scientific reports

OPEN

A machine learning framework for predicting shear strength properties of rock materials

Daxing Lei[1,2✉], Yaoping Zhang[1,2], Zhigang Lu[1,2], Guangli Wang[1], Zejin Lai[1,2], Min Lin[3] & Yifan Chen[4]

The shear strength characteristics of rock materials, specifically internal friction angle and cohesion, are critical parameters for the design of rock structures. Accurate strength prediction can significantly reduce design time and costs while minimizing material waste associated with extensive physical testing. This paper utilizes experimental data from rock samples in the Himalayas to develop a novel machine learning model that combines the improved sparrow search algorithm (ISSA) with Extreme Gradient Boosting (XGBoost), referred to as the ISSA-XGBoost model, for predicting the shear strength characteristics of rock materials. To train and validate the proposed model, a dataset comprising 199 rock measurements and six input variables was employed. The ISSA-XGBoost model was benchmarked against other models, and feature importance analysis was conducted. The results demonstrate that the ISSA-XGBoost model outperforms the alternatives in both training and test datasets, showcasing superior predictive accuracy ($R^2 = 0.982$ for cohesion and $R^2 = 0.932$ for internal friction angle). Feature importance analysis revealed that uniaxial compressive strength has the greatest influence on cohesion, followed by P-wave velocity, while density exerts the most significant impact on internal friction angle, also followed by P-wave velocity.

Keywords Rock materials, Machine learning, Shear strength, Internal friction angle, Cohesion, Extreme gradient boosting (XGBoost)

The mechanical properties of rock materials are a fundamental focus for geotechnical engineers and geologists involved in rock engineering[1–3]. Among these properties, the internal friction angle (φ) and cohesion (c) are the primary parameters defining the shear strength of rock materials[4,5]. They are critical for the rational design and safe operation of engineering applications, including rock slopes, underground chambers, and foundations (see Fig. 1). For example, using the wrong internal friction angle and cohesion can lead to overestimation or underestimation of the failure probability of rock materials[6]. Therefore, reasonable and accurate prediction of shear strength characteristics is helpful to reduce the construction risk and provide sufficient countermeasures for the design of engineering[7–9].

Traditionally, cohesion and internal friction angle are obtained by testing standard specimens using triaxial compression equipment[10]. This method is widely regarded as the most accurate and is universally accepted. Over the years, numerous experts have significantly advanced the understanding of cohesion and internal friction angle in rock materials[11–13]. Through extensive experiments, numerical analyses, and theoretical approaches, researchers have explored the mechanical behavior of rock materials, resulting in a wealth of published studies[14–21]. For instance, Gu et al.[22] investigated the evolution of mechanical parameters such as deformation modulus, internal friction angle and cohesion of Shuangjiangkou granite under different stress paths. Similarly, Hashiba and Fukui[23] examined the loading-rate dependence of force and internal friction angle from a small amount of rock sample. However, experimentalists often face safety risks from close observation, and test results are susceptible to biases caused by unfavorable field conditions. In practice, due to the associated cost and time requirements (such as the early stages of a project), geotechnical engineers need to evaluate φ and c without triaxial test results[24,25]. To address these limitations, researchers have sought to develop inexpensive and efficient indirect estimation methods. Many researchers[26–33] have explored the use of parameters such as uniaxial compressive strength (UCS), uniaxial tensile strength (UTS), sound velocity, gamma-ray data, and porosity to estimate φ and

[1]School of Resources and Civil Engineering, Gannan University of Science and Technology, Ganzhou 341000, China. [2]Key Laboratory of Intelligent and Green Development of Tungsten & Rare Earth Resources, Provincial Department of Education, Ganzhou 341000, Jiangxi, China. [3]Anhui Lujiang Longqiao Mining Co., LTD, Hefei 230000, China. [4]School of Resources and Safety Engineering, Central South University, Changsha 410083, Hunan, China. ✉email: daxinglei17@163.com
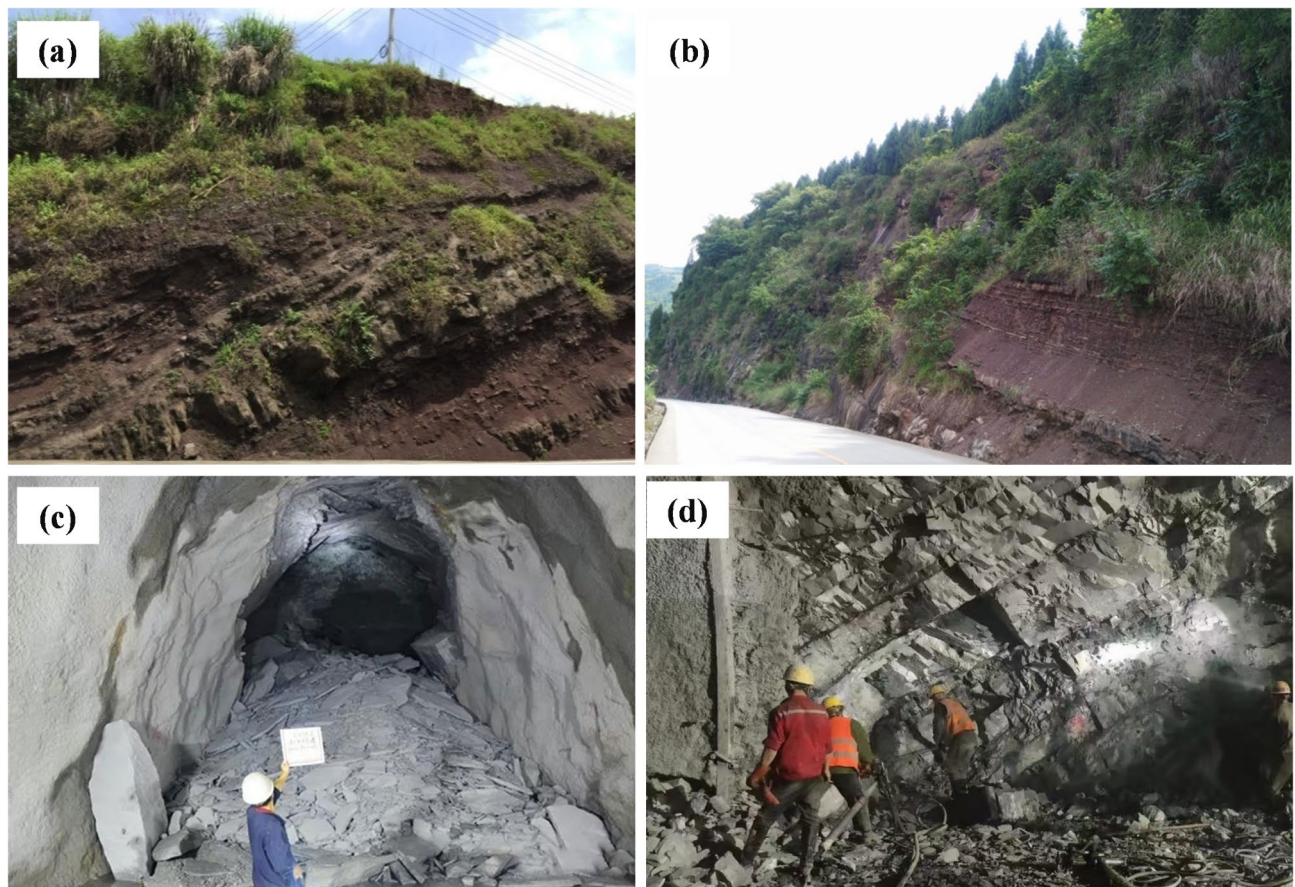
Fig. 1. Photograph of (**a**) slope of mining surface, (**b**) slope of mountainous highway, (**c**) underground cavern rock engineering, (**d**) underground mining engineering in China, (photograph by Daxing Lei).

c in the absence of triaxial test results. Nevertheless, the complex and highly nonlinear relationships between these influencing factors and strength characteristics pose significant challenges. Traditional regression models, constrained by their reliance on fitting methods, can only account for a limited number of variables and fail to capture the intricate interdependencies among multiple factors.

Machine learning (ML) technologies offer high efficiency and precision, making them well-suited for analyzing the nonlinear and complex relationships among multiple input parameters[34–39]. ML-based predictive models not only substantially reduce experimental workload but also outperform regression methods in handling regression problems with significantly higher accuracy[40–44]. The application of ML in modeling complex problems has been extensively validated[45–49]. For example, based on the genetic programming (GP), Shen and Jimenez[24] applied GP to predict the internal friction angle and cohesion of sandstone in the absence of triaxial data. The results show that the proposed ML model can provide good prediction performance. Taking the P-wave velocity of rock samples as the input parameter, Kainthola et al.[15] applied linear regression analysis and adaptive neurofuzzy inference system (ANFIS) technology to establish an ML model of rock materials. Similarly, Hiba et al.[50] employed a neural network model to predict φ and c from the logging data of two existing wells. Their study also included sensitivity analyses for three input parameters—neutron porosity (NPHI), compressional time (DTC), and bulk density (ROHB)—to assess their relative importance. Table 1 summarizes studies on the initial applications of ML methods for predicting cohesion and internal friction angle. Nonetheless, ML-based methods have not yet been widely used to predict cohesion and internal friction angle[51]. In limited research, scholars have adopted ML technologies such as artificial neural network (ANN), particle swarm optimization (PSO) and ANFIS to establish ML models and preliminarily prove their feasibility. However, existing research has mainly used some straightforward and ML algorithms, while the applicability of more advanced algorithms, such as the integrated algorithms of XGBoost and improved sparrow search algorithm (ISSA), in evaluating internal friction angle and cohesion has not been explored.

Using the triaxial test data from the Himalayan region, this study proposes a novel ML model combining the ISSA with XGBoost to predict internal friction angle and cohesion. The proposed model's performance was evaluated by comparison with the XGBoost model (without ISSA tuning) and four other ML models. Furthermore, a feature importance analysis was conducted to support geotechnical engineers with limited ML expertise in interpreting the results. This study aims to provide a more efficient and reliable prediction of internal friction angle and cohesion, which is also the key to improve the quality of related building design.

Model	Parameters	Rock materials	Performance
ANN[52]	Drilling rate of penetration ROP; Weight on bit WOB; Drill pipe pressure SPP; Torque; Drilling fluid pumping rate	/	Correlation coefficient (R) values were 0.85 and 0.89 for friction angle and cohesion respectively
ANN[50]	Bulk density (ROHB), compressional time (DTC), and neutron porosity (NPHI)	Carbonate	Average absolute percentage error (AAPE) = 1.1% for the friction angle and AAPE = 2.4% for the cohesion.
Genetic algorithm (GA)-ANN[53]	P-wave velocity, uniaxial compressive strength and brazilian tensile strength	Limestone	R^2 0.967 (c)
GP[24]	UCS, UTS, σ_3	Sandstone	/
ANFIS[54]	P-wave velocity	Limestone, quartzite, slate and quartz mica schist.	/
ANFIS-PSO[55]	Drilling specific energy (DSE) features	Dolomite, shale, marl	R^2 0.8724 (c) R^2 0.8142 (φ)
ANFIS- GA[55]	Drilling specific energy (DSE) features	Dolomite, shale, marl	R^2 0.8154 (c) R^2 0.6914 (φ)
Support vector machine (SVM)[56]	P-wave velocity, Density, UCS, UTS	Limestone, quartzite, slate and quartz mica schist.	: ($R^2 = 0.977$) and ($R^2 = 0.916$)
Lasso regression (LR)[56]	P-wave velocity, Density, UCS, UTS	Limestone, quartzite, slate and quartz mica schist.	: $R^2 = 0.928$ and : $R^2 = 0.606$
Ridge regression (RR)[56]	P-wave velocity, Density, UCS, UTS	Limestone, quartzite, slate and quartz mica schist.	: $R^2 = 0.961$ and : $R^2 = 0.822$
Decision tree (DT)[56]	P-wave velocity, Density, UCS, UTS	Limestone, quartzite, slate and quartz mica schist.	: $R^2 = 0.934$ and : $R^2 = 0.607$

Table 1. ML models for the estimation of cohesion and internal friction angle.

