y}$ is the mean target value within the node. Based on a randomly selected subset of features and their potential split thresholds, the algorithm evaluates all candidate splits and chooses the one that maximizes impurity reduction. The optimization objective is to maximize the information gain, as defined in Eq. (4):

$$\Delta I = I(t) - \frac{N_L}{N}I(t_L) - \frac{N_R}{N}I(t_R) \tag{4}$$

Where ΔI denotes the information gain; t_L and t_R represent the left and right child nodes generated by the split, respectively; N_L and N_R are the sample sizes of these child nodes; and $I(t_L)$ and $I(t_R)$ are their corresponding impurities. The recursive splitting process continues until predefined stopping criteria are met, such as a minimum number of samples in a node, reaching the maximum tree depth, or impurity falling below a specified threshold. At this point, the node becomes a leaf, with its output value defined as the mean target value of all samples within the node.

During the prediction phase, the RFR aggregates the outputs of individual decision trees through ensemble averaging, as illustrated in Fig. 5. For a new input sample x, each tree T_k produces a predicted value $\widehat{y}_k(x)$ based on its internal splitting rules. The final prediction of the RFR is the arithmetic mean of all individual tree predictions. This averaging process effectively mitigates the high variance associated with fully grown single trees, resulting in more stable and reliable predictions. Moreover, RFR inherently provides a measure of feature importance, typically computed by averaging and normalizing the total reduction in node impurity contributed by each feature across all trees. A higher feature importance score indicates a greater contribution of that feature to reducing prediction uncertainty.

In RFR models, hyperparameters such as the number of trees and their maximum depth critically affect model learning capacity and generalization performance. Improper hyperparameter settings can result in underfitting or overfitting. Traditional methods, including grid search and random search, are computationally expensive and inefficient. Conversely, BO can adaptively guide the search process based on historical evaluation results, allowing for the identification of optimal hyperparameter combinations with fewer iterations. This approach ensures the resulting model achieves both high predictive accuracy and strong generalization, thereby providing a reliable basis for subsequent analysis. Therefore, this study employs BO to tune the hyperparameters of the RFR.

Bayesian optimization (BO)

BO is a global optimization algorithm that utilizes prior information about the objective function to guide the selection of new sample points, thereby significantly improving search efficiency. It is widely applied in machine learning tasks such as hyperparameter tuning and neural architecture search[30]. BO begins by defining a prior distribution over the objective function. Using observed data and this prior, it applies Bayes' theorem to derive the posterior distribution of the objective function. The next sample point is then selected based on this posterior distribution[31]. By effectively incorporating information from previous samples, Bayesian Optimization models the objective function's shape to identify parameter sets that optimize the global objective. The mathematical formulation of Bayes' theorem is provided in Eq. (5).

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \tag{5}$$

Where x denotes the sample input, corresponding to the hyperparameters of the random forest in this study; y represents the sample output, specifically the loss function used herein; $p(y|x)$ is the posterior probability distribution of y; $p(x|y)$ denotes the likelihood of observing x given y; and $p(y)$ indicates the prior probability distribution of y.

The basic procedure of BO is as follows: (i) construct a probabilistic surrogate model M of the objective function based on the existing observed dataset $(x_{1:t}, y)$; (ii) use an acquisition function, guided by the surrogate

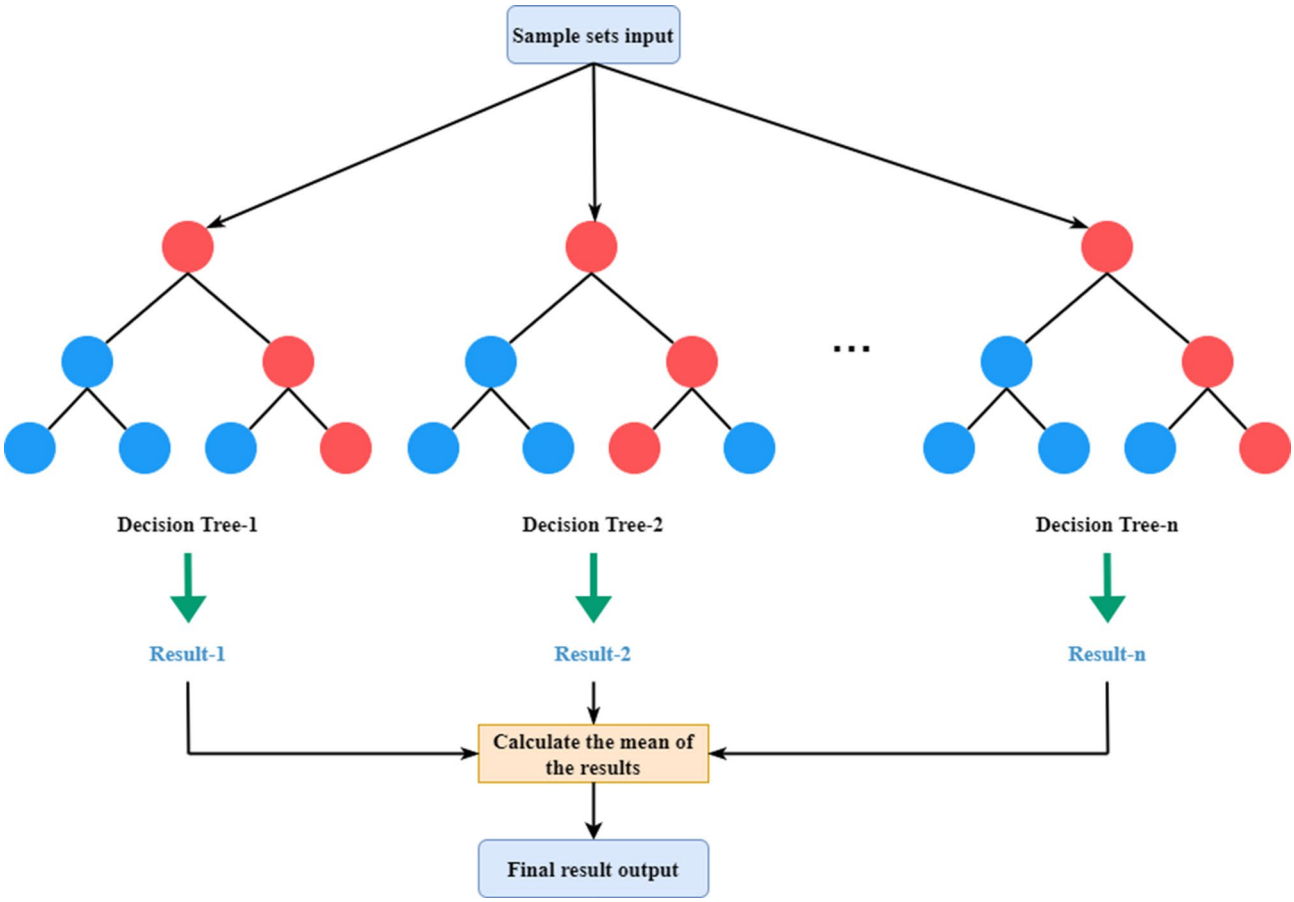

Fig. 5. Schematic diagram of a random forest.

model M, to select the next set of hyperparameters x_{t+1}; (iii) incorporate the new hyperparameter set x_{t+1} into the observed dataset; and (iv) repeat steps (i) to (iii) until the maximum number of iterations is reached, at which point the optimal hyperparameter set is returned. Thus, the core components of BO are the probabilistic surrogate model and the acquisition function.

Probabilistic surrogate model

Gaussian Processes (GP) are commonly used probabilistic surrogate models in BO. GPs assume that the relationship between the input x and the observed output y follows a Gaussian distribution and construct a joint probability distribution over inputs and outputs by computing the covariance matrix among sample points in the dataset. As a classical surrogate model, GP adaptively learns complex input-output relationships to approximate the objective function. However, it is limited in handling missing data and noise. In contrast, the Tree-structured Parzen Estimator (TPE) provides greater flexibility in model construction and effectively overcomes these limitations. TPE defines $p(x|y)$ using two density functions, as shown in Eq. (6).

$$p(x \mid y) = \begin{cases} l(x), & y<y^* \\ g(x), & y \geq y^* \end{cases} \tag{6}$$

Here, $y*$ denotes a threshold defined by $\gamma = p(y < y*)$, where γ represents the quantile of the observed outcomes y; $l(x)$ is the density function derived from observations with y values below $y*$; and $g(x)$ is the density function corresponding to the remaining observations.

Acquisition function

In the TPE algorithm, Expected Improvement (EI) serves as the acquisition function to identify the next sampling point. The formulation of EI is presented in Eq. (7).

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y)\, p(y \mid x)\, dy = \int_{-\infty}^{y^*} (y^* - y) \frac{p(x \mid y)\, p(y)}{p(x)}\, dy \tag{7}$$

Where $p(x)$ is calculated according to Eq. (8):

$$p(x) = \int p(x \mid y)p(y)dy = \gamma l(x) + (1-\gamma)g(x) \tag{8}$$

Substituting Eqs. (6) and (8) into Eq. (7) yields:

$$EI_{y*}(x) = \frac{\gamma y * l(x) - l(x)\int_{-\infty}^{y*} p(y)dy}{\gamma l(x) + (1-\gamma)g(x)} \propto (\gamma + \frac{g(x)}{l(x)}(1-\gamma))^{-1} \tag{9}$$

In each iteration of BO, the algorithm selects the point $x*$ that maximizes the EI as the next candidate sample. As indicated by Eq. (9), the EI value is proportional to the ratio $l(x)/g(x)$. Consequently, the sample point corresponding to the maximum value of this ratio is chosen as the subsequent sampling location.

BO-RFR model

This study employs BO to tune the hyperparameters of the RFR algorithm, with the process flow illustrated in Fig. 6. The prediction of WCFZ height using the BO-RFR model can be summarized in the following steps:

Step 1: The dataset for the WCFZ height, comprising 147 samples from the Carboniferous-Permian system and 111 samples from the Jurassic system, is split into training and testing sets at an 8:2 ratio. Descriptive statistics and outlier detection are performed on five input features: mining height, mining depth, working face slope length, hard rock ratio coefficient, and encoded mining method, to ensure data quality. The target variable is the height of the WCFZ.

Step 2 Multiple regression decision trees are constructed using the training set. Each tree is built on a bootstrap-sampled subset of the data. At each node, the optimal split point is selected from a randomly chosen subset of features to minimize the variance of the target variable within the node.

