



ARTICLE

# Dual-Objective XGBoost Prediction Model for the Cementation Performance of MICP-Treated Sandy Soil in Small-Sample Scenarios

Pingang Tang<sup>1</sup>, Guang Zhu<sup>1</sup>, Junjun Xu<sup>1</sup>, Chaojian Yue<sup>1</sup>, Wei He<sup>1</sup>, Jun Sun<sup>2</sup> and Chen Zeng<sup>3,\*</sup>

<sup>1</sup>China State Construction International Engineering Limited, Wuhan, China

<sup>2</sup>School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan, China

<sup>3</sup>School of Civil Engineering and Architecture, Wuhan Polytechnic University, Wuhan, China

\*Corresponding Author: Chen Zeng. Email: zengchen@whpu.edu.cn

Received: 15 April 2026; Accepted: 25 June 2026; Published: 23 October 2026

**ABSTRACT:** Microbially Induced Carbonate Precipitation (MICP) is an environmentally friendly technique for sandy soil stabilization. However, the cementation performance is governed by multiple coupled factors and complex experimental procedures, making accurate prediction challenging. In this study, a dual-objective XGBoost prediction model suitable for small-sample scenarios is developed from 77 sets of laboratory data to rapidly estimate unconfined compressive strength (UCS) and calcium carbonate content (CCC) separately. A mechanism-guided feature engineering strategy is adopted to construct three cross features, including urease activity coupled with curing time, calcium carbonate content combined with curing time, and urea-calcium concentration, together with five key influencing parameters. Five-fold cross-validation is used to ensure model stability. In the UCS model, soil particle size fraction (29.94%) and urease activity (20.85%) dominate, while in the CCC model, soil particle size fraction (22.47%) and urease activity (17.63%) prevail, both align well with fundamental MICP mechanisms. The CCC model achieves a coefficient of determination ( $R^2$ ) of 0.6342 and a mean absolute error (MAE) of 2.92%, showing reliable predictive ability. The UCS model achieved an  $R^2$  of 0.7991 and a MAE of 844.83 kPa. However, due to the mathematical amplification of relative error, a small portion of low-strength specimens produced abnormally high MAPE (116.60%), limiting the formal engineering design of the UCS model. SHAP (SHapley Additive exPlanations) analysis is further employed to enhance model interpretability and to quantitatively clarify the marginal contributions and interaction effects of the input features. The proposed framework offers a valuable reference for parameter analysis and mechanistic interpretation of MICP-treated soils. At the same time, the larger prediction deviation for UCS highlights the intrinsic uncertainty of strength evolution in such complex multi-factor systems.

**KEYWORDS:** Microbially induced carbonate precipitation; small-sample prediction; XGBoost; feature engineering; unconfined compressive strength; calcium carbonate content

## 1 Introduction

With the advancement of infrastructure construction into complex geological regions in China, challenges in sand foundation reinforcement have become increasingly prominent [1]. Traditional methods such as grouting and dynamic compaction are increasingly incompatible with the global trend of green and low-carbon geotechnical engineering [2,3]. As an eco-friendly soil-improvement technology, microbially induced carbonate precipitation (MICP) has attracted widespread attention for sandy-soil stabilization, owing to its low environmental disturbance, controllable cementation strength, and broad applicability [4]. Notably,

MICP has recently been extended to the sustainable stabilization of various industrial solid wastes, showing great potential for large-scale geoenvironmental applications [5].

Despite these advances, the inherent complexity of MICP poses substantial challenges for accurate performance prediction. Its cementation behavior is jointly governed by microbial traits, soil properties, reaction conditions, and environmental parameters, all of which exhibit highly nonlinear and coupled relationships [6]. This complexity hinders the rapid and precise prediction of MICP performance in engineering practice.

In recent years, machine learning has been gradually applied to MICP performance modeling due to its strong capacity to capture nonlinear relationships [7–10]. Ensemble algorithms such as extreme gradient boosting (XGBoost) have demonstrated promising performance in predicting geomechanical parameters [11–13], verifying the feasibility of data-driven methods for MICP analysis.