Dataset and preprocessing
Data collection
The Himalayas are one of the youngest structurally active complex geological chains, and the physical and mechanical properties of the rock materials there show high uncertainty. A large number of rock engineering activities have been developed in the Himalayan region. Therefore, it is necessary to study the cohesion and internal friction angle of rock materials in this area to guide the design of rock structures.

This paper collected experimental results from the literature[54] as the dataset for developing ML models. As presented in Fig. 2, the rock materials tested in this dataset include limestone, quartzite, slate, and quartz mica schist collected from the Luhri area, Himanchal Pradesh, India. A total of 597 rock samples underwent tests, including uniaxial compression, tensile strength, triaxial testing, and longitudinal wave velocity measurements. The average of the three test results was analyzed as a single value, and finally 199 sets of valid data were obtained. Further details of these tests and rock samples can be found in the literature[54]. Previous study[57] has demonstrated that incorporating four key mechanical properties—P-wave velocity, density, uniaxial compressive strength, and tensile strength (TS)—as input variables significantly enhances the predictive performance of ML models for shear strength parameters. Accordingly, these four properties were used as input variables in this study, with the shear strength parameters of rock materials (φ and c) designated as output variables.

Data preprocessing
In general, the application of ML modeling requires data analysis and pre-processing. The statistical analysis of input and output variables in the dataset is shown in Fig. 3 and Table 2. The box diagram provides a comprehensive visual analysis of the input and output parameters (P-wave velocity, density, UCS, TS, c, φ) of the four rock types included in the dataset, showing the corresponding distribution characteristics. Table 2 complements this with detailed statistical metrics for each parameter. The analysis reveals that the six input and output variables across the four rock materials cover a wide value range, with no significant outliers detected.

The preprocessing of dataset mainly consists of normalized and segmented datasets. Normalization is the primary process for standardizing data in a dataset[58]. This step guarantees that all input variables are treated fairly when modeling and that no one parameter is overestimated or underestimated. Similar to unifying measurements of length—converting yards, inches, or feet to the standard unit of meters—normalization scales all input variables to a common range. This process enables the ML model to more effectively and efficiently capture complex nonlinear relationships. Normalization methods commonly used in ML include the minimum-maximum normalization method[59], the Z-score normalization method[60] and the robust scaling method[61]. In this paper, the minimum-maximum normalization method is used to re-scale the data to a range between 0 and 1. The mathematical formula for this method is provided in Eq. 1. The impact of alternative normalization techniques on model predictions is further investigated in the following section.

$$\bar{X} = \frac{X - X_{min}}{X_{max} - X_{\min}}$$

(1)

where X and $\bar{X}$ represent the original and normalized values respectively. X_{max} and X_{min} represent the maximum and minimum values of the original values, respectively.

Following is the data splitting of dataset. As a popular method for model validation, data splitting method randomly divides datasets into training set and test set[62]. The ML model is used on the training set to learn the training, by separating another part of the dataset independent of the training process (i.e., the test set) for validation[63]. Currently, there is no industry standard or specification for data segmentation ratio. Common
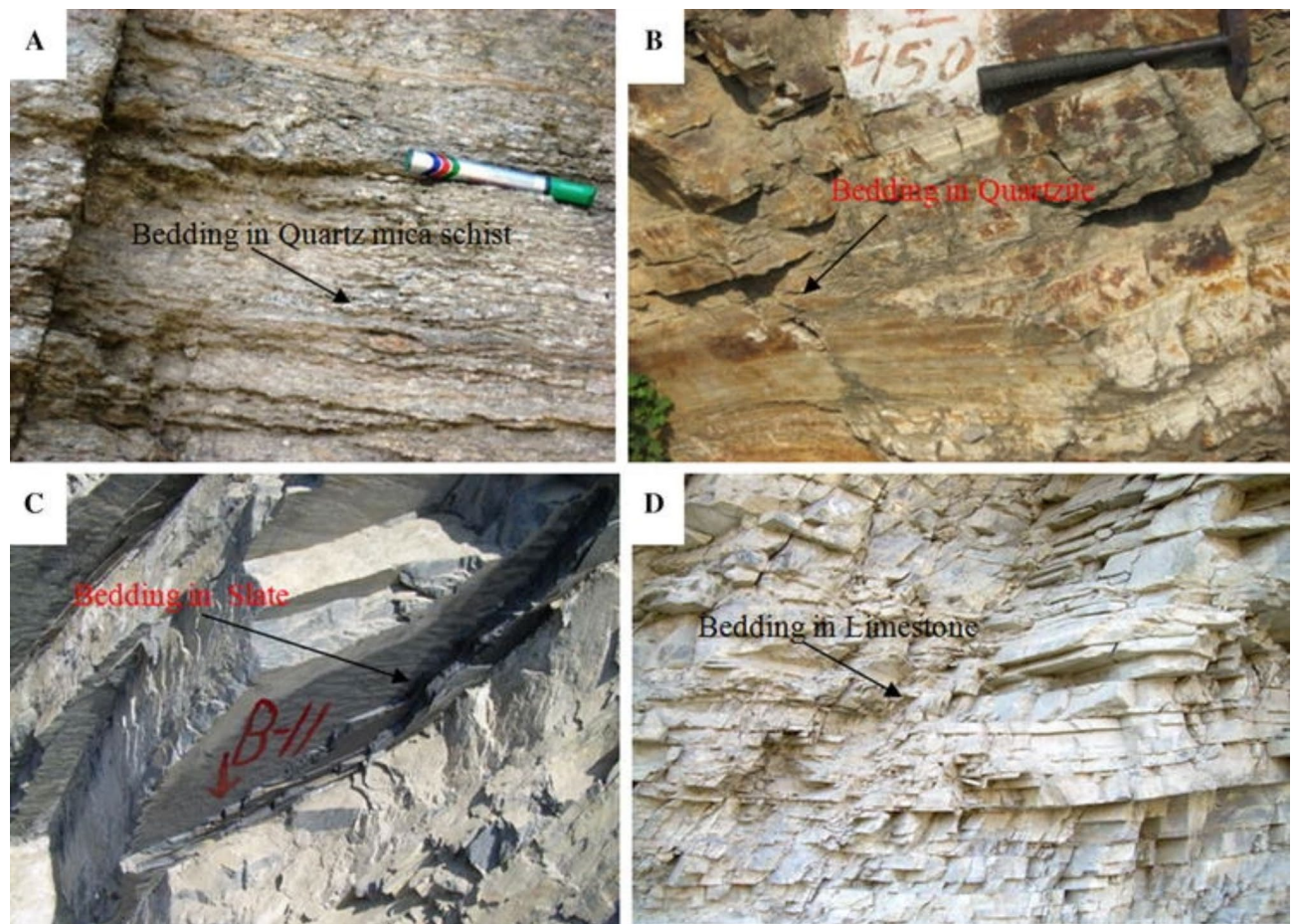
3

Fig. 2. Four rock materials taken from the Luhri area, Himanchal Pradesh, India. (**a**) quartz mica schist (**b**) quartzite (**c**) slate (**d**) limestone[54].
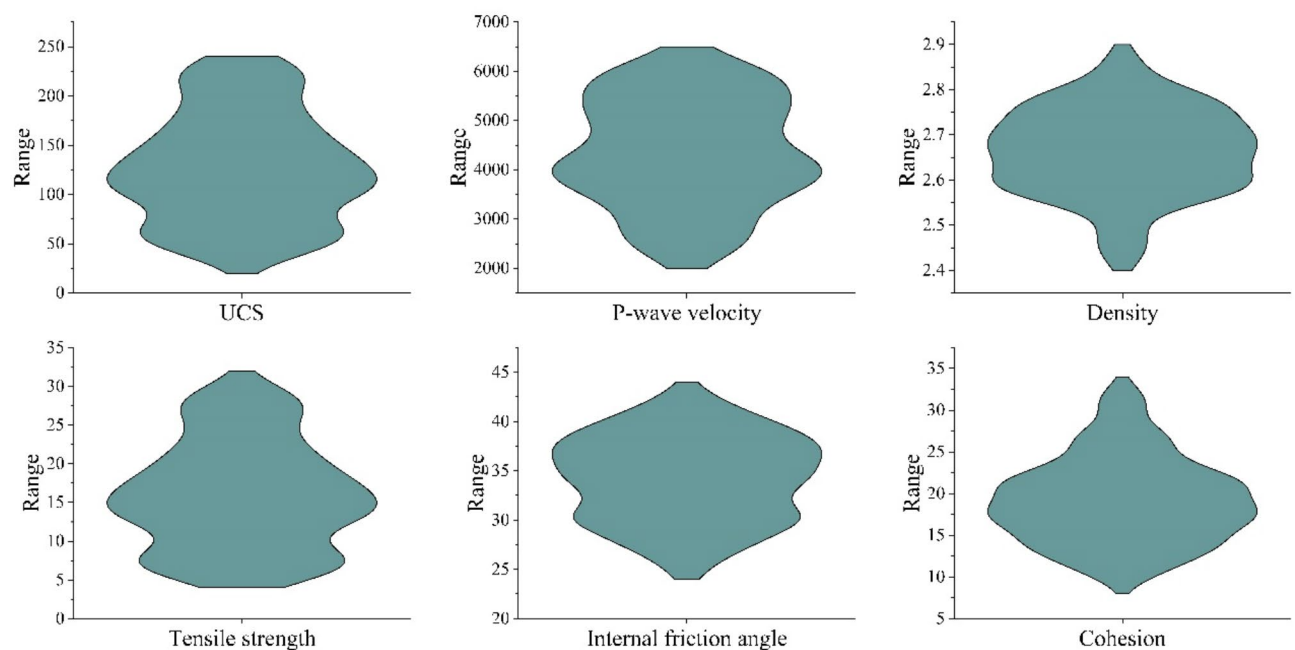

Fig. 3. Violin plot of each parameter.

	Skewness	Kurtosis	Coefficient of Variation	Minimum	Median	Maximum
P-wave velocity	−0.0683	−0.9759	0.2541	2209.34	4240.93	6328.14
Density	−0.2039	−0.2480	0.0368	2.41	2.67	2.89
UCS	0.2974	−0.8894	0.4336	40.97	120.9	237.76
TS	0.2911	−0.8727	0.4340	5.2	15.13	29.85
c	0.4770	−0.2314	0.2679	9.96	18.7	32.11
φ	−0.1208	−0.8783	0.129	24.57	34.55	43.35

Table 2. The detailed statistics of the input and output parameters.

data splitting ratios include 70:30 and 80:20. This represents 70 (80) percent of the dataset used for training and 30 (20) percent for testing. The predictive performance of the model can be effectively evaluated through data segmentation, which ensures that ML models are trained on representative samples while still being rigorously tested on new data that has never been seen before[64]. Therefore, in this paper, the data splitting ratio of 70:30 is used to randomly extract 139 data points and 60 data points from the dataset to create training and testing ML models respectively.