Step 3 For new input samples, each decision tree independently produces a prediction. The final RFR prediction is the average of all individual tree outputs, effectively reducing variance and enhancing prediction stability.

Step 4 BO is applied to intelligently tune key RFR hyperparameters, such as the number of trees and maximum depth. This approach leverages TPE to model the objective function probabilistically, directing the search within the hyperparameter space and efficiently identifying optimal parameter combinations.

Step 5 The optimized hyperparameters are used to retrain the model, yielding the final BO-RFR model. Its performance is then evaluated on an independent test set and compared with pre-optimization results and other benchmark models to quantify the improvement.

The principle of SHAP (Shapley additive exPlanations)

SHAP, grounded in cooperative game theory, provides a unified framework for interpreting feature importance in machine learning models[32]. In a cooperative game, the Shapley value offers a fair allocation scheme to distribute the total coalition payoff among the participants. Translating this concept to model interpretation, each individual prediction is treated as a game, with features as players and the model's prediction for that instance as the total payoff. The SHAP value corresponds to the equitable share of the payoff assigned to each feature. This allocation simulates all possible scenarios of features joining the predictive feature subset in different orders: for a given feature, its marginal contribution is systematically evaluated by measuring the change in prediction when the feature is added to every possible subset of the remaining features. The SHAP value is then computed as the weighted average of these marginal contributions across all feature subsets. The SHAP value for feature i is formally defined by Eq. (10):

$$\phi_i(f,x) = \sum_{S \subseteq \{1,2,\ldots,p\} \setminus \{i\}} \frac{|S|!(p-|S|-1)!}{p!}[f(S \cup \{i\}) - f(S)] \tag{10}$$

Where f denotes the trained machine learning model, x represents the specific sample to be explained, p is the total number of features, and S denotes any subset of the feature set excluding feature i; $f(S)$ denotes the expected prediction value of the model for sample x using only the feature subset S. In practice, this is typically estimated by marginalizing over the missing features outside S using their corresponding values across all samples in the dataset. The term $[f(S \cup \{i\}) - f(S)]$ reflects the marginal contribution of adding feature i to the coalition S. The weighting factor $[|S|!(p-|S|-1)!/p!]$ averages over all possible permutations of feature orderings to ensure a fair allocation.

This computational approach ensures fairness in the results, as it accounts not only for the individual effect of the feature but also thoroughly evaluates the joint impact of its complex interactions with all other features on the final prediction. However, directly computing the Shapley value using this formula is computationally expensive, exhibiting exponential complexity. For tree-based models, such as the RFR employed in this study, Lundberg et al. (2020) proposed the TreeSHAP algorithm[33], which exploits the tree structure to compute SHAP values exactly in polynomial time, significantly improving computational efficiency. TreeSHAP recursively traverses each node of the decision tree, estimating the marginal contribution of features along decision paths through conditional probability computations, thereby circumventing exhaustive coalition enumeration.

In this study, SHAP analysis provides explanations at two levels for the WCFZ height predictions produced by the BO-RFR model. For individual samples, each feature, such as mining height or hard rock ratio coefficient,

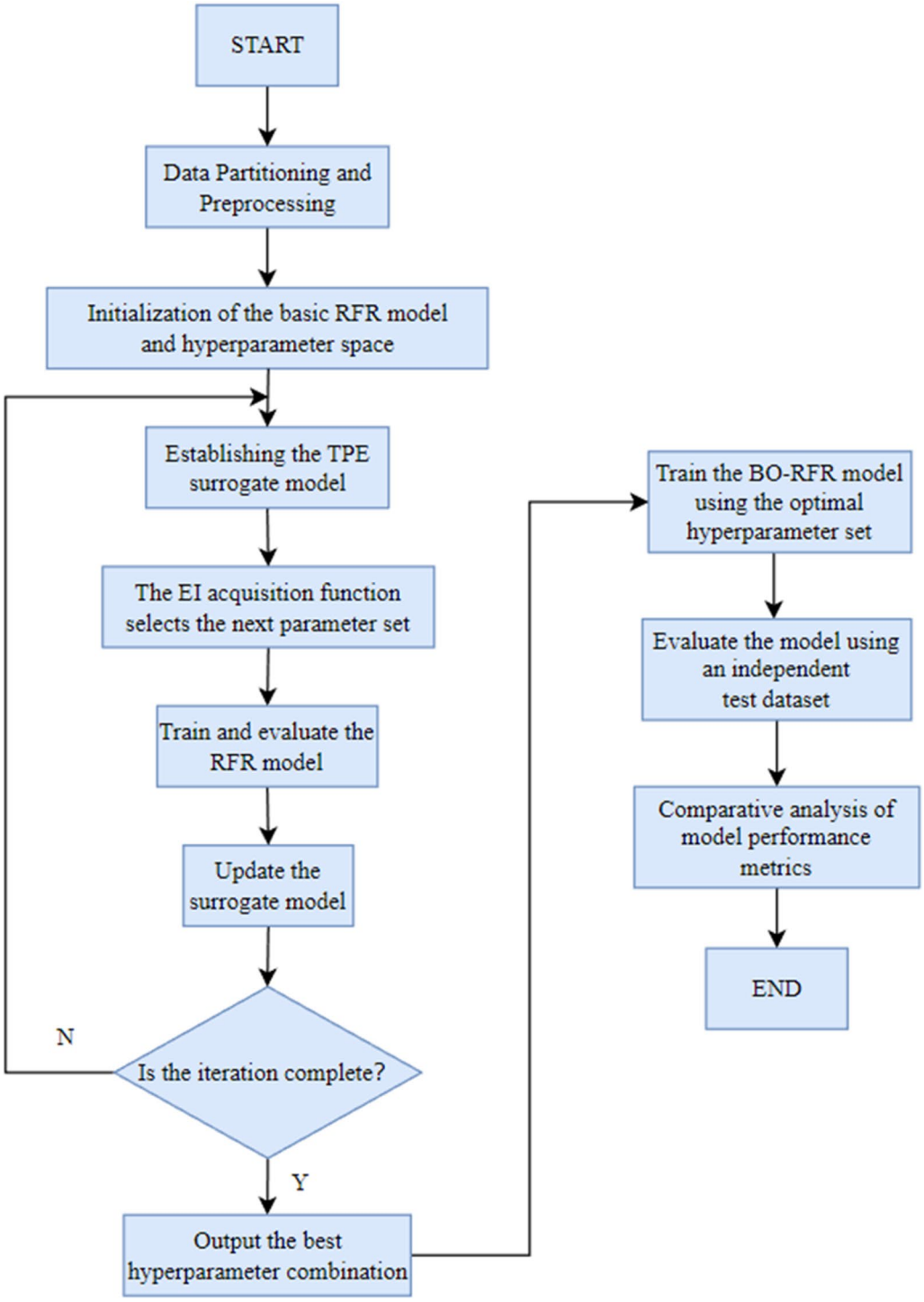

Fig. 6. Flowchart of the BO-RFR model.

has a corresponding SHAP value ϕ_i. A positive ϕ_i indicates that, under the specific conditions of the sample, the feature positively contributes to the predicted WCFZ height; conversely, a negative value indicates an inhibitory effect. The sum of all feature SHAP values ϕ_i explains the deviation of the sample's predicted value from the mean prediction across all samples. Aggregating SHAP values across all samples yields the global feature importance, commonly measured as the mean absolute SHAP value $|\varphi_i|$. This metric offers a more direct and interpretable measure of the average impact of features on model output magnitude compared to traditional importance scores based on impurity reduction.

Selection of model evaluation metrics

To ensure prediction accuracy and enhance generalization capability, the dataset was split into training and testing sets at an 8:2 ratio. The optimal training model was identified through 5-fold cross-validation, followed by performance evaluation on the test set. To rigorously evaluate model stability under small-sample conditions, this study employed a repeated cross-validation approach. Specifically, 20 independent repetitions ($N=20$) of 5-fold cross-validation were conducted on the training set. In each repetition, a distinct random seed was used to shuffle the data and generate fold partitions, ensuring the independence of each iteration. The evaluation metrics used to assess the BO-RFR model predictions include the coefficient of determination (R^2), root mean square error (*RMSE*), mean absolute error (*MAE*), and mean absolute percentage error (*MAPE*). These metrics collectively quantify the model's goodness of fit. Specifically, R^2 measures the proportion of variance in the dependent variable explained by the independent variables, ranging from 0 to 1, with values closer to 1 indicating a better fit; its calculation is provided in Eq. (11). *RMSE*, *MAE*, and *MAPE* assess the deviations between predicted and observed values, where smaller values reflect higher prediction accuracy and improved model performance. Their respective formulas are presented in Eqs. (12), (13), and(14).

$$R^2(x) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{13}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{14}$$

Where y_i denotes the actual target value, $\hat{y}_i$ represents the predicted target value, $\overline{y}$ is the mean of the target values, and n is the total number of samples.

Results and analysis

Statistical analysis of WCFZ height and influencing factors

Linear correlation analysis based on the pearson correlation coefficient

The Pearson correlation coefficient is used to measure the linear relationship between two continuous variables. Its value ranges from − 1 to 1, where 1 indicates a perfect positive correlation, −1 indicates a perfect negative correlation, and 0 indicates no correlation. We conducted Pearson correlation analyses to examine the relationships between the WCFZ height and various influencing factors under two coal-bearing stratigraphic conditions. The results are presented in Table 1. After distinguishing between the coal-bearing strata, mining height was identified as the factor most strongly linearly correlated with the WCFZ height. The Pearson correlation coefficients between mining height and WCFZ height in the Carboniferous-Permian and Jurassic coal seams were 0.706 and 0.734, respectively, significantly higher than those of other factors. As shown in Fig. 7(a), the linear fit coefficients (R^2) between mining height and WCFZ height for the Carboniferous-Permian and Jurassic coal seams are 0.53 and 0.50, respectively, indicating a strong linear relationship. According to the classification criteria proposed by Mokoka, a correlation coefficient below 0.50 is generally interpreted as indicative of a low level of correlation[34]. In Table 1, aside from mining height, the Pearson correlation coefficients of other factors are all below 0.50. Consistently, the R^2 values of linear fits shown in Fig. 7 are also relatively low, suggesting a weak linear correlation between these factors and WCFZ height. However, this does not imply that these factors have no significant association with WCFZ height. Consequently, a grey relational analysis

Coal seam	Influencing factors	Pearson correlation coefficient	*p*-value
Carboniferous-Permian	Mining height	0.706	0.000
	Mining depth	0.287	0.000
	Working face slope length	0.270	0.001
	Hard rock ratio coefficient	0.290	0.002
Jurassic	Mining height	0.734	0.000
	Mining depth	0.481	0.000
	Working face slope length	−0.255	0.007
	Hard rock ratio coefficient	0.186	0.115

Table 1. Correlation analysis between WCFZ height and influencing factors under different coal seam conditions.