Nevertheless, current prediction research still faces three key technical barriers. First, MICP experiments are time-consuming and resource-intensive, leaving most engineering datasets limited to small samples [14]. Existing models largely rely on large-scale data and lack robust validation frameworks tailored for small-sample scenarios, leading to inadequate adaptation and poor generalization [14–17]. Second, prevailing studies predominantly employ raw variables and fail to construct interpretable features that explicitly encode core MICP biochemical couplings, resulting in black-box models with limited mechanistic insight and extrapolative power [7,12]. Third, most models focus on predicting a single index, such as UCS, lacking a unified framework for the independent yet complementary prediction of UCS and CCC, which are both critical for comprehensive engineering assessment [2,6].

To address these barriers, this study develops a mechanism-guided, small-sample-adaptive dual-model XGBoost framework for predicting MICP cementation performance. Based on the mechanism, a feature engineering strategy is adopted to construct interpretable cross features, including urease activity combined with curing time and urea combined with calcium chloride. Grouting time is introduced as an engineering-oriented feature to enhance practical applicability. The XGBoost algorithm with robust regularization is integrated to improve stability and generalization under data-limited conditions. Separate models are established for UCS and CCC prediction, complemented by SHAP analysis to quantify feature contributions and interaction effects. The overall workflow is illustrated in Fig. 1 and comprises the following steps.

- (1) Data collection and preprocessing. Seventy-seven sets of laboratory MICP sandy soil curing data are collected, and data quality is optimized through missing value imputation and outlier filtering.
- (2) Feature engineering. Cross features such as urease activity  $\times$  curing time and urea  $\times$  calcium chloride are constructed based on MICP reaction mechanisms, together with grouting times to reflect field construction practice.
- (3) Model construction. Independent XGBoost models for UCS and CCC are built, with parameter isolation to avoid mutual interference.
- (4) Model validation. Five-fold cross-validation is employed to evaluate stability, and multiple metrics ( $R^2$ , MAE, RMSE, MAPE, MdAPE) are used for a comprehensive assessment.
- (5) Result analysis. Feature importance analysis identifies dominant factors, and discussions are conducted in conjunction with MICP mechanisms to propose engineering application recommendations.

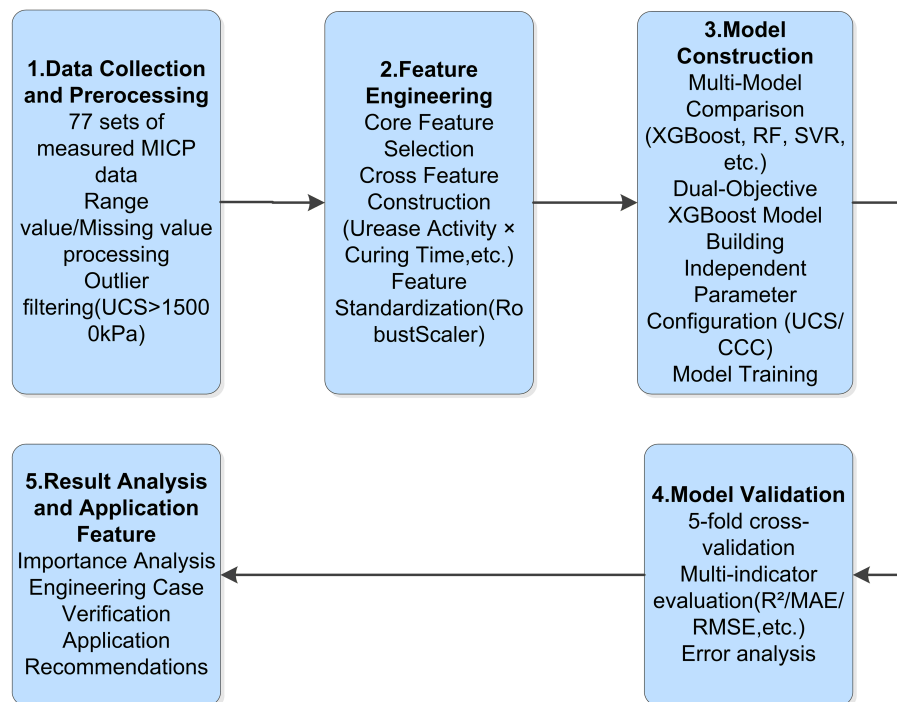


**Figure 1:** Schematic flowchart of the research approach, including data collection, feature engineering, model construction, validation, and result analysis.