The data splitting method has potential shortcomings and biases in evaluating model prediction performance. In order to overcome these challenges, a 5-fold cross-validation method is implemented in this paper. As a popular statistical method, it provides a comprehensive and robust way to evaluate the predictive performance of ML models[65]. It not only reduces computation time, but also avoids any bias due to random data splitting (i.e. avoids underfitting and overfitting)[66]. In each fold, specify a different part for training and the rest for testing. This process is repeated in all folds until each part is utilized. Such a cross-validation process ensures that the ML model can fully learn on the training set. In the end, the best performing 'optimal parameters' are selected and passed on to the ML model to help it avoid overfitting.

Performance evaluation is an important part of the ML model[67]. Statistical evaluation indices are indispensable for quantifying the accuracy and reliability of model predictions. To evaluate the prediction performance, four commonly used statistical indices, defined in Eqs. 2–4 were used. Definitions and detailed statistical significance of these indicators can be found in the literatures[68,69]. In general, the prediction performance of the model is the best when these statistical evaluation indices reach the corresponding ideal value ($R^2 = 1$, RMSE = 0, MAE = 0)[70].

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(S_o - S_p\right)^2}{\sum_{i=1}^{N} \left(S_o - \overline{S_o}\right)^2} \tag{2}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |S_o - S_p| \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left(S_o - S_p\right)^2}{N}} \tag{4}$$

where N denotes the number of data. S_o and S_p are the actual and predicted results, respectively. $\overline{S_o}$ is the average of S_o.

Methodology
Extreme gradient boosting
XGBoost is an advanced variant of gradient boosted decision tree (GBDT) proposed by Chen and Guestrin[71]. The algorithm reduces the error of the prediction of the previous step by continuously generating new regression trees, gradually reduces the error between the predicted value and the true value, and then improves the prediction effect of the model[72]. By providing parallel tree boosting, the model can solve nonlinear problems rapidly and accurately in an effective way, and has been widely used in several fields. In recent years, the concept of XGBoost has been introduced to nonlinear problems that require high precision. Details about the XGBoost can be easily found in the following papers[73–75].

The schematic of XGBoost is presented in Fig. 4. XGBoost uses a regression tree (CART) as the base learner. CART is a binary tree where each leaf node represents a numerical prediction, and each internal node signifies a conditional judgment based on eigenvalues. During the training process, XGBoost iteratively corrects model errors by adding new regression trees. The introduction of regularization and second-order gradient optimization enhances both the training efficiency and prediction accuracy of the model. Through continuous iterations, multiple low-precision trees are combined to form a high-precision predictive model.

The principle of the XGBoost algorithm is briefly described below and its prediction function is shown in Eq. 5.
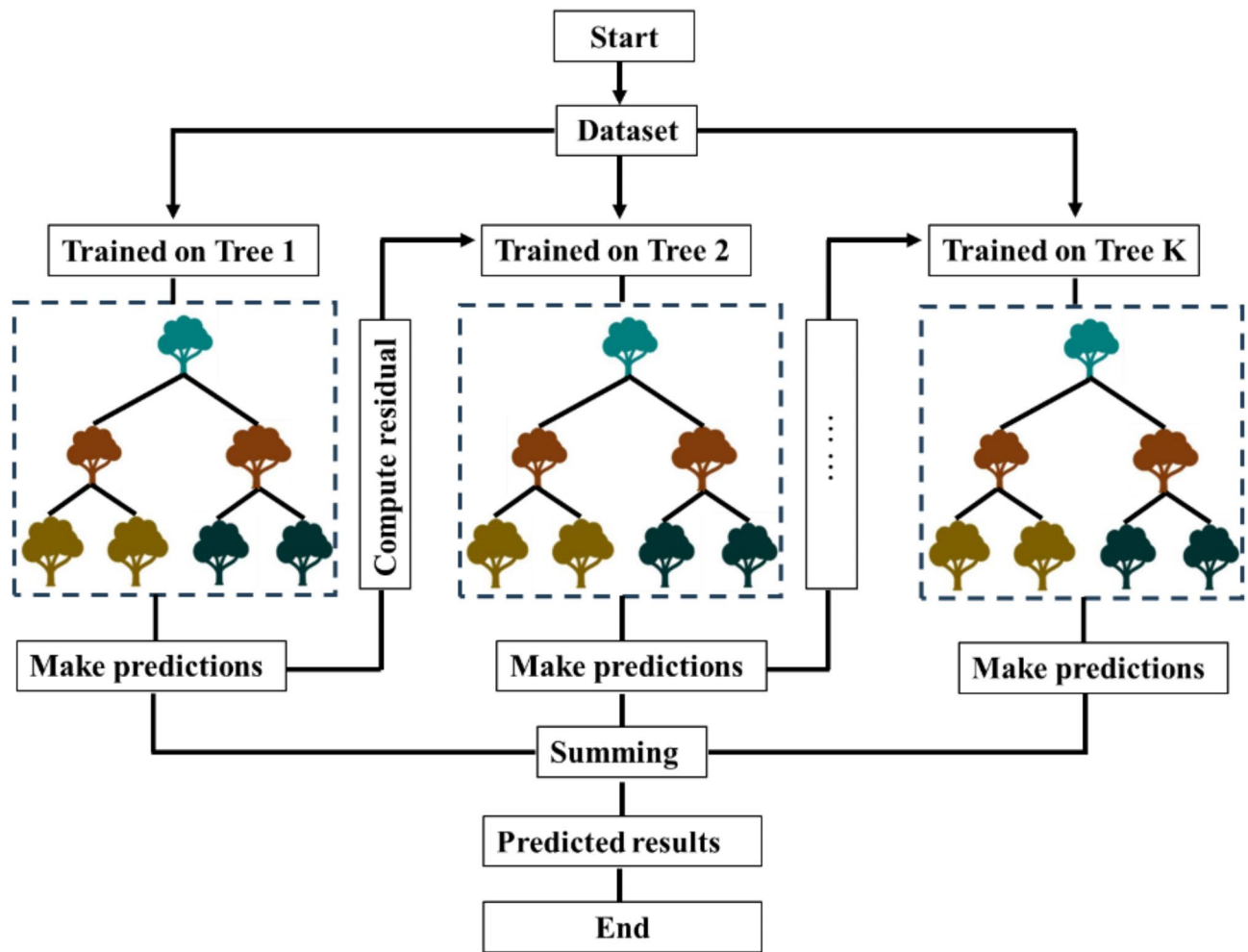
$$\hat{y_i} = \sum_{k=1}^{K} f_k (x_i), f_k \in \Gamma \tag{5}$$

Fig. 4. Graphical representation of XGBoost model.

where $\overset{\wedge}{y_i}$ is the predicted result value of the ith sample in the dataset, K is the total number of regression trees, x_i is the ith sample, and Γ is the space of regression trees.

As shown in Eq. 6, the XGBoost adds a regular term $\Omega(f_k)$ (i.e., penalty function) to the objective function *Obj* to reduce overfitting and increase variety. Term $\sum_{i=1}^{n} 1\left(y_i, \overset{\wedge}{y_i}\right)$ in Eq. 6 characterizes the fit of the model, i.e. how well the predictions match the actual results, while term $\Omega(f_k)$ measures the complexity of the model.

$$Obj = \sum_{i=1}^{n} 1\left(y_i, \overset{\wedge}{y_i}\right) + \sum_{k=1}^{K} \Omega(f_k) \qquad (6)$$

where y_i represents the true result value of the ith sample.

The penalty function can be rewritten as:

$$\Omega(f_k) = \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \gamma T \qquad (7)$$

where γ is the complexity cost of introducing additional leaf nodes, w_j^2 is the weight of the jth leaf node, λ is the regular term, and T is the number of leaf nodes.

The model was trained using the additive training method, as shown in Eq. 8. Additive training refers to the process of model training in which new tree models are gradually added to improve the overall model's prediction ability by adjusting the prediction results of existing models. This process can also be understood as the gradient boosting method and is one of the core training ideas used by XGBoost.

$$Obj^{(t)} = \sum_{i=1}^{n} 1\left(y_i, \overset{\wedge(t-1)}{y_i} + f_k\left(x_i\right)\right) + \Omega\left(f_k\right) \tag{8}$$

where $\overset{\wedge(t-1)}{y_i}$ is the prediction for the (t-1)th sample at the ith iteration and f_t is used to reduce the loss function.

Equation 8 is optimized with a 2nd-order Taylor expansion to obtain the final objective function as shown in Eq. 9. The parameters are continuously updated through Eq. 9 until the conditions are satisfied.

$$Obj^{(t)} = \sum_{k=1}^{K} \left[g_i f_k\left(x_i\right) + \frac{1}{2}h_i(f_k\left(x_i\right))^2\right] + \Omega\left(f_k\right) \tag{9}$$

where g_i and h_i denote the first and second derivatives obtained from the loss function, respectively.

Improved sparrow search algorithm

The sparrow search algorithm (SSA), a meta-heuristic machine learning algorithm, represents an innovative development in population intelligence optimization techniques[76]. SSA achieves parameter optimization by emulating natural sparrow behaviors, such as foraging and anti-predation strategies. Compared to other population intelligence algorithms, SSA is distinguished by its robust optimization capabilities and exceptional stability, making it widely applicable across diverse fields, including engineering, mathematics, and computer science[77–79].

Depending on the classification, sparrow populations contain both producers and scroungers. As shown in Fig. 5, producers are responsible for locating food and guiding the population to food sources, while scroungers depend on the producers to access these resources. The process is outlined as follows:

As shown in Eq. 10, a sparrow population can be mathematically represented as a two-dimensional matrix of size N x D. Each element in Eq. 10 represents the decision variable at the jth position of the ith sparrow, and each row vector represents the set of sparrows at the ith position.

$$Population = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \cdots & \cdots & \cdots & \cdots \\ x_{M,1} & x_{M,2} & \cdots & x_{M,D} \end{bmatrix} \tag{10}$$

where M is the population size and D is the dimension of the search space.

Each $x_{i,j}$ is given a random value within the specified upper and lower bound. The fitness function is employed to compute the fitness value for each sparrow location in the population and determine the location of the best sparrow (e.g., x^{Gbest}) in the population. The sparrow with the better fitness value is prioritized for food and acts as a producer to lead the entire population to run to the food source.