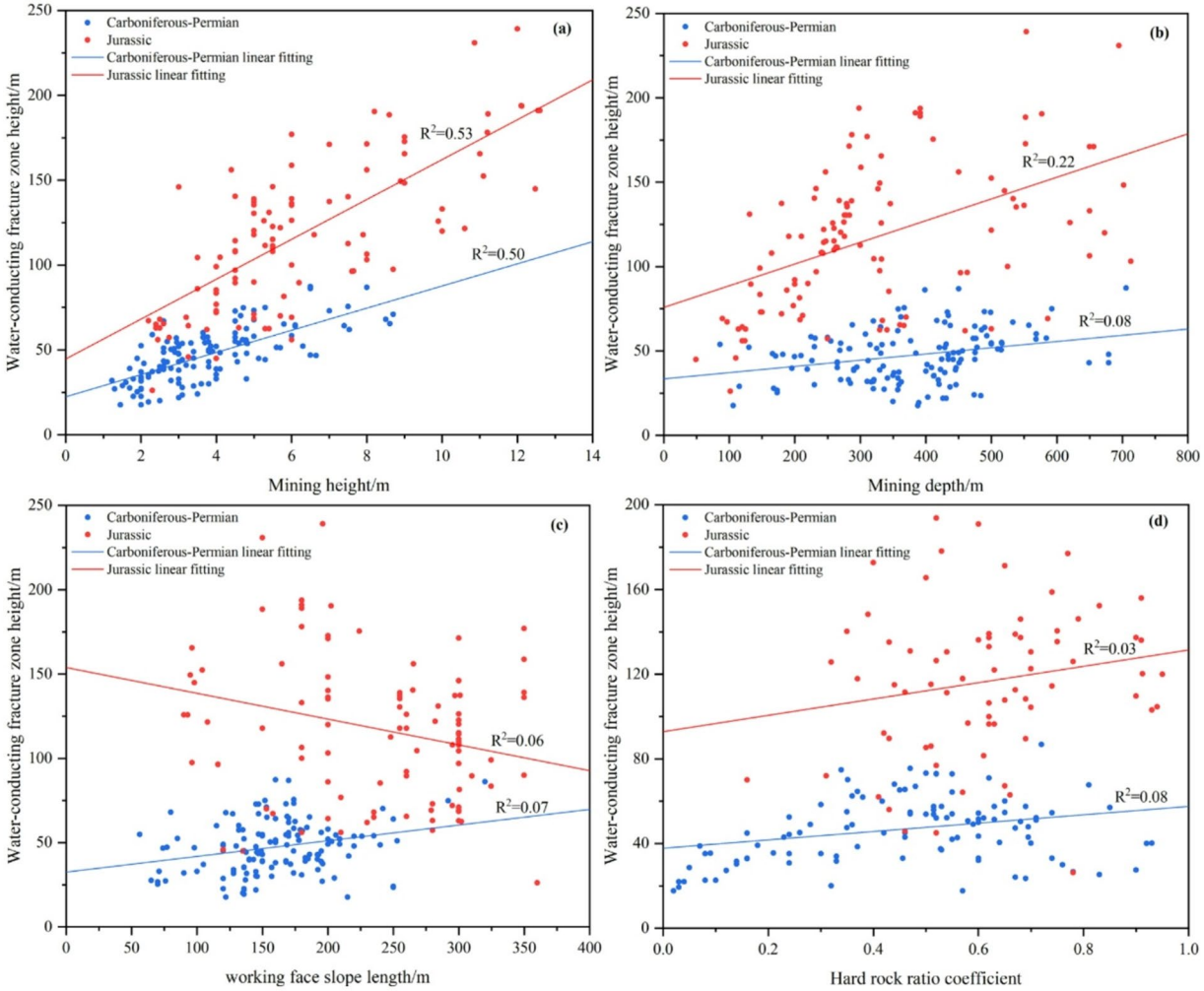

Fig. 7. Scatter plots between WCFZ height and influencing factors under different coal seam conditions.

was subsequently employed to further investigate the degree of association between the influencing factors and WCFZ height.

Trend correlation analysis based on grey relational degree

Conventional correlation analyses typically require large sample sizes and assume that data follow typical distribution patterns. In contrast, grey relational analysis is a robust multivariate statistical method well-suited for small samples and systems with incomplete information. Its core principle lies in evaluating the similarity of geometric shapes between data sequences to determine the strength of their association. The closer the shapes and the more consistent the trends, the higher the degree of correlation; conversely, less similarity indicates a weaker association[35]. Using the WCFZ height as the reference sequence and the influencing factors as comparison sequences, the grey relational analysis results under two coal-bearing stratigraphic conditions are presented in Fig. 8. In the Carboniferous-Permian coal seams, the ranking of correlation degrees among the factors is as follows: mining height (0.89) > working face slope length (0.87) > mining depth (0.86) > hard rock ratio coefficient (0.85) > mining method (0.67). All factors exhibited grey relational degrees above 0.65, indicating a strong association with WCFZ height. In the Jurassic coal seams, the order of correlation degrees is mining height (0.77) > hard rock ratio coefficient (0.71) > mining depth (0.69) > working face slope length (0.66) > mining method (0.52). Except for the mining method, which displayed a moderate association, the other factors maintained relatively strong correlations with the WCFZ height.

It is noteworthy that the degrees of association between the influencing factors and the WCZF height are significantly stronger in the Carboniferous-Permian strata compared to the Jurassic strata. It reveals that the development of WCFZ in the Carboniferous-Permian and Jurassic coal seams follows distinctly different controlling mechanisms. In the Carboniferous-Permian coal seams, the development of the WCFZ is a relatively regular and controlled process. The overburden consists of hard, well-layered marine-terrestrial transitional deposits, whose mechanical behavior more closely resembles that of a continuous medium[36]. In such a system, the transmission of influence from factors to the WCFZ height is direct and highly predictable. Therefore,

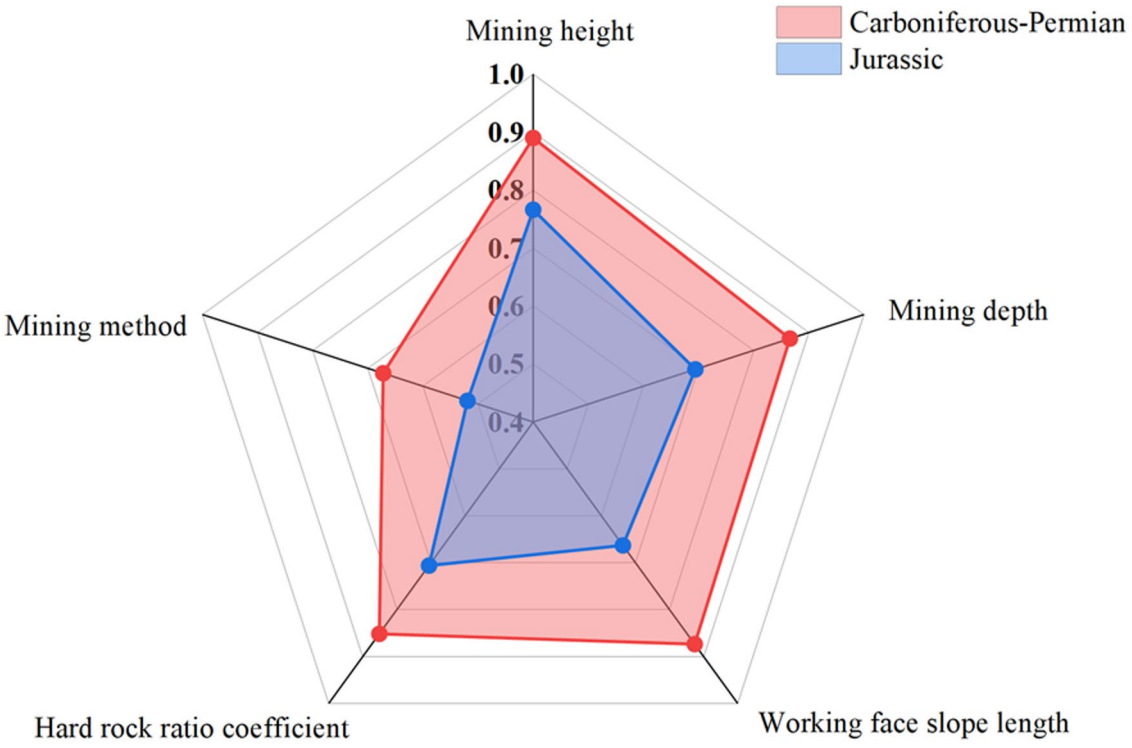

Fig. 8. Grey relational degrees of influencing factors with the WCFZ height.

Coal seam	model	R^2	RMSE	MAE	MAPE
Carboniferous-Permian	BO-RFR	0.895 ± 0.010	4.331 ± 0.198	3.256 ± 0.145	0.070 ± 0.004
	BO-XGBoost	0.885 ± 0.015	4.518 ± 0.296	3.556 ± 0.274	0.078 ± 0.007
	BO-GBDT	0.867 ± 0.023	4.843 ± 0.421	3.609 ± 0.352	0.078 ± 0.009
	BO-Adaboost	0.817 ± 0.010	5.709 ± 0.150	4.261 ± 0.170	0.094 ± 0.004
Jurassic	BO-RFR	0.924 ± 0.014	10.659 ± 1.004	8.793 ± 0.867	0.089 ± 0.009
	BO-XGBoost	0.868 ± 0.057	13.850 ± 2.891	10.239 ± 2.167	0.089 ± 0.021
	BO-GBDT	0.902 ± 0.020	12.101 ± 1.237	9.057 ± 0.965	0.084 ± 0.011
	BO-Adaboost	0.845 ± 0.010	15.322 ± 0.523	13.225 ± 0.745	0.129 ± 0.007

Table 2. Comparison of model performance based on 20 repetitions of 5-fold cross-validation (mean ± standard deviation).

changes in individual factors can accurately reflect changes in the outcome, resulting in generally high grey relational degrees. In contrast, the synchrony of trends between factors and WCFZ height in the Jurassic coal seams is weaker. This highlights the high complexity and inherent instability in the development of WCFZ within these strata. The overburden here comprises weak, homogeneous continental deposits, more akin to loose media or fractured rock masses, where system failure exhibits abrupt and nonlinear characteristics[8]. Minor variations in factors such as the working face slope length may cause qualitative changes in failure mode, thereby disrupting the simple synchronous relationship between WCFZ height and any single factor.