## 2 Experimental Data and Model Methods

### 2.1 Source of Experimental Data and Feature Selection

The data used in this study were compiled from domestic and international published literature, consisting of 77 groups of laboratory MICP soil solidification experiments. The dataset covers varying soil particle sizes, urease activity levels, and cement solution concentrations, as well as different curing times, showing good data representativeness and coverage [18].

#### 2.1.1 Input Features

Guided by the MICP reaction mechanism and existing research [6,14,19], five basic input features were first screened. A mechanism-based feature engineering approach was then applied to build cross features, which can capture nonlinear synergistic effects unavailable from individual indicators. Two independent feature subsets were configured for the two prediction targets, respectively.

For the UCS prediction model, seven features were finalized, consisting of five original parameters: soil particle size (mm), urease activity (mM/min), urea concentration (mol/L), calcium chloride concentration (mol/L), and curing time (d). Three cross features were generated, namely urease activity combined with curing time, calcium carbonate content combined with curing time, and urea concentration combined with calcium chloride concentration.

For the CCC prediction model, six features were determined, including six basic indices: soil particle size (mm), urease activity (mM/min), urea concentration (mol/L), calcium chloride concentration (mol/L), curing time (d), and grouting times. One cross feature, urease activity coupled with curing time, was retained to quantify the cumulative hydrolytic potential.

The physical definitions of the derived features are illustrated below. Soil particle size varies from 0.04 to 2.5 mm, covering fine, medium, and coarse sand and controlling pore structure and bacterial attachment efficiency. Urease activity ranges between 5.0 and 25.0 mM/min and governs the rate of urea hydrolysis. The cross feature of urease activity combined with curing time estimates the total catalytic input during the whole reaction process and serves as a critical controlling factor for calcium carbonate production [6]. Urea concentration varies from 0.5 to 2.0 mol/L, and calcium chloride concentration ranges from 0.2 to 1.0 mol/L; their combination describes the stoichiometric matching condition for carbonate nucleation [14]. Grouting times are classified according to curing length and represent practical construction arrangements to improve the uniform distribution of precipitates.

### 2.1.2 Output Target Variable

Unconfined compressive strength (UCS, kPa) was tested by a standard compression apparatus, with original values ranging from 82 to 14,000 kPa. One outlier over 15,000 kPa was eliminated according to engineering criteria [20], leaving 76 effective samples. Calcium carbonate content (CCC, %) falls in the range of 5% to 30% [18], and all 77 samples are available for subsequent modeling.

## 2.2 Data Preprocessing

For range-type features such as soil particle size and urease activity, the mean normalization method is adopted. The formula is as follows:

$$\text{mean} = \frac{\min + \max}{2} \quad (1)$$

$$\text{std} = \frac{\max - \min}{2} \quad (2)$$

where mean is the mean of range features, std is the standard deviation of range features, min and max are the minimum and maximum values of range features, respectively.

Missing value imputation adopts a group-based median imputation strategy. Prior to all preprocessing operations, the dataset is split into training and test sets. All median values for imputation are calculated exclusively from the training set according to curing time grouping. Then the missing values within each group are filled using the corresponding group median derived from the training set. Since the dataset is integrated from multiple published studies without a unified experimental grouping standard, we only adopt grouping by curing time to improve the accuracy of median imputation.

Samples are grouped by curing time. Missing value imputation adopts a group-based median imputation strategy. The median of each feature by grouping is calculated according to curing time. Then the missing values within groups are filled. Missing values without corresponding grouping data are filled using the global median. Since the dataset is integrated from different literature and lacks unified experimental grouping, only grouping by curing time is used to improve the accuracy of median imputation for missing values.