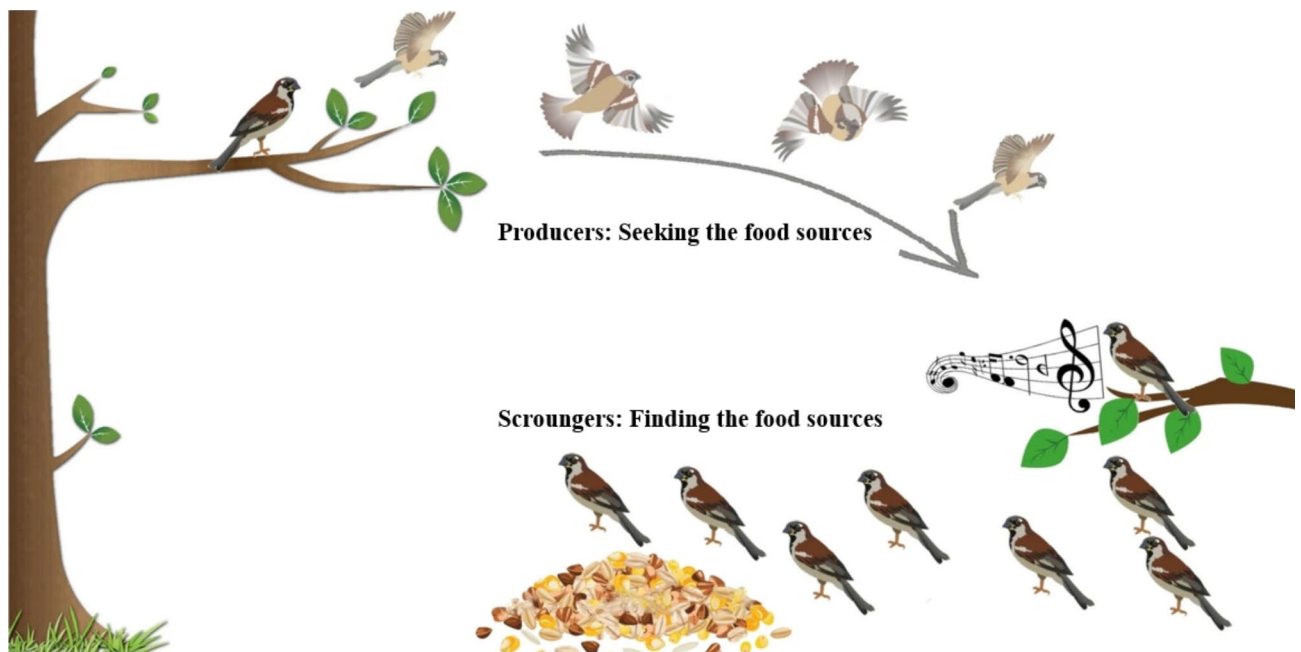As shown in Eq. 11, the location of producers is updated to[80]:

Fig. 5. Graphical representation of sparrow search algorithm (SSA), modified from[77].

$$x_{i,j}^{g+1} = \begin{cases} x_{i,j}^g + Q \times L & r \geqslant ST \\ x_{i,j}^g \times e^{\frac{-i}{\alpha \times G \text{max}}} & r < ST \end{cases} \tag{11}$$

where g represents the current number of iterations, G_{max} is the maximum number of iterations considered in the search process; Q is a random number obeying a normal distribution, α is a uniform random number in (0, 1); the alarm threshold r belongs to [0, 1], the safety threshold ST belongs to [0.5, 1], and L is a matrix that is 1 × d and has elements assigned to value 1.

As shown in Eq. 12, the location of scroungers is updated to:

$$x_{i,j}^{g+1} = \begin{cases} x_j^{Pbest} - \left| x_{i,j}^g - x_j^{Pbest} \right| \times A^+ \times L & i \leqslant \frac{M}{2} \\ Qe^{\frac{x^{\text{worst}} - x_{i,j}^g}{i^2}} & i > \frac{M}{2} \end{cases} \tag{12}$$

where x_j^{Pbest} and x_j^{Worst} are the best and worst positions of the discoverer, respectively, and A^+ is a 1 × d matrix whose elements are randomly assigned values − 1 and 1.

When the sparrow is foraging, a randomly selected portion of the population (usually taken as 10–20%) of sparrows will be responsible for vigilance. In a dangerous situation, either the producer or the scrounger will abandon the current food and fly to a new safe location. The positions of these scouters are updated as shown in Eq. 13.

$$x_{i,j}^{g+1} = \begin{cases} x_{i,j}^g + K \left(\frac{\left| x_{i,j}^g - x_j^{\text{worst}} \right|}{(f(x_i^g) - f(x^{\text{worst}})) + \varepsilon} \right) & f(x_i^g) = f(x^{Gbest}) \\ x_j^{Gbest} + \beta \times \left| x_{i,j}^g - x_j^{Gbest} \right| & f(x_i^g) > f(x^{Gbest}) \end{cases} \tag{13}$$

where $f(x_i^g)$ is the fitness value of the ith sparrow at the gth iteration, and $f(x^{\text{worst}})$ is the fitness value of the worst sparrow in the population; β is a step control parameter; K is a random number ranging from − 1 to 1, and ε is a small random value to avoid the denominator to be 0. x^{Gbest} is the best-positioned sparrow in the population.

Although the SSA algorithm has significant advantages in terms of search accuracy, stability and convergence speed, it still has some serious flaws. For example, the convergence strategy of SSA is to jump directly to the neighborhood of the current optimal solution, which will be underpowered at the late stage of the search for optimality and underpowered for local search. To solve this problem, the Lévy flight strategy is introduced into the location of scroungers to improve the global search capability. As shown in Eq. 14, the introduction of the Lévy flight strategy modifies Eq. 12. This optimized algorithm is known as the Improved sparrow search algorithm (ISSA).

$$x_{i,j}^{g+1} = \begin{cases} x_j^{Pbest} + H \left| x_{i,j}^g - x_j^{Pbest} \right| \times A^+ \times L & i \leqslant \frac{M}{2} \\ Qe^{\frac{x_j^{\text{worst}} - x_{i,j}^g}{i^2}} & i > \frac{M}{2} \end{cases} \tag{14}$$

where H is a random number determined by Eqs. 15–17.

$$H = \mu / |v|^{1/\delta}, \delta = \frac{3}{2} \tag{15}$$

$$\mu \sim N\left(0, \sigma_\mu^2\right), v \sim \left(0, \sigma_v^2\right) \tag{16}$$

$$\sigma_\mu = \left\{ \frac{\Gamma(1+\delta) \sin\left(\frac{\pi\delta}{2}\right)}{\delta \Gamma\left[\frac{1+\delta}{2}\right] 2^{\frac{\delta-1}{2}}} \right\}^{1/\delta}, \sigma_v = 1 \tag{17}$$

Model development

Generally, ML models that do not incorporate optimization algorithms often have convergence problems. Moreover, the artificial determination of model hyperparameters is subjective and unfavorable for application. XGBoost model improves the computing speed and accuracy to the extreme on the basis of efficient implementation of the gradient boosting decision tree algorithm, but the step-by-step growth strategy leads to unnecessary consumption of computer operating resources. The construction process of the ML model essentially lies in the determination of hyperparameters[81]. In this section, the ISSA algorithm is used to optimize the hyperparameters of the XGBoost model. Accordingly, a new hybrid ML model (i.e., ISSA-XGBoost model) is established. To the best of the authors' knowledge, the application of the ISSA algorithm in improving hybrid ML models for predicting cohesion and internal friction angle has not been reported yet.

The modeling steps of the hybrid ISSA-XGBoost model are outlined in Fig. 6. After collecting the raw dataset, the raw data are processed using a suitable normalization method (e.g., Eq. 1). Randomly divide the dataset into the test set and training set. Then, the ISSA algorithm is initialized and the search space as well as the model hyperparameters are set. Set the range of parameters to be optimized in the XGBoost model to generate the initial population of sparrows. Next, based on the resulting sparrow population, iterate with statistical evaluation indices such as R^2 or RMSE as the fitness function to calculate the positional fitness of each sparrow. The obtained
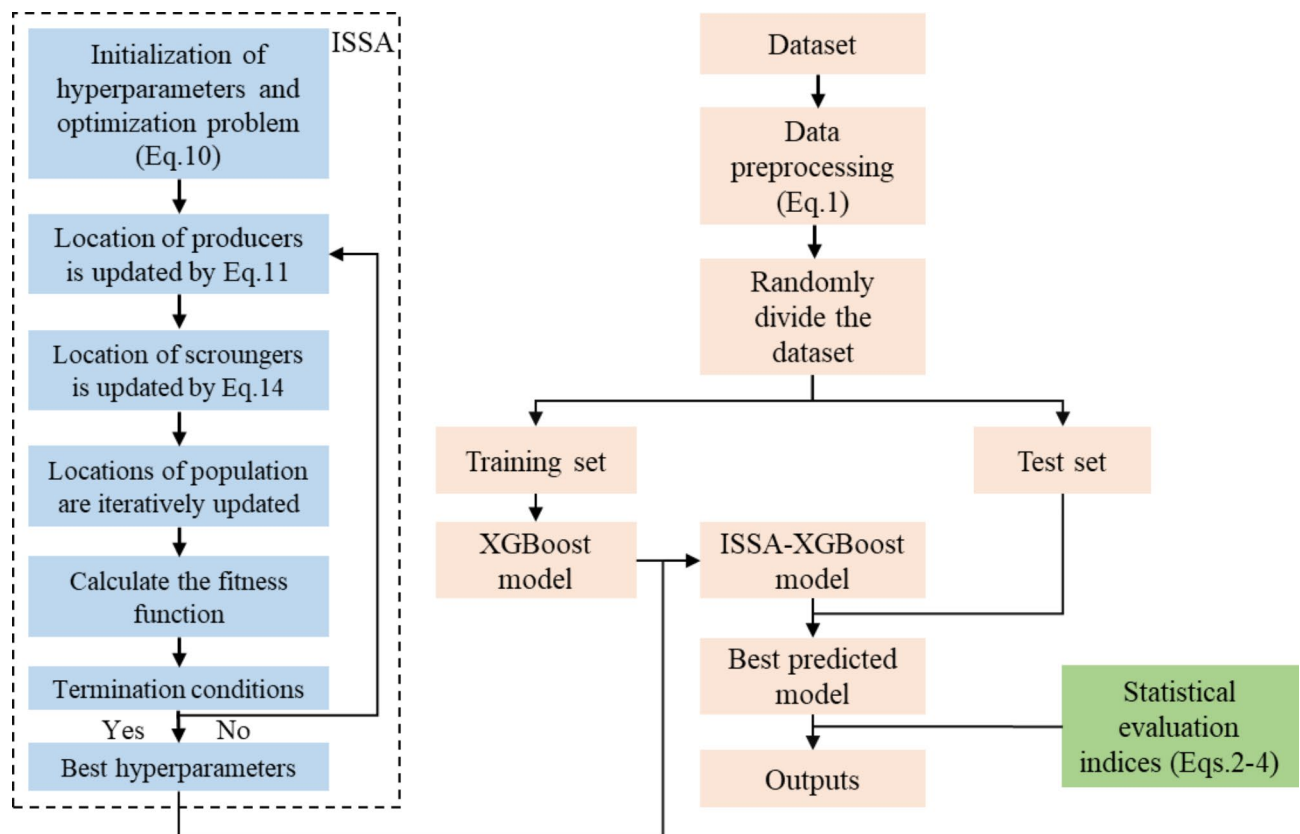
Fig. 6. Framework example of the proposed ISSA-XGBoost model.

fitness values are sorted and the current global best position is localized. The sparrow position is continuously updated according to Eqs. 10–17. Finally, when the number of iterations satisfies the termination condition, terminate the iteration and output the parameters corresponding to the best sparrow position.

Results and analysis
Model performance
For comparison, the predictions of the XGBoost model are also presented. In this case, the XGBoost model was only validated with 5-fold cross-validation method and did not use any optimization algorithm for hyperparameter optimization. The prediction performance of the proposed hybrid ML model and the XGBoost model for both the training and test sets is illustrated in Figs. 7 and 8.

Each point in Figs. 7 and 8 represents a predicted sample, with the X-axis denoting the data index and the Y-axis representing either the cohesion or internal friction angle. From these figures, it can be seen that the prediction results of the ISSA-XGBoost model are closer to the experimental results both on the training set and the test set. Compared with the XGBoost model, the ISSA-XGBoost model is closer to the experimental results. This indicates that the ISSA-XGBoost model is very accurate in modeling the internal friction angle and cohesion. Table 3 shows the statistical evaluation indices of the two ML models on the training set and test set, respectively. It can be seen that for both cohesion and internal friction angle, the R^2 values on the training set are very close to the R^2 values on the test set. This result indicates that the ISSA-XGBoost model is well-trained. It is reasonable that the R^2 values on the test set are slightly lower. Regarding cohesion and internal friction angle, the ISSA-XGBoost model significantly outperforms the XGBoost model (on both the training and test sets). The excellent prediction accuracy, as demonstrated in Table 3, underscores the potential of the proposed hybrid ML model as a reliable tool for predicting cohesion and the internal friction angle.