The combined results from Pearson correlation coefficients and grey relational analysis indicate a strong linear relationship between mining height and the development height of WCFZ. Additionally, mining depth, working face slope length, hard rock ratio coefficient, and mining method also show strong nonlinear correlations with WCFZ height. Consequently, selecting these five factors for predicting the height of WCFZ is appropriate.

Results of model prediction

Evaluation of model robustness based on repeated cross-validation

To ensure the stability of model predictions and quantify performance variability under small-sample conditions, this study employed a repeated 5-fold cross-validation framework for model evaluation and comparison. Table 2 reports the mean and standard deviation of performance metrics for the BO-RFR model and comparative models across 20 repetitions of 5-fold cross-validation (Comprehensive results for various coal seam conditions are detailed in Supplementary Information 1). As shown in Table 2, the BO-RFR model attained the highest

No.	Random forest hyperparameters	Ranges
1	n_estimators	(1,500)
2	max_depth	(1,20)
3	max_features	(1,5)
4	min_samples_split	(2,10)
5	min_samples_leaf	(1,10)

Table 3. Hyperparameter combinations and their ranges for random forest optimization.

Coal seam	n_estimators	max_depth	max_features	min_samples_split	min_samples_leaf
Carboniferous-Permian	72	17	3	2	1
Jurassic	81	10	3	3	1

Table 4. Optimal hyperparameter combinations selected by the BO algorithm.

mean R^2 and the lowest mean error, with the smallest or near-smallest standard deviations among all models. These results demonstrate that BO-RFR not only achieves superior predictive accuracy but also exhibits minimal sensitivity to data partition randomness, reflecting optimal robustness and confirming its reliability as a robust predictive tool.

Comparison of predictive accuracy of different models on the test set

This study developed RFR models to predict the height of the WCFZ based on sample data from the Carboniferous-Permian and Jurassic coal seams, respectively. To improve the accuracy and stability of the models, BO was applied to optimize the RFR hyperparameters. The hyperparameter combinations and their respective ranges considered for optimization are listed in Table 3, encompassing five parameters, including the number of decision trees and maximum tree depth. The optimal hyperparameter sets identified by the BO algorithm are presented in Table 4.

In the optimal hyperparameter combinations obtained through the BO algorithm, the max_depth for the Carboniferous-Permian model (17) is significantly greater than that of the Jurassic model (10). This difference corresponds to the much more complex lithological composition and mechanical structure of the overburden in the former. The Carboniferous-Permian overburden is typically characterized by marine-terrestrial transitional deposits, featuring a complex and variable lithological assemblage. The distribution of hard rock layers plays a significant controlling role in fracture development, leading to a pronounced nonlinear and complex interaction between the height of the WCFZ and influencing factors. A deeper tree can better capture the intricate coupling relationships between the hard rock ratio coefficient, mining height, and other relevant factors.

In contrast, the Jurassic coal seam overburden predominantly consists of continental clastic rocks, primarily medium-hard to weak sandstones and mudstones, with relatively simple stratification and less pronounced mechanical property contrasts. Consequently, the development pattern of the WCFZ height is mainly governed by macroscopic factors such as mining height and mining depth, not requiring an overly complex model architecture. Similarly, the Jurassic model requires more trees, which may be attributed to a greater coefficient of variation in its data: the WCFZ height distribution is more dispersed under shallow burial depths and soft rock conditions (see Fig. 3). Increasing the number of trees aids variance reduction through ensemble averaging, thus enhancing prediction stability. Conversely, the Carboniferous-Permian data distribution is relatively concentrated, requiring slightly fewer trees.

The optimal hyperparameter combinations obtained through BO were incorporated into the RFR model, resulting in the BO-RFR model for predicting the height of the WCFZ. Two BO-RFR models were then trained separately using 117 Carboniferous-Permian and 88 Jurassic WCFZ height samples, producing two optimized models. The training results, shown in Fig. 9, demonstrate close agreement between predicted and observed values. Specifically, Fig. 9(a) presents the Carboniferous-Permian training results with an R^2 of 0.955 and an *RMSE* of 3.344; Fig. 9(b) shows the Jurassic results with an R^2 of 0.942 and an *RMSE* of 10.436. As illustrated, the BO-RFR model tends to overestimate height for samples with relatively low observed values and underestimate height for those with higher observed values.

Following the completion of model training, the predictive performance of the BO-RFR model was assessed using test samples for the height of the WCFZ. The predictions from the BO-RFR model were compared with those from the RFR, BO-XGBoost, BO-GBDT, and BO-AdaBoost models. Prediction results for the Carboniferous-Permian and Jurassic coal seams are presented in Figs. 10 and 11, respectively, with corresponding evaluation metrics summarized in Table 5. As shown in the figure, the BO-RFR model achieves the highest R^2 among all machine learning models, while exhibiting the lowest values for the three error metrics—*RMSE*, *MAE*, and *MAPE*. This indicates that the predictive performance of the BO-RFR model is significantly superior to that of the other models. Furthermore, BO substantially enhanced the RFR model's predictive capability. For the Carboniferous-Permian coal seam, the BO-RFR model improved R^2 from 0.876 to 0.905, while *RMSE*, *MAE*, and *MAPE* decreased from 4.703, 3.347, and 0.072 to 4.122, 3.196, and 0.067, respectively. This improvement was

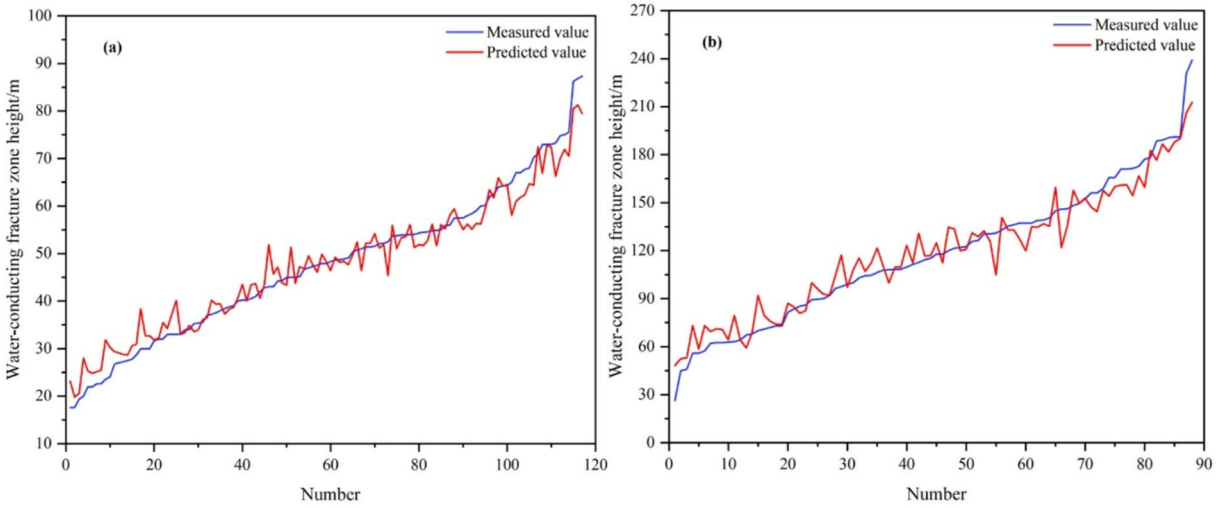

Fig. 9. BO-RFR model training results.

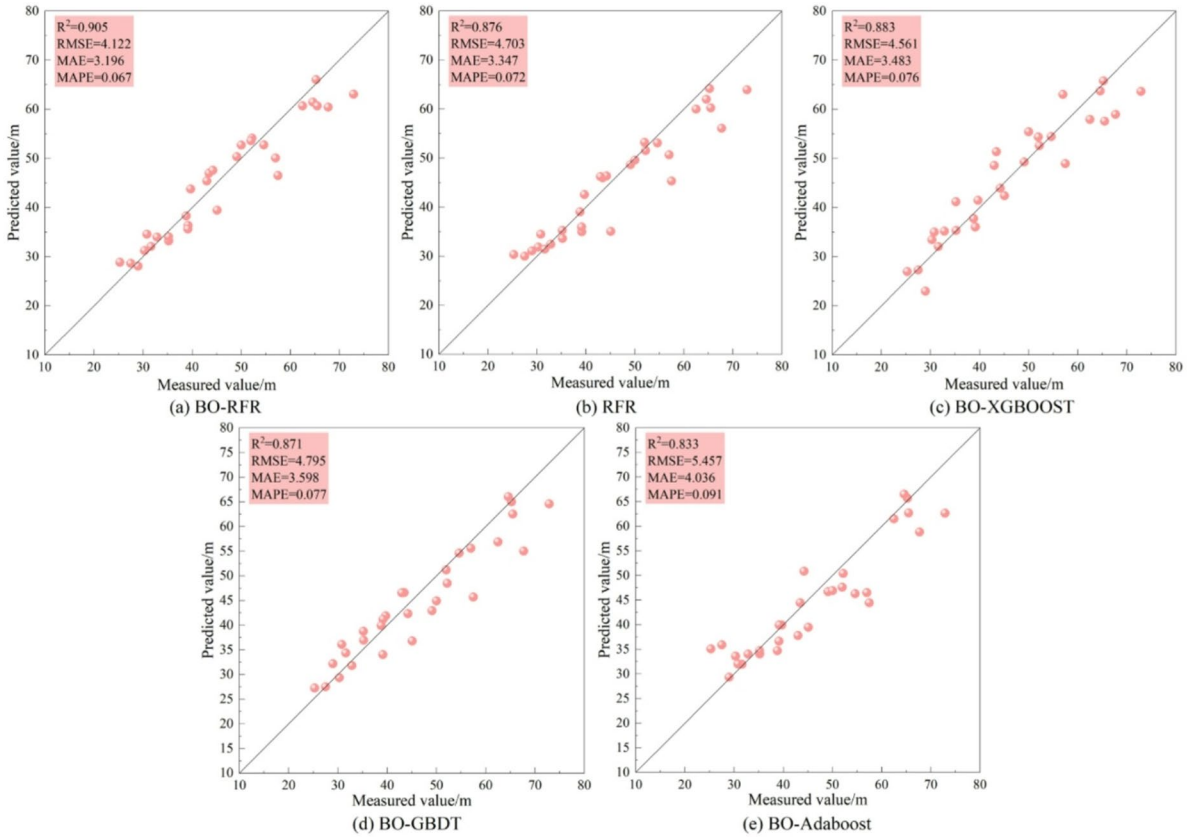

Fig. 10. Comparison of predicted WCFZ heights for Carboniferous-Permian coal seams using different models.

even more pronounced for the Jurassic coal seam, where R^2 increased from 0.925 to 0.949, and *RMSE*, *MAE*, and *MAPE* decreased from 10.594, 8.798, and 0.089 to 8.799, 7.198, and 0.073, respectively.