Outlier filtering is implemented using a fixed threshold method based on engineering experience for UCS. Samples with UCS values exceeding 15,000 kPa are identified as outliers and directly excluded from the dataset, as such high strength exceeds the conventional range of MICP-treated sandy soil under standard experimental conditions. No outliers are removed for CCC, since all CCC values fall within the reasonable engineering interval of 5%–30%.

Data standardization uses the RobustScaler method to process input features. This method has strong robustness to outliers and is suitable for small-sample data scenarios. The calculation formula is as follows.

$$X_{\text{scaled}} = \frac{X - \text{median}}{IQR} \quad (3)$$

where, median is the median of the feature,  $IQR$  is the interquartile range, which is the difference between the upper quartile and the lower quartile.

## 2.3 Model Selection and Principles

### 2.3.1 Basis for Model Selection

In this study, XGBoost is selected as the core predictive model, primarily based on comprehensive considerations for several reasons. First, XGBoost is well-suited for few-shot scenarios, effectively suppressing overfitting through mechanisms such as L1/L2 regularization and column sampling, resulting in stronger robustness in the MICP data scenario with 77 sample groups [9]. Second, the gradient boosting framework can accurately capture the nonlinear coupling relationship between multiple factors and cementation performance. Alnmr et al. [19] employed the XGBoost model to predict the unconfined compressive strength and elastic modulus of expansive soil, achieving  $R^2$  values of 0.94 and 0.91 in the test set, respectively, verifying its strong fitting capability in predicting soil mechanical parameters and demonstrating superior predictive performance compared to traditional linear regression models [8]. Additionally, a built-in feature importance assessment function is included, allowing for the quantification of the contribution of each feature and addressing the lack of interpretability in “black box” models [12]. Independent parameter configuration for different prediction targets is supported, facilitating the simultaneous prediction of UCS and CCC dual objectives and avoiding interference between objectives.

### 2.3.2 XGBoost Model Principle

XGBoost iteratively trains weak learners, approximating the true value based on the residual optimization of the previous round, with the following mathematical expression:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F} \quad (4)$$

Among them,  $l(y_i, \hat{y}_i)$  is the loss function,  $\Omega(f)$  is the regularization term, used to control model complexity and prevent overfitting [6]. In this study, the squared error loss function is adopted for both the UCS and CCC prediction models, as it is well-suited for continuous numerical prediction tasks and provides stable gradient information for the gradient boosting framework.

### 2.3.3 Baseline Machine Learning Models

To benchmark the performance of the proposed XGBoost framework, three widely adopted baseline regression algorithms were implemented for comparative analysis.

A standard Linear Regression (LR) model using Ordinary Least Squares (OLS) was employed as a fundamental baseline. In contrast, a non-linear Support Vector Regression (SVR) model with a Radial Basis Function (RBF) kernel was utilized, configured with a penalty parameter  $C$  of 1.0, an epsilon of 0.1, and the kernel coefficient gamma set to ‘scale’. Additionally, an ensemble Random Forest Regressor (RF) was constructed, comprising 100 decision trees ( $n_{\text{estimators}} = 100$ ). For this model, the maximum tree depth was unrestricted ( $\text{max\_depth} = \text{None}$ ), the minimum samples required to split a node was 2, and the number

of features considered for the best split was set to the square root of the total feature count (`max_features = 'sqrt'`). All baseline models were trained and evaluated using the identical 5-fold cross-validation procedure applied to the XGBoost model to ensure a consistent and fair comparative analysis.

## 2.4 Model Parameter Settings

To achieve optimal performance for dual-objective prediction, grid search and cross-validation are used to configure independent parameters for the UCS and CCC prediction models, with specific parameter settings shown in [Table 1](#).