Comparison with previous studies
In this section, the proposed ML model is evaluated against four other ML models: the lasso regression (LR), the ridge regression (RR), the support vector machine (SVM), and the decision tree (DT). After training on a dataset split into 70% training and 30% test sets, the models were evaluated on the test set. The three statistical evaluation indices of each model are shown in Table 4. For cohesion, the differences in prediction performance among the models are relatively small. The best performer is the proposed model of this paper, whose three statistical evaluation indices are RMSE = 0.322, MAE = 0.450, R^2 = 0.982. For the internal friction angle, the proposed model significantly outperforms the other four models. Compared with the worst-performing LR model (R^2 = 0.606, RMSE = 2.7255, MAE = 2.3064), the three statistical indicators of the proposed model are
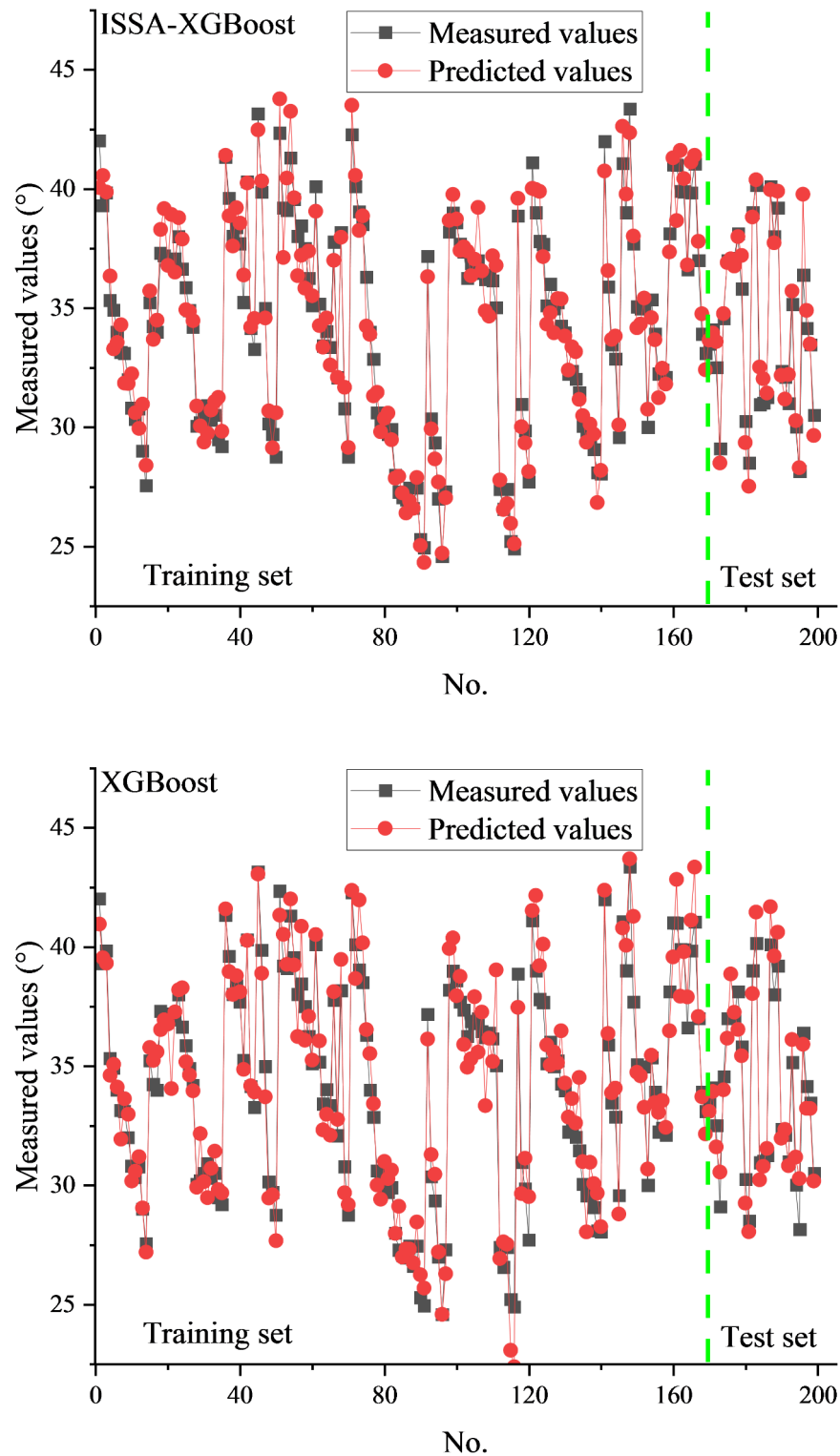
Fig. 7. Comparisons between the predicted and measured internal friction angles.

as high as $R^2 = 0.932$, RMSE = 0.920, MAE = 0.625. In conclusion, the proposed model demonstrates superior performance, not only compared to the four alternative ML models but also relative to the XGBoost model.

Discussion
Data preprocessing methods
Data preprocessing methods play a critical role in maintaining data consistency and integrity, enabling ML models to achieve optimal predictive performance. The effectiveness of ML models depends heavily on the
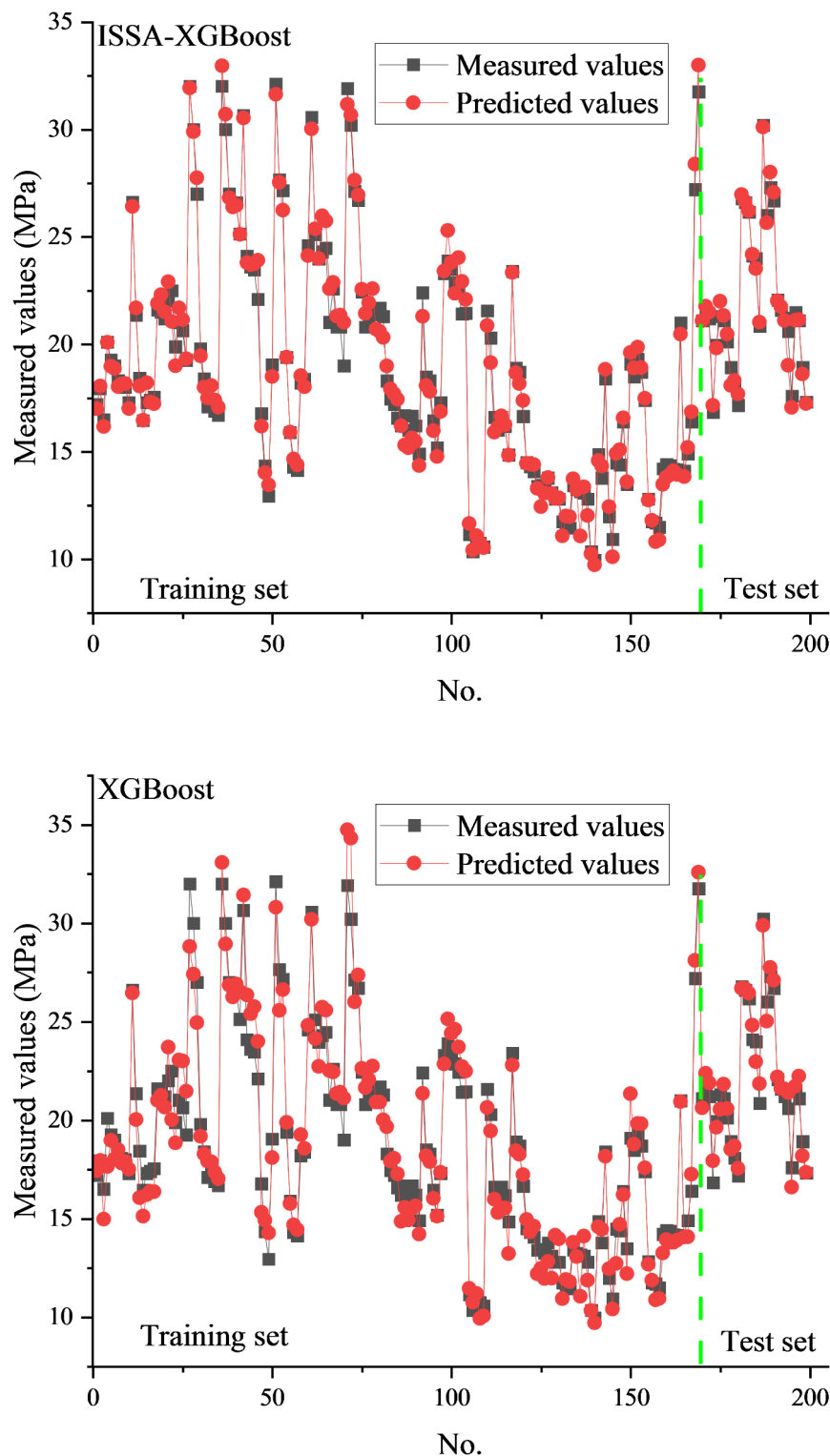
Fig. 8. Comparisons between the predicted and measured cohesions.

chosen algorithm and dataset. Once these factors are established, data preprocessing becomes a decisive element in influencing model performance. Proper data preprocessing not only eliminates magnitude discrepancies in raw data—avoiding issues such as the "big numbers eat decimals" phenomenon—but also significantly enhances the computational efficiency of ML models[82]. In this section, we will discuss the effect of three common data preprocessing methods on the ML model performance.

The minimum-maximum normalization method, zero-mean normalization method, and arctangent normalization method were selected to evaluate the effects of different data preprocessing techniques on the

Parameters	Model	Dataset	R^2	RMSE	MAE
Internal friction angle	ISSA-XGBoost	Training set	0.957	0.943	0.770
		Test set	0.932	0.920	0.625
	XGBoost	Training set	0.931	1.192	0.922
		Test set	0.919	1.002	0.831
Cohesion	ISSA-XGBoost	Training set	0.985	0.473	0.639
		Test set	0.982	0.322	0.450
	XGBoost	Training set	0.953	0.895	1.135
		Test set	0.964	0.549	0.642

Table 3. Prediction performances for each model.

Model		RMSE	MAE	R^2
LR	Internal friction angle	2.7255	2.3064	0.606
	Cohesion	1.4896	1.1454	0.928
RR	Internal friction angle	2.7256	2.3003	0.607
	Cohesion	1.2412	1.0335	0.934
SVM	Internal friction angle	1.2906	0.9094	0.916
	Cohesion	0.8253	0.5577	0.977
DT	Internal friction angle	2.2963	1.7655	0.822
	Cohesion	1.0560	0.8389	0.961
ISSA-XGBoost model	Internal friction angle	0.920	0.625	0.932
	Cohesion	0.322	0.450	0.982

Table 4. Prediction performances for each model.

predictive performance of the ISSA-XGBoost model. The formula for the minimum-maximum normalization method is presented in Eq. 1, while the zero-mean normalization and arctangent normalization methods are detailed in Eqs. 18 and 19. Equation 18 transforms the raw dataset into a standard normal distribution with unit standard deviation and zero mean. And Eq. 19 transforms the raw dataset to range [−1,1].

$$\bar{X} = \frac{X - \mu}{\sigma} \tag{18}$$

$$\bar{X} = 2\frac{\arctan X}{\pi} \tag{19}$$

where μ and σ are the mean and standard deviation of the dataset, respectively.