To provide a more intuitive illustration of the accuracy and applicability of the proposed BO-RFR model, its predictions were compared with results obtained from widely used empirical formulas for the height of WCFZ in the industry[14]. As indicated in Eqs. (15) and (16), the former and latter equations correspond to empirical formulas for WCFZ height under hard and medium-hard lithological conditions, respectively.

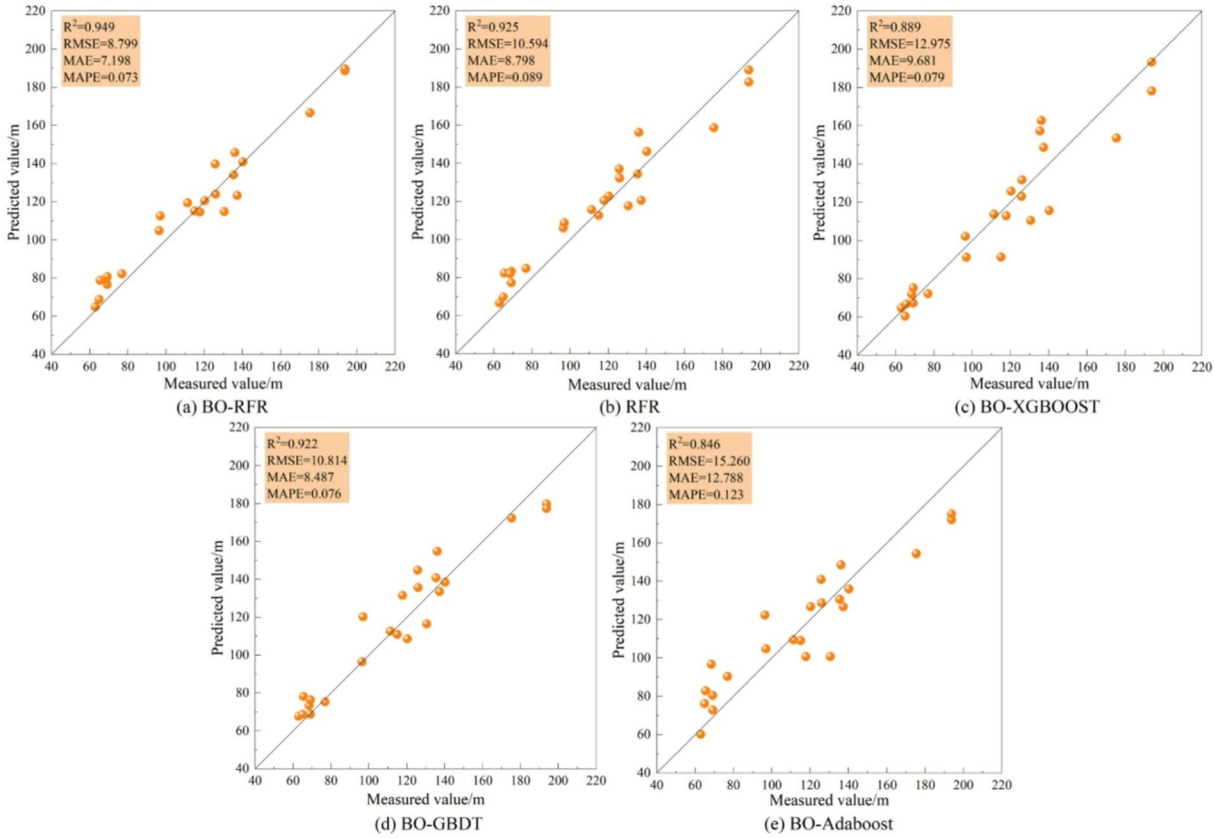

Fig. 11. Comparison of predicted WCFZ heights for Jurassic coal seams using different models.

Coal seam	Model	R^2	RMSE	MAE	MAPE
Carboniferous-Permian	BO-RFR	0.905	4.122	3.196	0.067
	RFR	0.876	4.703	3.347	0.072
	BO-XGBoost	0.883	4.561	3.483	0.076
	BO-GBDT	0.871	4.795	3.598	0.077
	BO-Adaboost	0.833	5.457	4.036	0.091
	Empirical formulas	0.640	21.093	19.460	0.484
Jurassic	BO-RFR	0.949	8.799	7.198	0.073
	RFR	0.925	10.594	8.798	0.089
	BO-XGBoost	0.878	13.557	11.276	0.099
	BO-GBDT	0.914	11.411	8.283	0.074
	BO-Adaboost	0.854	14.846	12.913	0.131
	Empirical formulas	0.642	63.351	55.502	0.450

Table 5. Comparison of evaluation metrics for different models.

$$H_W = 30\sqrt{\sum M} + 10 \tag{15}$$

$$H_W = 20\sqrt{\sum M} + 10 \tag{16}$$

Where ΣM represents the cumulative mining height.

The results of the empirical formulas are presented in Table 5. It is evident that there is a significant deviation in the results obtained from these empirical formulas when compared to those derived from machine learning algorithms. This discrepancy arises because the empirical formulas consider mining height as the sole input variable, leading to an oversimplification of the complex geomechanical issues involved. Additionally, the predictive errors for Jurassic coal seams ($RMSE = 63.351$) are significantly higher than those for Carboniferous-

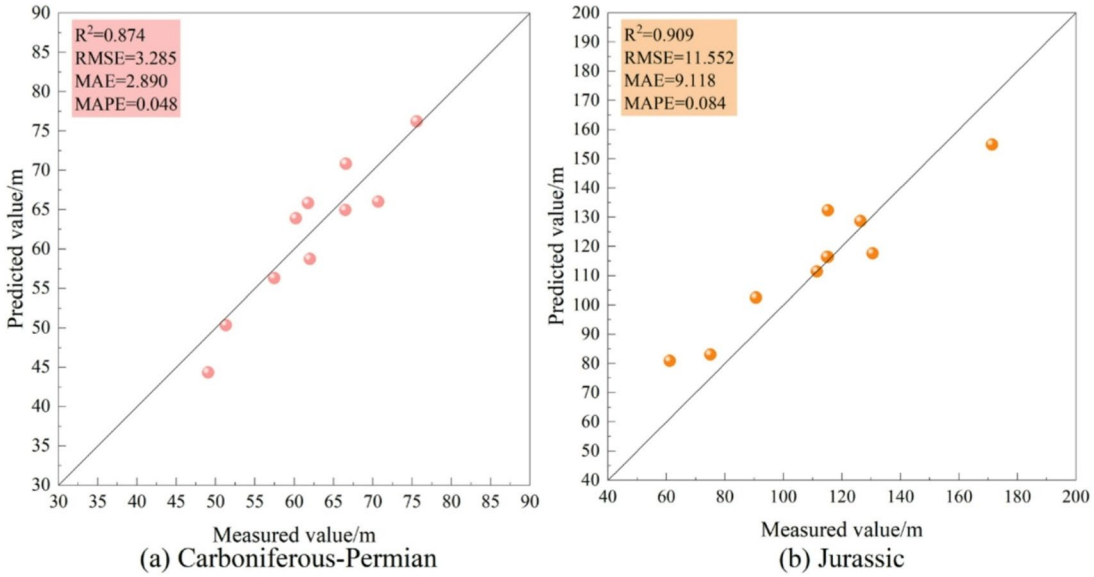

Fig. 12. Comparison between predicted and measured values in blind sample validation.

Coal seam	Sample size	R^2	RMSE	MAE	MAPE
Carboniferous-Permian	10	0.874	3.285	2.890	0.048
Jurassic	10	0.909	11.552	9.118	0.084

Table 6. Prediction performance of the BO-RFR model on the blind validation dataset.

Permian seams (*RMSE* = 21.093). This is attributable to the fact that the empirical formulas were developed based on data collected several decades ago from the Carboniferous-Permian mining regions in central and eastern China. As coal mining has expanded into the western regions and deeper deposits of China, both geological and mining conditions have undergone substantial changes. Therefore, the applicability of these empirical formulas for predicting the height of WCFZ in Jurassic coal seams is inherently limited.

Blind sample validation results

To further evaluate the adaptability of the BO-RFR model to unknown geological conditions, we applied it to a completely independent blind validation set that includes 10 samples each from the Carboniferous-Permian and Jurassic coal seams[37,38] (The complete blind validation dataset can be found in Supplementary Information 2). The prediction results are illustrated in Fig. 12, and the evaluation metrics are summarized in Table 6. Compared to the conventional test set (Table 5), the model exhibited a slight decrease in R^2 on the blind validation set, while the error metrics—*RMSE*, *MAE*, and *MAPE*—showed an increase. This is expected, as the geological backgrounds of the blind validation samples differ from those of the primary modeling dataset. Nevertheless, the model maintained high R^2 values (all exceeding 0.87) and acceptable absolute errors across both coal seam groups in the blind validation set. These findings provide strong evidence that the BO-RFR model developed in this study not only achieves high accuracy within the distribution of the training data but also demonstrates robust generalization capabilities across diverse regions and geological conditions.

Interpretability analysis based on SHAP

Feature impact effect analysis

To analyze the influence of individual features on the model's predictions, SHAP dependence plots were generated, as shown in Fig. 13. This figure illustrates the continuous relationship between the values of continuous numerical features—specifically, mining height, mining depth, working face slope length, and the hard rock ratio coefficient—and their contributions to predictions, as represented by SHAP values. Subplots (a)–(d) correspond to the Carboniferous-Permian formation, while subplots (e)–(h) represent the Jurassic formation. Each scatter point in the figure represents an individual sample, with the horizontal axis denoting feature values, and the vertical axis indicating the corresponding SHAP values. The scatter distribution visually illustrates how each feature's contribution to the BO-RFR model predictions varies with changes in feature values.