The original dataset was processed using Eqs. 18 and 19 before undergoing the same modeling process. The statistical evaluation indices for each model are presented in Fig. 9. The results indicate that all three data preprocessing methods achieve predictions closely aligned with the actual values, demonstrating the ISSA-XGBoost model's capability to accurately simulate the shear strength parameters of rocks. Among the preprocessing methods, Eq. 1 delivers significantly higher accuracy compared to Eqs. 18 and 19. Notably, even the method with the lowest prediction accuracy (Eq. 18 in Fig. 9) outperforms the four ML models listed in Table 4 in predicting both internal friction angle and cohesion.

Feature importance score

Feature importance analysis serves as a critical reference for assessing the contribution of input parameters to the model's predictions[83,84]. Figure 10 shows the feature importance score of each input variable in the developed model to cohesion and internal friction angle. A higher feature importance score indicates a relatively greater influence of the corresponding input variable on the output variable[85]. The results reveal that the input variables exert different impacts on the two properties. For the cohesion, the effects of the four input variables are 0.96 (P-wave velocity), −0.036 (density), 0.97 (UCS), 0.96 (TS). For the internal friction angle, the scores 0.33 (P-wave velocity), 0.69 (density), 0.3 (UCS), 0.27 (TS). Among the input parameters, UCS demonstrates the greatest relative importance for cohesion, whereas density has the highest relative importance for the internal friction angle.

Limitations

Internal friction angle and cohesion hold significant potential for economic benefits such as cost optimization and reduced time investment when determined using ML techniques. This paper introduces a novel ML model that integrates the strengths of XGBoost and ISSA, providing a reliable hybrid approach for predicting internal friction angle and cohesion. Although this paper has yielded valuable insights, its limitations should not be
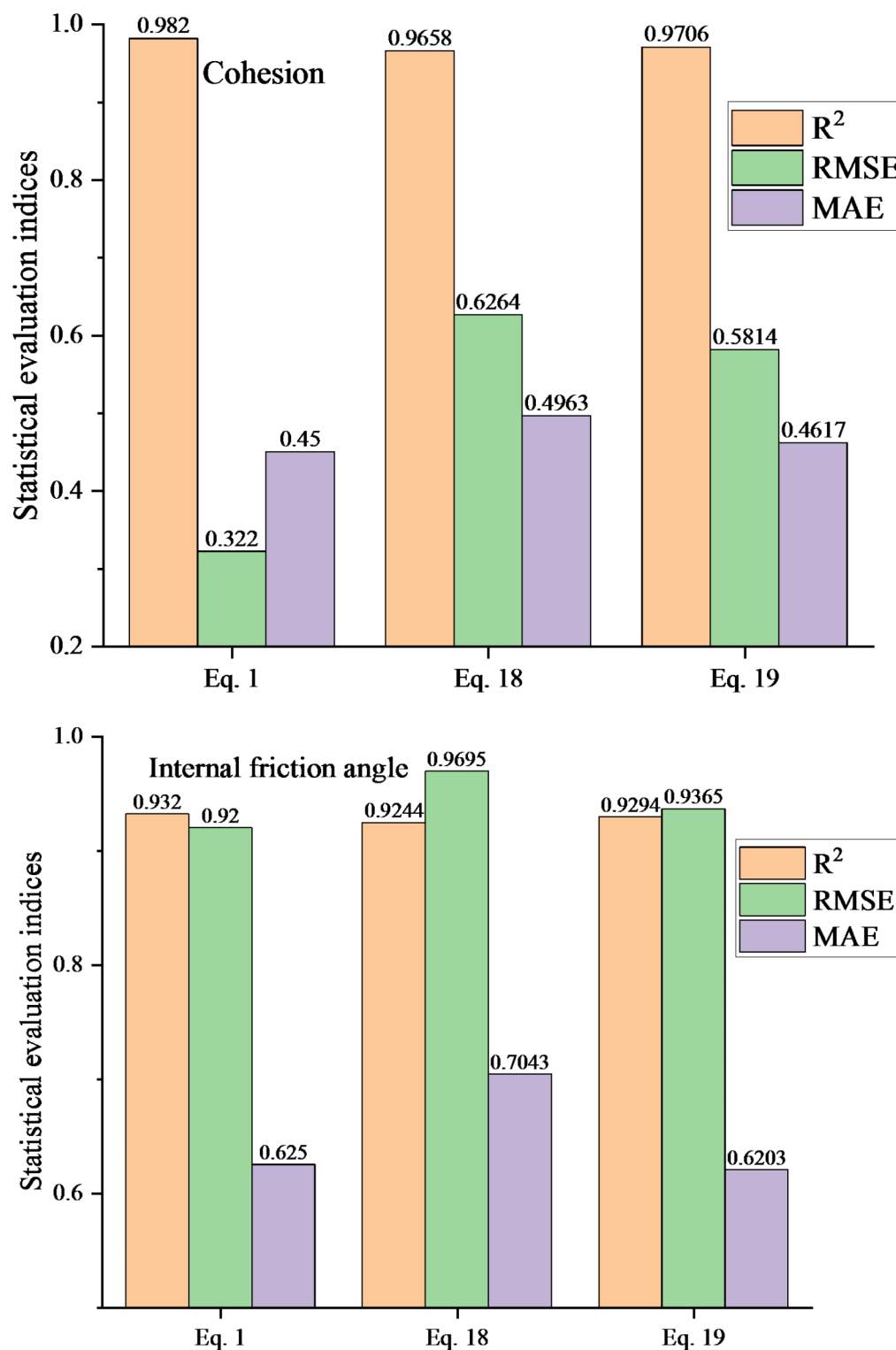
Fig. 9. Predictive indicator results for different data preprocessing method.

overlooked. For example, errors inherent to ML models and the variability introduced by experimental results are unavoidable[86,87]. In practical engineering, moisture content has a great influence on internal friction angle and cohesion[88,89]. However, due to the absence of moisture content data in the compiled dataset, it was not included as an input variable in the proposed ML model. Despite this omission, the model still achieves satisfactory predictive performance. A possible explanation is that, within the selected dataset described in Sect. 2, the moisture content of the samples is consistent. Consequently, the ML model delivers reliable predictions even without moisture content as an input.
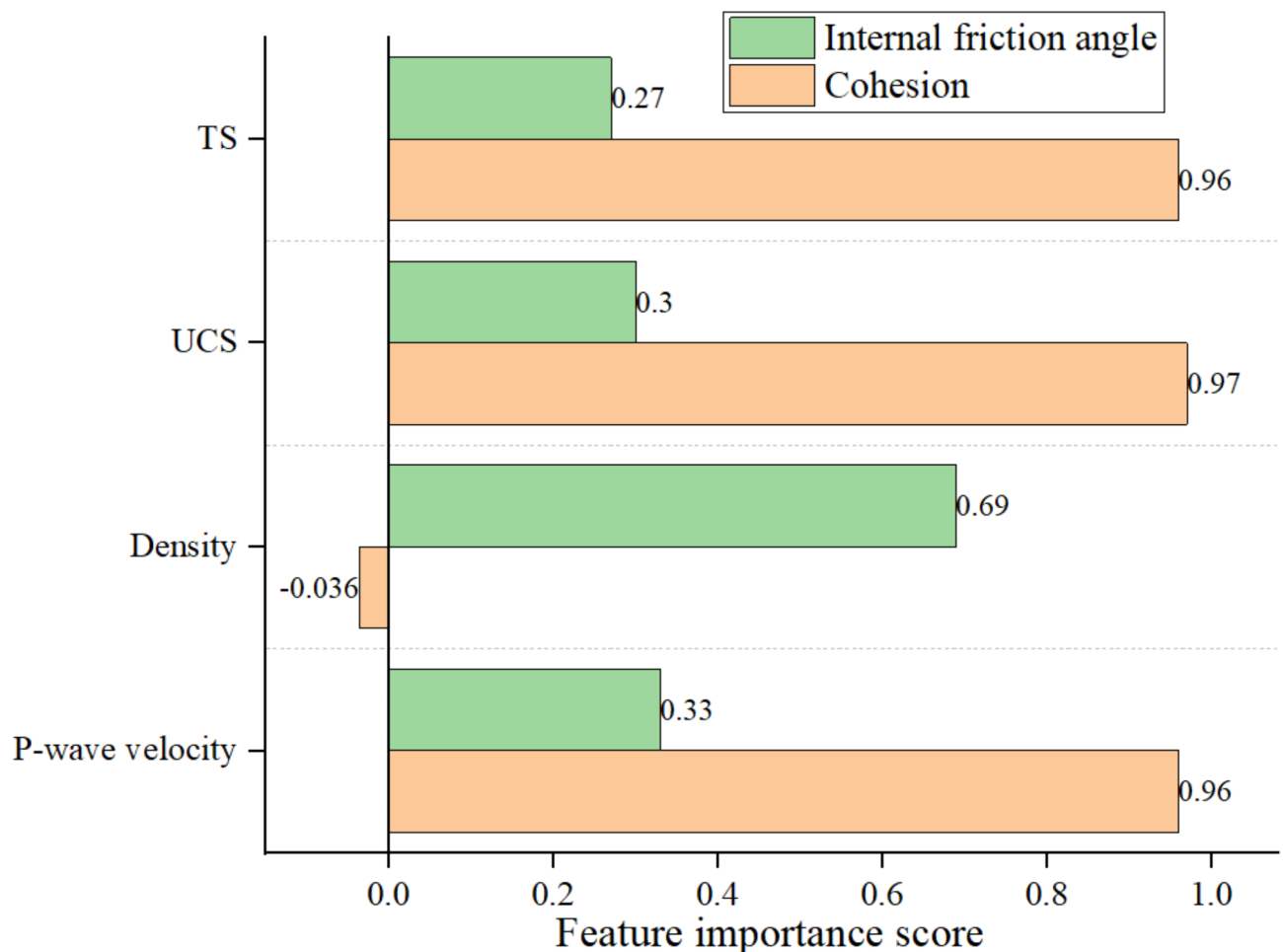
Fig. 10. Feature importance score of inputs.

Furthermore, as more data becomes available, the generalization ability and prediction accuracy of the constructed ML model can be further enhanced[90]. In future research, expanding the dataset to include data from various types of rock materials will be a valuable step. Additionally, the input variables, such as uniaxial compressive strength and tensile strength, currently have a limited value range. Expanding these ranges could improve the model's generalization capability. Lastly, to bridge the gap between computational predictions and practical applications, we aim to develop a user-friendly graphical user interface.

Conclusions

1) Utilizing ML technology, cohesion and internal friction angle of rock materials can be accurately estimated using extensive historical data. The proposed model, built on four input variables, demonstrates strong generalization ability and high prediction accuracy. Compared with the actual observed values, the new model gives reliable prediction results with R^2, RMSE and MAE values of $0.932(\varphi)$, $0.982(c)$, $0.920(\varphi)$, $0.322(c)$, $0.625(\varphi)$, $0.450(c)$, respectively.
2) The proposed ISSA-XGBoost model was compared with five other ML models. For parameter internal friction angle, the performance ranking is ordered as ISSA-XGBoost > XGBoost > SVM > DT > RR > LR. For parameter cohesion, the ranking is slightly different: ISSA-XGBoost > SVM > XGBoost > DT > RR > LR.
3) The feature importance analysis shows that the four input parameters affect cohesion in the following order from strongest to weakest: UCS > P-wave velocity = TS > density. The four input parameters affect the internal friction angle in the following order from strongest to weakest: density > P-wave velocity > UCS > TS.

Data availability

The data used to support the findings of this study are available from the corresponding author upon request.