Overall, the SHAP value trends can be classified into two types: monotonically increasing and initially increasing before stabilizing. As shown in Figs. 13(a) and 13(e), the SHAP values exhibit a typical monotonically increasing pattern with mining height, indicating that greater mining heights significantly enhance the model output. Similar increasing trends are observed for mining depth in the Carboniferous-Permian formation and

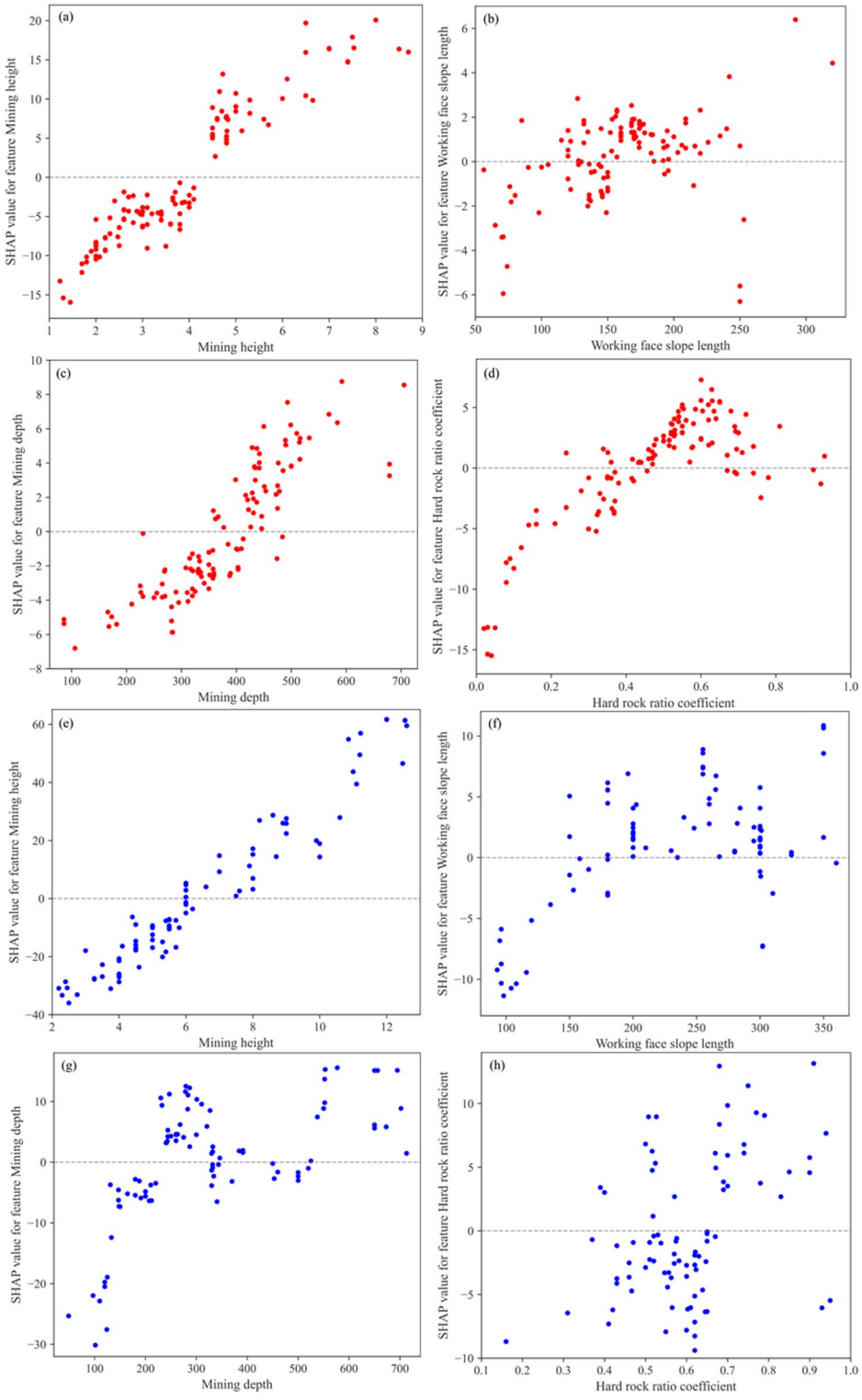

Fig. 13. SHAP feature dependence plots of the BO-RFR model.

the hard rock ratio coefficient in the Jurassic. Conversely, the hard rock ratio coefficient in the Carboniferous-Permian formation displays an initially increasing then stabilizing pattern. Specifically, SHAP values increase with the hard rock ratio coefficient up to a threshold of 0.6, beyond which they plateau. Comparable patterns are also observed for mining depth and working face slope length in the Jurassic formation. For the discrete categorical variable of mining method, the horizontal axis of the feature dependence plot (feature values) lacks a meaningful continuous order, rendering it unsuitable for the continuous representation described above.

Instead, the effect of this feature is characterized through subsequent SHAP global feature importance analysis and SHAP value distribution plots.

Global feature importance analysis based on SHAP

To investigate the specific influence patterns of input features on the predicted height of WCFZ, the distributions of feature SHAP values were plotted, as shown in Fig. 14. Each point in the plot represents a sample, with its horizontal position indicating the direction and magnitude of the corresponding feature's contribution to the model prediction. The vertical axis lists features ordered by their global importance in descending order. The color gradient from blue to red represents the original feature values within each sample, where blue denotes low values and red denotes high values.

In both coal-bearing strata, mining height exhibits the largest absolute SHAP values and the widest distribution range, with most SHAP values being positive, indicating that this feature is the primary driver for elevating the predicted WCFZ height. For the Carboniferous-Permian formation, the SHAP values of the hard rock ratio coefficient mostly range between 0 and 5, suggesting that this feature generally contributes positively to the predicted WCFZ height in the BO-RFR model. Similarly, in the Jurassic formation, the SHAP values for mining depth predominantly range from −5 to 20, with high feature values (represented by red points) showing a noticeable right-skewed distribution. This trend implies that increasing mining depth significantly raises the predicted height of WCFZ. In contrast, SHAP values for mining depth in the Carboniferous-Permian formation are concentrated between −5 and 5, with points of varying colors distributed fairly evenly around zero. This pattern reflects a relatively stable but limited effect of mining depth on the model output. A similar trend is observed for the hard rock ratio coefficient in the Jurassic formation. For both coal-bearing strata, the SHAP values of working face slope length and mining method are primarily distributed narrowly around zero, indicating these features have minimal impact on the model predictions.

To quantitatively evaluate the overall average impact of each input feature on the model's prediction of WCFZ height, we computed the mean absolute SHAP value for each feature across the sample set. These values were then ranked in descending order as global importance indicators, as presented in Fig. 15. This metric reflects the average expected change in model prediction when a given feature is removed.

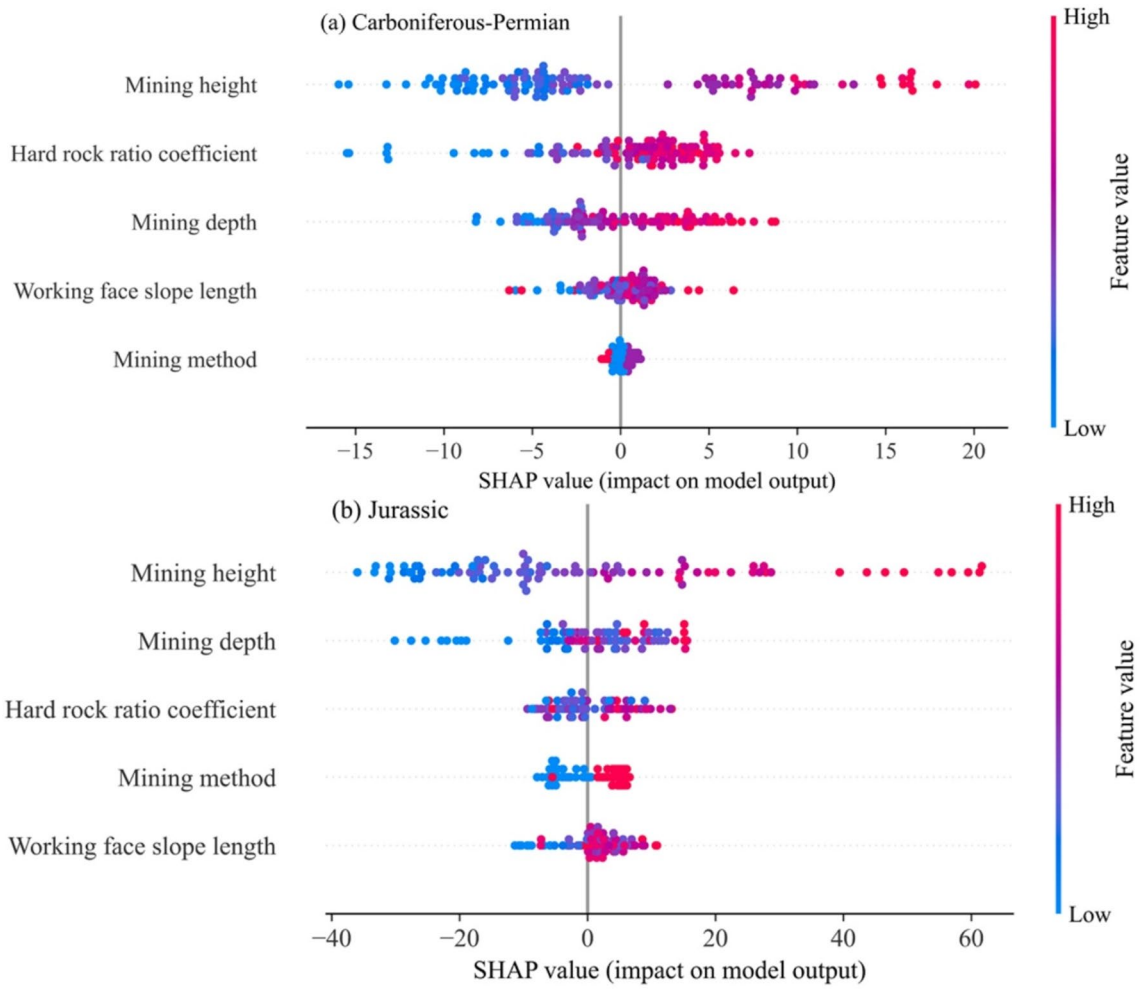

Fig. 14. Distribution of SHAP values for features in the BO-RFR model.