References

1. Castro-Filgueira, U., Alejano, L. R. & Ivars, D. M. Particle flow code simulation of intact and fissured granitic rock samples. *J. Rock. Mech. Geotech.* **12** (5), 960–974. https://doi.org/10.1016/j.jrmge.2020.01.005 (2020).
2. Gomes Silva, G. & Cerqueira Silva, R. Mechanical evaluation and optimal mixing ratios in co-disposal of phosphate mining tailings and waste rock. *Sci. Rep.* **14** (1), 31584. https://doi.org/10.1038/s41598-024-72537-2 (2024).
3. Xie, S. J. et al. A statistical damage constitutive model considering whole joint shear deformation. *Int. J. Damage Mech.* **29** (6), 988–1008. https://doi.org/10.1177/1056789519900778 (2020).
4. Maľa, M., Greif, V. & Ondrásik, M. Pore structure evolution in andesite rocks induced by freeze-thaw cycles examined by non-destructive methods. *Sci. Rep.* https://doi.org/10.1038/s41598-022-12437-5 (2022).
5. Siddig, O., Ibrahim, A. F. & Elkatatny, S. Estimation of rocks' failure parameters from drilling data by using artificial neural network. *Sci. Rep.* https://doi.org/10.1038/s41598-023-30092-2 (2023).
6. Li, D. Q., Chen, Y. F., Lu, W. B. & Zhou, C. B. Stochastic response surface method for reliability analysis of rock slopes involving correlated non-normal variables. *Comput. Geotech.* **38** (1), 58–68. https://doi.org/10.1016/j.compgeo.2010.10.006 (2011).
7. Sopaci, E. & Akgün, H. Geotechnical assessment and engineering classification of the Antalya Tufa rock, Southern Turkey. *Eng. Geol.* **197**, 211–224. https://doi.org/10.1016/j.enggeo.2015.08.029 (2015).
8. Xie, S. et al. Nonlinear shear constitutive model for peak shear-type joints based on improved Harris damage function. *Arch. Civ. Mech. Eng.* https://doi.org/10.1007/s43452-020-00097-z (2020).
9. Chen, Y. et al. Crack propagation and scale effect of random fractured rock under compression-shear loading. *J. Mater. Res. Technol.* **23**, 5164–5180. https://doi.org/10.1016/j.jmrt.2023.02.104 (2023).
10. Sheshde, E. A., Cheshomi, A. & Maleki, R. Estimation of shear strength parameters of intact limestones using miniature triaxial test. *Eng. Geol.* **334**, 107509. https://doi.org/10.1016/j.enggeo.2024.107509 (2024).
11. Kalantari, S., Baghbanan, A. & Hashemalhosseini, H. An analytical model for estimating rock strength parameters from small-scale drilling data. *J. Rock. Mech. Geotech.* **11** (1), 135–145. https://doi.org/10.1016/j.jrmge.2018.09.005 (2019).
12. Huang, Q. Z. et al. Deformation failure and damage evolution law of weathered granite under triaxial compression. *Nondestruct Test. Eval.* **39** (7), 1900–1924. https://doi.org/10.1080/10589759.2023.2283710 (2024).
13. Sopacı, E. & Akgün, H. Geotechnical assessment and engineering classification of the Antalya Tufa rock, Southern Turkey. *Eng. Geol.* **197**, 211–224. https://doi.org/10.1016/j.enggeo.2015.08.029 (2015).
14. Xie, S. J. et al. A damage constitutive model for shear behavior of joints based on determination of the yield point. *Int. J. Rock. Mech. Min. Sci.* https://doi.org/10.1016/j.ijrmms.2020.104269 (2020).
15. Taheri, A., Zhang, Y. & Munoz, H. Performance of rock crack stress thresholds determination criteria and investigating strength and confining pressure effects. *Constr. Build. Mater.* **243**, 118263. https://doi.org/10.1016/j.conbuildmat.2020.118263 (2020).
16. Aladejare, A. E. & Wang, Y. Influence of rock property correlation on reliability analysis of rock slope stability: from property characterization to reliability analysis. *Geosci. Front.* **9** (6), 1639–1648. https://doi.org/10.1016/j.gsf.2017.10.003 (2018).
17. Barton, N. Shear strength criteria for rock, rock joints, rockfill and rock masses: problems and some solutions. *J. Rock. Mech. Geotech.* **5** (4), 249–261. https://doi.org/10.1016/j.jrmge.2013.05.008 (2013).
18. Xie, S. et al. Shear strength model of joints based on Gaussian smoothing method and macro-micro roughness. *Comput. Geotech.* **143**, 104605. https://doi.org/10.1016/j.compgeo.2021.104605 (2022).
19. Yadav, S., Saldana, C. & Murthy, T. G. Porosity and geometry control ductile to brittle deformation in indentation of porous solids. *Int. J. Solids Struct.* https://doi.org/10.1016/j.ijsolstr.2016.03.031 (2016).
20. Lin, H. et al. Comparative analysis of rock damage models based on different distribution functions. *Geotech. Geol. Eng.* **40**, 1–28. https://doi.org/10.1007/s10706-021-01899-5 (2022).
21. Xie, S., Lin, H. & Duan, H. A novel criterion for yield shear displacement of rock discontinuities based on renormalization group theory. *Eng. Geol.* **314**, 107008. https://doi.org/10.1016/j.enggeo.2023.107008 (2023).
22. Gu, L., Feng, X-T., Kong, R., Yang, C. & Xia, Y. Evolution of mechanical parameters of Shuangjiangkou granite under different loading cycles and stress paths. *J. Rock. Mech. Geotech.* **16** (4), 1113–1126. https://doi.org/10.1016/j.jrmge.2023.09.005 (2024).
23. Hashiba, K. & Fukui, K. New multi-stage triaxial compression test to investigate the loading-rate dependence of rock strength. *Geotech. Test. J.* https://doi.org/10.1520/gtj20140061 (2014).
24. Shen, J. & Jimenez, R. Predicting the shear strength parameters of sandstone using genetic programming. *Bull. Eng. Geol. Environ.* **77** (4), 1647–1662. https://doi.org/10.1007/s10064-017-1023-6 (2018).
25. Xie, S., Lin, H., Chen, Y. F. & Ma, T. X. Modified Mohr-Coulomb criterion for nonlinear strength characteristics of rocks. *Fatigue Fract. Eng. M.* **47** (6), 2228–2242. https://doi.org/10.1111/ffe.14278 (2024).
26. Karaman, K., Cihangir, F., Ercikdi, B., Kesimal, A. & Demirel, S. Utilization of the Brazilian test for estimating the uniaxial compressive strength and shear strength parameters. *J. S Afr. Inst. Min. Metall.* **115**, 185–192. https://doi.org/10.17159/2411-9717/2015/v115n3a3 (2015).
27. Çobanoğlu, İ. & Çelik, S. B. Estimation of uniaxial compressive strength from point load strength, Schmidt hardness and P-wave velocity. *Bull. Eng. Geol. Environ.* **67** (4), 491–498. https://doi.org/10.1007/s10064-008-0158-x (2008).
28. Moon, K. & Yang, S-B. Cohesion and internal friction angle estimated from Brazilian tensile strength and unconfined compressive strength of volcanic rocks in Jeju Island. *J. Korean Geotech. Soc.* **36** (2), 17–28 (2020).
29. Xie, S., Lin, H., Duan, H., Yao, R. & Ma, T. A novel triaxial strength criterion for rocks based on the ultimate strength and its application. *Geoenergy Sci. Eng.* **246**, 213590. https://doi.org/10.1016/j.geoen.2024.213590 (2025).
30. Sivakugan, N., Das, B. M., Lovisa, J. & Patra, C. R. Determination of C and Φ of rocks from indirect tensile strength and uniaxial compression tests. *Int. J. Geotech. Eng.* **8** (1), 59–65. https://doi.org/10.1179/1938636213Z.00000000053 (2014).
31. Jaeger, J. C., Cook, N. G. & Zimmerman, R. *Fundamentals of Rock Mechanics* (Wiley, 2009).
32. Xie, S., Lin, H., Duan, H. Y., Liu, H. W. & Liu, B. H. Numerical study on cracking behavior and fracture failure mechanism of fractured rocks under shear loading. *Comp. Part. Mech.* **11** (2), 903–920. https://doi.org/10.1007/s40571-023-00660-4 (2024).
33. Abbas, A. K., Flori, R. E. & Alsaba, M. Estimating rock mechanical properties of the Zubair shale formation using a Sonic wireline log and core analysis. *J. Nat. Gas Sci. Eng.* **53**, 359–369. https://doi.org/10.1016/j.jngse.2018.03.018 (2018).
34. Xie, S. et al. Prediction of shear strength of rock fractures using support vector regression and grid search optimization. *Mater. Today Commun.* **36**, 106780. https://doi.org/10.1016/j.mtcomm.2023.106780 (2023).
35. Pouyanfar, N. et al. Machine learning-assisted rheumatoid arthritis formulations: A review on smart pharmaceutical design. *Mater. Today Commun.* https://doi.org/10.1016/j.mtcomm.2024.110208 (2024).
36. Ghorbani, E. & Yagiz, S. Predicting disc cutter wear using two optimized machine learning techniques. *Arch. Civ. Mech. Eng.* https://doi.org/10.1007/s43452-024-00911-y (2024).
37. Xie, S. J. et al. Hybrid machine learning models to predict the shear strength of discontinuities with different joint wall compressive strength. *Nondestruct Test. Eva.* https://doi.org/10.1080/10589759.2024.2381083 (2024).
38. Takaew, P., Cecilia Xia, J. & Doucet, L. S. Machine learning and tectonic setting determination: bridging the gap between Earth scientists and data scientists. *Geosci. Front.* **15** (1), 101726. https://doi.org/10.1016/j.gsf.2023.101726 (2024).
39. Rayhani, M., Tatar, A., Shokrollahi, A. & Zeinijahromi, A. Exploring the power of machine learning in analyzing the gas minimum miscibility pressure in hydrocarbons. *Geoenergy Sci. Eng.* **226**, 211778. https://doi.org/10.1016/j.geoen.2023.211778 (2023).
40. Fathipour-Azar, H., Wang, J. F., Jalali, S. M. E. & Torabi, S. R. Numerical modeling of geomaterial fracture using a cohesive crack model in grain-based DEM. *Comp. Part. Mech.* **7** (4), 645–654. https://doi.org/10.1007/s40571-019-00295-4 (2020).

41. Xie, S. et al. A new integrated intelligent computing paradigm for predicting joints shear strength. *Geosci. Front.* https://doi.org/10.1016/j.gsf.2024.101884 (2024).

42. Irazábal, J., Salazar, F. & Vicente, D. J. A methodology for calibrating parameters in discrete element models based on machine learning surrogates. *Comp. Part. Mech.* **10** (5), 1031–1047. https://doi.org/10.1007/s40571-022-00550-1 (2023).

43. Sun, Z. et al. Pipeline deformation monitoring based on long-gauge fiber-opticsensing systems: Methods, experiments, and engineering applications. *Measurement* **248**, 116911. https://doi.org/10.1016/j.measurement.2025.116911 (2025).

44. Weidner, L. & Walton, G. The influence of training data variability on a supervised machine learning classifier for structure from motion (SfM) point clouds of rock slopes. *Eng. Geol.* **294**, 106344. https://doi.org/10.1016/j.enggeo.2021.106344 (2021).

45. Hu, M. S., Rutqvist, J. & Steefel, C. I. Mesh generation and optimization from digital rock fractures based on neural style transfer. *J. Rock. Mech. Geotech.* **13** (4), 912–919. https://doi.org/10.1016/j.jrmge.2021.02.002 (2021).

46. Shakirov, A., Molchanov, A., Ismailova, L. & Mezghani, M. Quantitative assessment of rock lithology from gamma-ray and mud logging data. *Geoenergy Sci. Eng.* https://doi.org/10.1016/j.geoen.2023.211664 (2023).

47. Iraji, S., Soltanmohammadi, R., Matheus, G. F., Basso, M. & Vidal, A. C. Application of unsupervised learning and deep learning for rock type prediction and petrophysical characterization using multi-scale data. *Geoenergy Sci. Eng.* **230**, 212241. https://doi.org/10.1016/j.geoen.2023.212241 (2023).

48. Durga Kannaiah, P. V. & Maurya, N. K. Machine learning approaches for formation matrix volume prediction from well logs: insights and lessons learned. *Geoenergy Sci. Eng.* **229**, 212086. https://doi.org/10.1016/j.geoen.2023.212086 (2023).

49. Yuvaraj, P., Murthy, A. R., Iyer, N. R., Samui, P. & Sekar, S. Prediction of fracture characteristics of high strength and ultra high strength concrete beams based on relevance vector machine. *Int. J. Damage Mech.* **23** (7), 979–1004. https://doi.org/10.1177/1056789514520796 (2014).

50. Hiba, M., Ibrahim, A. F., Elkatatny, S. & Ali, A. Application of machine learning to predict the failure parameters from conventional well logs. *Arab. J. Sci. Eng.* **47** (9), 11709–11719. https://doi.org/10.1007/s13369-021-06461-2 (2022).

51. Mahmoodzadeh, A. et al. Machine learning techniques to predict rock strength parameters. *Rock. Mech. Rock. Eng.* **55** (3), 1721–1741. https://doi.org/10.1007/s00603-021-02747-x (2022).

52. Siddig, O., Ibrahim, A. F. & Elkatatny, S. Estimation of rocks' failure parameters from drilling data by using artificial neural network. *Sci. Rep.* **13** (1), 3146. https://doi.org/10.1038/s41598-023-30092-2 (2023).

53. Khandelwal, M. et al. Implementing an ANN model optimized by genetic algorithm for estimating cohesion of limestone samples. *Eng. Comput.* **34** (2), 307–317. https://doi.org/10.1007/s00366-017-0541-y (2018).

54. Kainthola, A. et al. Prediction of strength parameters of Himalayan rocks: A statistical and Anfis approach. *Geotech. Geol. Eng.* **33** (5), 1255–1278. https://doi.org/10.1007/s10706-015-9899-z (2015).

55. Mohammadi Behboud, M., Ramezanzadeh, A., Tokhmechi, B., Mehrad, M. & Davoodi, S. Estimation of Geomechanical rock characteristics from specific energy data using combination of wavelet transform with ANFIS-PSO algorithm. *J. Pet. Explor. Prod. Te.* **13** (8), 1715–1740. https://doi.org/10.1007/s13202-023-01644-z (2023).

56. Shahani, N. M. et al. Predicting angle of internal friction and cohesion of rocks based on machine learning algorithms. *Mathematics* **10** (20), 3875 (2022).

57. Chen, W. et al. Assessing cohesion of the rocks proposing a new intelligent technique namely group method of data handling. *Eng. Comput.* **36** (2), 783–793. https://doi.org/10.1007/s00366-019-00731-2 (2020).

58. Kosova, F., Altay, Ö. & Ünver, H. Structural health monitoring in aviation: a comprehensive review and future directions for machine learning. *Nondestruct. Test. Eval.* https://doi.org/10.1080/10589759.2024.2350575 (2024).

59. Zhang, K. et al. Displacement prediction of step-like landslides based on feature optimization and VMD-Bi-LSTM: a case study of the Bazimen and Baishuihe landslides in the three Gorges, China. *Bull. Eng. Geol. Environ.* **80** (11), 8481–8502. https://doi.org/10.1007/s10064-021-02454-5 (2021).

60. Hodneland, E. et al. Impact of MRI radiomic feature normalization for prognostic modelling in uterine endometrial and cervical cancers. *Sci. Rep.* https://doi.org/10.1038/s41598-024-66659-w (2024).

61. Daghistani, F. & Abuel-Naga, H. Evaluating the influence of sand particle morphology on shear strength: A comparison of experimental and machine learning approaches. *Appl. Sci.* https://doi.org/10.3390/app13148160 (2023).

62. Nam, C. Prediction of mechanical properties of high-entropy ceramics by deep learning with compositional descriptors. *Mater. Today Commun.* https://doi.org/10.1016/j.mtcomm.2023.105949 (2023).

63. Tariq, Z., Gudala, M., Yan, B. C., Sun, S. Y. & Mahmoud, M. A fast method to infer nuclear magnetic resonance based effective porosity in carbonate rocks using machine learning techniques. *Geoenergy Sci. Eng.* https://doi.org/10.1016/j.geoen.2022.211333 (2023).

64. Kannaiah, P. V. D. & Maurya, N. K. Machine learning approaches for formation matrix volume prediction from well logs: insights and lessons learned. *Geoenergy Sci. Eng.* https://doi.org/10.1016/j.geoen.2023.212086 (2023).

65. Mousavi, M., Gandomi, A. H., Holloway, D., Berry, A. & Chen, F. Machine learning analysis of features extracted from time-frequency domain of ultrasonic testing results for wood material assessment. *Constr. Build. Mater.* https://doi.org/10.1016/j.conbuildmat.2022.127761 (2022).

66. Zhou, S. et al. Estimating dynamic compressive strength of rock subjected to freeze-thaw weathering by data-driven models and non-destructive rock properties. *Nondestruct. Test. Eva.* https://doi.org/10.1080/10589759.2024.2313569 (2024).

67. Zhou, C. et al. A novel framework for landslide displacement prediction using MT-InSAR and machine learning techniques. *Eng. Geol.* **334**, 107497. https://doi.org/10.1016/j.enggeo.2024.107497 (2024).

68. Xie, S., Han, Z. & Lin, H. A quantitative model considering crack closure effect of rock materials. *Int. J. Solids Struct.* **251**, 111758. https://doi.org/10.1016/j.ijsolstr.2022.111758 (2022).

69. Xie, S. J. et al. Constitutive modeling of rock materials considering the void compaction characteristics. *Arch. Civ. Mech. Eng.* https://doi.org/10.1007/s43452-022-00378-9 (2022).

70. Xie, S., Han, Z., Hu, H. & Lin, H. Application of a novel constitutive model to evaluate the shear deformation of discontinuity. *Eng. Geol.* https://doi.org/10.1016/j.enggeo.2022.106693 (2022).

71. Chen, T. & Guestrin, C. (eds) Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; (2016).

72. Benemaran, R. S. Application of extreme gradient boosting method for evaluating the properties of episodic failure of borehole breakout. *Geoenergy Sci. Eng.* https://doi.org/10.1016/j.geoen.2023.211837 (2023).

73. Golkarian, A., Khosravi, K., Panahi, M. & Clague, J. J. Spatial variability of soil water erosion: comparing empirical and intelligent techniques. *Geosci. Front.* https://doi.org/10.1016/j.gsf.2022.101456 (2023).

74. Furtado, C. et al. A methodology to generate design allowables of composite laminates using machine learning. *Int. J. Solids Struct.* https://doi.org/10.1016/j.ijsolstr.2021.111095 (2021).

75. Xie, J. W., Huang, J. S., Zeng, C., Huang, S. & Burton, G. J. A generic framework for geotechnical subsurface modeling with machine learning. *J. Rock. Mech. Geotech.* **14** (5), 1366–1379. https://doi.org/10.1016/j.jrmge.2022.08.001 (2022).

76. Xue, J. K. & Shen, B. A novel swarm intelligence optimization approach: sparrow search algorithm. *Syst. Sci. Control Eng.* **8** (1), 22–34. https://doi.org/10.1080/21642583.2019.1708830 (2020).

77. Awadallah, M. A., Al-Betar, M. A., Doush, I. A., Makhadmeh, S. N. & Al-Naymat, G. Recent versions and applications of sparrow search algorithm. *Arch. Comput. Methods Eng.* **30** (5), 2831–2858. https://doi.org/10.1007/s11831-023-09887-z (2023).

78. Joshi, D. A., Menon, R., Jain, R. K. & Kulkarni, A. V. Deep learning based concrete compressive strength prediction model with hybrid meta-heuristic approach. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2023.120925 (2023).

79. Li, B. & Qiu, J. Displacement prediction of open-pit mine slope based on SSA-ELM. *Front. Earth Sci.* https://doi.org/10.3389/feart.2023.1126394 (2023).
80. Yao, H., Song, G. & Li, Y. Displacement prediction of channel slope based on EEMD-IESSA-LSSVM combined algorithm. *Appl. Sci.* **13** (17), 9582 (2023).
81. Mai, H-V-T., Nguyen, M. H. & Ly, H-B. Development of machine learning methods to predict the compressive strength of fiber-reinforced self-compacting concrete and sensitivity analysis. *Constr. Build. Mater.* **367**, 130339. https://doi.org/10.1016/j.conbuildmat.2023.130339 (2023).
82. Xie, S., Lin, H., Ma, T., Peng, K. & Sun, Z. Prediction of joint roughness coefficient via hybrid machine learning model combined with principal components analysis. *J. Rock. Mech. Geotech.* https://doi.org/10.1016/j.jrmge.2024.05.059 (2024).
83. Zhang, W., Wu, C., Zhong, H., Li, Y. & Wang, L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on bayesian optimization. *Geosci. Front.* **12** (1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007 (2021).
84. Hu, X. et al. Predicting triaxial compressive strength of high-temperature treated rock using machine learning techniques. *J. Rock. Mech. Geotech.* **15** (8), 2072–2082. https://doi.org/10.1016/j.jrmge.2022.10.014 (2023).
85. Nguyen, T-T., Cao, B-T., Pham, V-V., Bui, H-G. & Do, N-A. Design optimization of quasi-rectangular tunnels based on hyperstatic reaction method and ensemble learning. *J. Rock. Mech. Geotech.* https://doi.org/10.1016/j.jrmge.2024.10.020 (2024).
86. Bilgehan, M. A comparative study for the concrete compressive strength Estimation using neural network and neuro-fuzzy modelling approaches. *Nondestruct Test. Eval.* **26** (1), 35–55. https://doi.org/10.1080/10589751003770100 (2011).
87. Xie, S., Lin, H., Duan, H. & Chen, Y. Modeling description of interface shear deformation: A theoretical study on damage statistical distributions. *Constr. Build. Mater.* **394**, 132052. https://doi.org/10.1016/j.conbuildmat.2023.132052 (2023).
88. Xie, S. J. et al. Statistical damage shear constitutive model of rock joints under seepage pressure. *Front. Earth Sci.* **8**, 16. https://doi.org/10.3389/feart.2020.00232 (2020).
89. Vásárhelyi, B. & Davarpanah, M. Influence of water content on the mechanical parameters of the intact rock and rock mass. *Periodica Polytechnica-Civil Eng.* **62** (4), 1060–1066. https://doi.org/10.3311/PPci.12173 (2018).
90. Sun, Z. et al. Investigation of electrical resistivity for fiber-reinforced coral aggregate concrete. *Constr. Build. Mater.* **414**, 135011. https://doi.org/10.1016/j.conbuildmat.2024.135011 (2024).

Acknowledgements

Author contributions

Conceptualization, Daxing Lei.; methodology, Daxing Lei.; software, Yaoping Zhang.; validation, Zhigang Lu.; investigation, Yifan Chen.; resources, Zhigang Lu.; data curation, Min Lin and Guangli Wang; writing—original draft preparation, Daxing Lei.; visualization, Guangli Wang.; funding acquisition, Daxing Lei. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.L.

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.