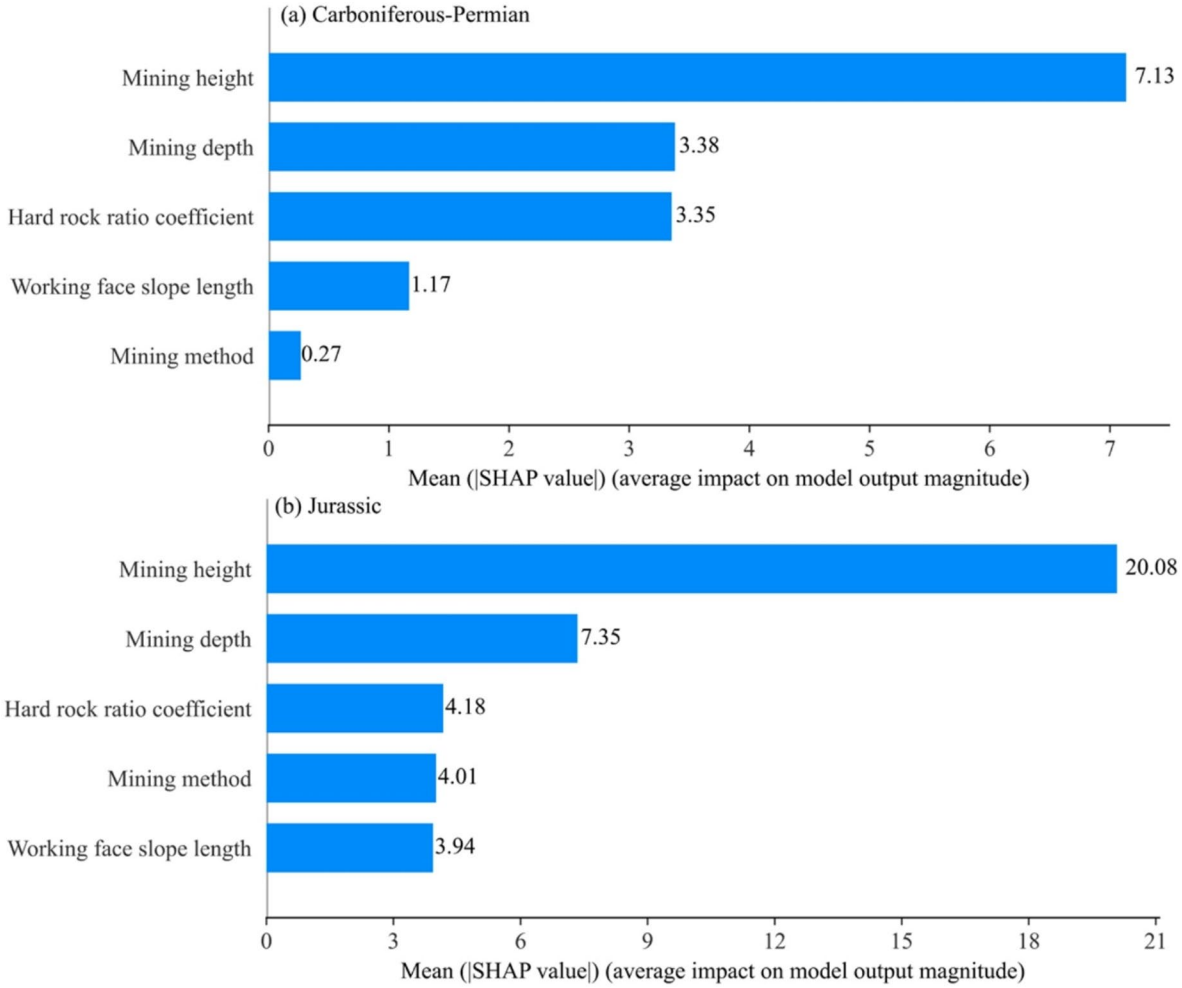

Fig. 15. Global feature importance ranking based on mean absolute SHAP values.

Mining height overwhelmingly ranks first in importance, significantly exceeding other features. Specifically, in the Carboniferous-Permian formation, the mean |SHAP| value of mining height is 2.3 times that of the second-ranked feature, the hard rock ratio coefficient. In the Jurassic formation, the mean |SHAP| of mining height is 2.7 times greater than that of mining depth. This provides robust data-driven evidence that, under the geological and mining conditions considered in this study, mining height is the most fundamental and consistent driving factor controlling the development of WCFZ. In the Carboniferous-Permian formation, the mean |SHAP| values for the hard rock ratio coefficient and mining depth are both relatively high and comparable, indicating that these two features are major factors influencing the model output with similar levels of importance. Conversely, in the Jurassic formation, mining depth has a mean |SHAP| value of 7.37, making it the most important feature after mining height. Other features play a secondary role in affecting model predictions.

Discussion

To comparatively investigate the development height of WCFZ under different coal seam stratigraphic conditions, this study collected 147 measured samples from Carboniferous-Permian strata and 111 from Jurassic strata. Five influencing factors—mining height, mining depth, working face slope length, hard rock ratio coefficient, and mining method—were selected for analysis. A BO-RFR model was then developed to predict the height of the WCFZ, and SHAP values were employed to quantitatively analyze the contribution magnitude and direction of various influencing factors. This study aims to systematically reveal the significant differences in the development mechanisms of WCFZ in Carboniferous-Permian and Jurassic coal seams based on the interpretable machine learning approach using SHAP.

SHAP values quantify the marginal contribution of each feature to the model's prediction, and their importance ranking essentially reveals the relative influence of various geological and engineering factors in controlling the development of WCFZ. Mining height consistently exhibits the highest SHAP importance across both coal-bearing strata. This is because mining height directly determines the scale of the caving zone in the goaf and the magnitude of mining-induced disturbance energy. This conclusion is fully consistent with the classical theory of mining-induced strata pressure.

In the Carboniferous-Permian formation, the SHAP importance of the hard rock ratio coefficient ranks second, exceeding that of mining depth. This reflects the complex lithological assemblage resulting from marine–terrestrial sedimentary interactions, where thick hard rock layers often constitute key strata that govern the failure behavior of overlying rocks. A higher hard rock ratio coefficient indicates more developed key strata with greater load-bearing capacity, effectively inhibiting the upward propagation of fractures. Consequently, the primary strategy for water control in Carboniferous-Permian coal seams is to investigate the structure and control mining height, which entails detailed characterization of hard key strata distributions and the application of modeling to determine safe critical mining heights under their influence, thereby optimizing working face design. By contrast, in the Jurassic strata, mining depth surpasses the hard rock ratio coefficient in SHAP importance, ranking second overall. This shift indicates a transition in controlling mechanisms under conditions of shallow burial and soft rock lithology. Here, mining depth predominantly represents the thickness of the overlying competent bedrock and the in-situ stress state. Its increased significance signifies that overall stratum stability and roof fall risk have supplanted key strata structural control as the dominant mechanical factor. This finding corroborates the prior data analysis based on the fracture-to-mining ratio presented in this study. As shown in Fig. 3, the Jurassic coal seam exhibits a significantly higher and more dispersed fracture-to-mining ratio, indicating that damage to the overlying rock mass responds more sensitively and variably to mining disturbances. In this context, mining depth—representing the overburden load and the stress state of the surrounding rock—emerges as a critical threshold factor governing the stability of this sensitive system. Therefore, the prominence of mining depth in the SHAP analysis is intrinsically consistent with the high damage efficiency and increased uncertainty characteristics revealed by the fracture-to-mining ratio. Accordingly, the principal strategy for Jurassic coal seams should shift to assess stress and prevent roof falls, considering mining depth—as an indicator of geostress and bedrock thickness—equally important to mining height in risk assessment. Models should be employed to identify high-risk areas at shallow depths and to critically evaluate the suitability of high-intensity mining techniques.

Across both coal-bearing formations, the SHAP importance of working face slope length and mining method is relatively low, consistent with engineering experience, as these represent operational parameters rather than intrinsic geological properties. Minor differences in their rankings—for instance, the greater influence of mining method compared to working face slope length in the Jurassic formation—may reflect variations in the degree to which different mining techniques disturb specific overlying rock structures.

Traditional hydrogeological studies have predominantly provided qualitative descriptions of the factors influencing WCFZ height, which are important but insufficient to quantify their relative contributions. SHAP values, however, offer the first precise quantification of each factor's contribution on a unified scale. For example, in the Jurassic formation, the mean absolute SHAP value of mining depth is approximately 1.61 times that of the hard rock ratio coefficient, clearly indicating that mining depth should be prioritized in exploration.

Notably, the ranking of influential factors obtained from the SHAP analysis in this study (Fig. 15) differs from the results of the Pearson correlation-based linear analysis (Table 1). For example, in the Jurassic coal seam, the linear correlation between working face slope length and WCFZ height is greater than that of the hard rock ratio coefficient; however, the SHAP importance of the former is lower than that of the latter. This discrepancy does not represent a contradiction but rather highlights differing analytical perspectives in interpreting complex systems. The Pearson correlation coefficient measures isolated linear relationships between two variables, whereas SHAP values—derived from the BO-RFR model—quantify the net marginal contribution of each feature to the predicted outcome within a nonlinear, multi-factor interaction context. These nonlinear interactions are particularly significant in the geological setting of shallow burial depth and soft rock strata characteristic of the Jurassic formation, where the hard rock ratio coefficient often exerts a critical yet nonlinear inhibitory effect on fracture development through threshold mechanisms. This effect is effectively captured by the BO-RFR model and emphasized in the SHAP analysis, which accounts for interactions among all considered factors. Prior grey relational analysis (Fig. 8) preliminarily identified strong nonlinear associations between various factors and the WCFZ height. The BO-RFR-SHAP framework developed herein further quantifies the actual weights of these factors from a causal contribution perspective within such complex relationships. Therefore, the observed differences between SHAP and Pearson outcomes fundamentally reflect an advancement from simple linear correlation to a mechanistic understanding of nonlinear system dynamics, underscoring the distinct advantages of integrating machine learning with interpretable analysis for addressing complexities in geological and engineering contexts.

This study not only delivers a highly accurate predictive tool but also contributes a mechanistically differentiated safety management methodology, offering clear practical value for mine water hazard prevention in China's major coal bases across the central-eastern and western regions. Moreover, it provides a scientific basis for refining relevant industry standards. Despite the strong predictive performance of the BO-RFR model for fracture zone height in the Carboniferous-Permian and Jurassic coal seams, some inherent limitations remain. The dataset, comprising 258 measured samples from representative typical mining areas in central-eastern and western China, presents good regional coverage. However, the samples are predominantly drawn from relatively mature large-scale mines, and coverage of geologically complex areas, such as highly intricate structural belts, is limited. While the model demonstrates robust generalization within the current dataset, its direct application to geologically distinct coal-bearing basins may result in performance degradation. We recommend conducting transfer learning or fine-tuning with a limited number of local measurements before applying the model to new regions to enhance geographic adaptability. Future research could incorporate additional factors such as the position and thickness of key strata, aquifer water pressure, volume, and enrichment characteristics to develop a coupled geology–mining–hydrogeology predictive framework that better reflects engineering realities.

Conclusions

This study systematically compares the development patterns of WCFZ height in Carboniferous-Permian and Jurassic coal seams, utilizing a BO-RFR model in conjunction with SHAP interpretability analysis, resulting in the following key conclusions:

(1) Compared to traditional empirical formulas and other machine learning models (RFR, BO-XGBoost, BO-GBDT, BO-Adaboost), the BO-RFR model achieves the highest predictive accuracy for WCFZ height in both coal seam types. This demonstrates the model's capability to effectively capture the complex nonlinear relationships between geological and mining factors, highlighting its strong applicability in engineering practice.

(2) SHAP analysis indicates that mining height consistently remains the most critical influencing factor. However, the importance rankings of the hard rock ratio coefficient and mining depth are reversed between the two coal seam types: in the Carboniferous-Permian formation, the hard rock ratio coefficient ranks as the second most important factor, whereas in the Jurassic formation, mining depth surpasses it in importance. This directly confirms, from a data-driven perspective, the differing dominant controlling mechanisms in each formation.

(3) For Carboniferous-Permian coal seams, water hazard prevention should follow the principle of investigating geological structures and controlling mining height, with a focus on detailed exploration of key strata. For Jurassic coal seams, the principle of assessing geostress and preventing roof falls should be applied, emphasizing geostress evaluation and overall stability. The developed BO-RFR-SHAP framework provides a reliable theoretical model and decision-support tool for the intelligent and precise prevention and control of mine water hazards.

Data availability

The data on the height of the water-conducting fracture zone and its influencing factors used in this study are available in Figshare with the identifier https://doi.org/10.6084/m9.figshare.30312373.

Received: 29 September 2025; Accepted: 28 January 2026
Published online: 04 February 2026

References

1. Gai, G. C., Qiu, M. & Shi, L. Q. Tectonic controls on the development of water-conducting fracture zones in the North China block. *Meas* **246**, 116726 (2025).
2. Qiu, M. et al. Water-richness evaluation method and application of clastic rock aquifer in mining seam roof. *Sci. Rep.* **14**, 6465 (2024).
3. Zheng, L. L. et al. Study of the development patterns of water-conducting fracture zones under karst aquifers and the mechanism of water inrush. *Sci. Rep.* **14**, 20790 (2024).
4. Lu, C. J., Xu, J. P., Li, Q., Zhao, H. & He, Y. Research on the development law of water-conducting fracture zone in the combined mining of jurassic and carboniferous coal seams. *Appl. Sci-Basel.* **12** (21), 11178 (2022).
5. Gao, W., Li, Y. C. & He, Q. Y. Determination of fractured water-conducting zone height based on microseismic monitoring: a case study in Weigiang Coalmine, Shaanxi, China. *Sustainability* **14** (14), 8385 (2022).
6. Feng, D. et al. Research on 3D development characteristics of water-conducting fractured zone based on field measurement. *Front. Earth Sci.* **10**, 808931 (2022).
7. Zhao, D. K. & Wu, Q. An approach to predict the height of fractured water-conducting zone of coal roof strata using random forest regression. *Sci. Rep.* **8**, 10986 (2018).
8. Guo, Q. B. et al. Study on numerical simulation of overburden fracture development characteristics and prediction of water-conducting fracture zone height in shallow coal seam mining. *Geofluids* **2025** (1), 8283522 (2025).
9. Long, T. W., Hou, E. K., Xie, X. S., Fan, Z. G. & Tan, E. M. Study on the damage characteristics of overburden of mining roof in deeply buried coal seam. *Sci. Rep.* **12**, 11141 (2022).
10. Wang, H. Z., Zhu, J. Z. & Li, W. P. An improved back propagation neural network based on differential evolution and grey wolf optimizer and its application in the height prediction of water-conducting fracture zone. *Appl. Sci-Basel.* **14** (11), 4509 (2024).
11. Wu, Z. Y., Chen, Y. & Luo, D. Y. Comparative study of five machine learning algorithms on prediction of the height of the water-conducting fractured zone in undersea mining. *Sci. Rep.* **14** (1), 21047 (2024).
12. Gao, Z. Y., Jin, L. X., Liu, P. T. & Wei, J. J. Height prediction of water-conducting fracture zone in jurassic coalfield of Ordos basin based on improved radial movement optimization algorithm Back-Propagation neural network. *Math* **12** (10), 1602 (2024).
13. Zhu, Z. J. & Guan, S. S. Prediction of the height of fractured water-conducting zone based on the improved cuckoo search algorithm-Extreme learning machine model. *Front. Earth Sci.* **10**, 860507 (2022).
14. Xu, C., Zhou, K. P., Xiong, X., Gao, F. & Zhou, J. Research on height prediction of water-conducting fracture zone in coal mining based on intelligent algorithm combined with extreme boosting machine. *Expert Syst. Appl.* **249**, 123669 (2024).
15. Yang, P. et al. Predicting the height of the water-conducting fractured zone based on a multiple regression model and information entropy in the Northern Ordos Basin, China. *Mine Water Environ.* **41** (1), 225–236 (2022).
16. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
17. Liu, Y. & Zhao, H. Variable importance-weighted random forests. *Quant. Biol. (Beijing China).* **5**, 338–351 (2017).
18. Ma, W., Lin, G. & Liang, J. L. Estimating dynamics of central hardwood forests using random forests. *Ecol. Modell.* **419**, 108947 (2020).
19. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genom* **99**, 323–329 (2012).
20. Hill, A. J., Herman, G. R. & Schumacher, R. S. Forecasting severe weather with random forests. *Mon Weather Rev.* **148**, 2135–2161 (2020).
21. Li, L. H., Jing, W. P. & Wang, H. H. Extracting the forest type from remote sensing images by random forest. *IEEE Sens. J.* **21** (16), 17447–17454 (2021).
22. Ji, H. L., Qi, L. L., Lyu, M., Lai, Y. H. & Dong, Z. Improved bayesian optimization framework for inverse thermal conductivity based on transient plane source method. *Entropy* **25** (4), 575 (2023).
23. Vincent, A. M. & Jidesh, P. An improved hyperparameter optimization framework for automl systems using evolutionary algorithms. *Sci. Rep.* **13**, 4737 (2023).
24. Kang, Z. R., Yang, D. M. & Shen, P. F. Prediction correction modeling of water-conducting fracture zones height due to repeated mining in close distance coal seams. *Sci. Rep.* **14** (1), 31611 (2024).
25. Yin, H. Y. et al. Height prediction and 3D visualization of mining-induced water-conducting fracture zone in Western Ordos basin based on a multi-factor regression analysis. *Energies* **15** (11), 3850 (2022).

26. Feng, D., Hou, E. K., Xie, X. S. & Hou, P. F. Research on water-conducting fractured zone height under the condition of large mining height in Yushen mining area, China. *Lithosphere* **2023** (1), 8918348 (2023).
27. Karabadji, N. E. et al. Accuracy and diversity-aware multi-objective approach for random forest construction. *Expert Syst. Appl.* **225**, 120138 (2023).
28. Ma, J. J., Pan, Q. & Guo, Y. N. Depth-first random forests with improved Grassberger entropy for small object detection. *Eng. Appl. Artif. Intell.* **114**, 105138 (2022).
29. Hwang, S. W. et al. Feature importance measures from random forest regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar. *J. Wood Sci.* **69**, 1 (2023).
30. Hebbal, A., Balesdent, M., Brevault, L., Melab, N. & Talbi, E. G. Deep Gaussian process for multi-objective bayesian optimization. *Optim. Eng.* **24**, 1809–1848 (2023).
31. Wang, X. L., Jin, Y. C., Schmitt, S. & Olhofer, M. Recent advances in bayesian optimization. *ACM Comput. Surv.* **55** (13s), 1–36 (2023).
32. Choi, J. E., Shin, J. W. & Shin, D. W. Vector SHAP values for machine learning time series forecasting. *J. Forecast.* **44** (2), 635–645 (2025).
33. Lundberg, S. M. et al. From local explanations to global Understanding with explainable AI for trees. *Nat. Mach. Intel.* **2** (1), 56–67 (2020).
34. Mukaka, M. M. & Statistics Corner A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **24** (3), 69–71 (2012).
35. Zheng, K. et al. Robust grey relational analysis-based accuracy evaluation method. *Appl. Sci-Basel.* **15** (9), 4926 (2025).
36. Chen, D. X., Sun, C. & Wang, L. G. Collapse behavior and control of hard roofs in steeply inclined coal seams. *Bull. Eng. Geol. Environ.* **80** (2), 1489–1505 (2021).
37. Liu, Y., Yuan, S. C., Yang, B. B., Liu, J. W. & Ye, Z. Y. Predicting the height of the water-conducting fractured zone using multiple regression analysis and GIS. *Environ. Earth Sci.* **78** (14), 422 (2019).
38. He, X., Zhao, Y. X., Zhang, C. & Han, P. H. A model to estimate the height of the water-conducting fracture zone for Longwall panels in Western China. *Mine Water Environ.* **39** (4), 823–838 (2020).

Acknowledgements

This study was supported by the Natural Science Foundation of Shandong Province (No. ZR2022QD060; No. ZR2020KE023; and No. ZR2021MD057). We thank anonymous reviewers and the editor for their effort to review this manuscript.

Author contributions

M.Q.: Conceptualization, Review. Y.W.: Methodology, Writing—original draft. C.T.: Validation, Investigation. M.S.: Data compilation.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-38043-3.

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.