**Table 1:** XGBoost model parameter settings.

| Parameter Name   | UCS Prediction Model | CCC Prediction Model | Parameter Description          |
|------------------|----------------------|----------------------|--------------------------------|
| Objective        | reg:squarederror     | reg:squarederror     | Loss function type             |
| max_depth        | 4                    | 4                    | Maximum depth of decision tree |
| learning_rate    | 0.05                 | 0.04                 | Learning rate                  |
| n_estimators     | 300                  | 500                  | Number of decision trees       |
| reg_alpha        | 0.5                  | 1.0                  | L1 regularization coefficient  |
| reg_lambda       | 1.0                  | 2.0                  | L2 regularization coefficient  |
| Subsample        | 0.8                  | 0.8                  | Sample sampling ratio          |
| colsample_bytree | 0.8                  | 0.8                  | Feature sampling ratio         |
| random_state     | 42                   | 42                   | Random seed                    |

The objective is set as `reg: squarederror` for both the UCS and CCC models, as the squared error loss function is suitable for the continuous numerical prediction tasks in this study. The `max_depth` is restricted to 4 for both models to balance model complexity and generalization ability in small-sample scenarios. The `learning_rate` and `n_estimators` are matched to ensure stable iteration and adequate learning ability, with values of 0.05 and 300 for UCS, and 0.04 and 500 for CCC, respectively. The `reg_alpha` and `reg_lambda` provide L1 and L2 regularization to further constrain model complexity, set at 0.5 and 1.0 for UCS, and 1.0 and 2.0 for CCC. The `subsample` and `colsample_bytree` parameters realize random sampling of samples and features during training, both set to 0.8 for both models to effectively reduce overfitting risk. The `random_state` is fixed at 42 to ensure result repeatability. Other parameters such as `n_jobs`, `verbosity`, and `eval_metric` are set to default values.

## 2.5 Model Validation Methods and Evaluation Metrics

### 2.5.1 Dataset Splitting

The preprocessed data is divided into training and test sets. For UCS prediction, stratified random sampling is adopted to ensure a balanced distribution of strength levels across sets, with the test set proportion being 20%. In the CCC prediction, a standard random split is used. Consequently, the UCS training set has 60 samples, and the test set has 16 samples, while the CCC training set has 61 samples and the test set has 16 samples.

### 2.5.2 Cross-Validation Method

A 5-fold cross-validation is performed exclusively on the training set to evaluate the model's stability and generalization ability under small-sample conditions [17]. The training data is partitioned into 5 non-overlapping subsets. In each iteration, 4 subsets are used for training and 1 for validation. This process repeats 5 times, and the average performance metrics are calculated to mitigate the impact of data partitioning bias [7].

### 2.5.3 Evaluation Metrics

To comprehensively assess the model's predictive performance, a multi-dimensional evaluation metric system is adopted. The calculation formulas for each evaluation metric are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i + 1 \times 10^{-8}} \right| \times 100\% \quad (8)$$

$$MdAPE = \text{median} \left( \left| \frac{y_i - \hat{y}_i}{y_i + 1 \times 10^{-8}} \right| \times 100\% \right) \quad (9)$$

where,  $y_i$  is the true value of the  $i$ -th sample,  $\hat{y}_i$  is the predicted value of the  $i$ -th sample,  $\bar{y}$  is the average of the true values, and  $n$  is the number of samples [12].

The  $R^2$  represents the model's ability to explain the variability of the data, with a range of [0, 1]. The closer it is to 1, the better the prediction performance.

The MAE represents the average absolute deviation between the predicted values and the true values, with lower values being better. The RMSE represents the root mean square deviation between the predicted values and the true values, and is sensitive to extreme errors, with lower values being better.

The MAPE represents the average of the relative errors and is suitable for comparing errors across data of different magnitudes.

The MdAPE represents the median of the relative errors, is less sensitive to outliers, and can better reflect the overall prediction stability of the model.

### 2.6 Feature Importance Analysis Method

The built-in feature importance evaluation function of the XGBoost model is used, which calculates the importance score for each feature based on the weights assigned at decision tree splits [12]. The higher the feature importance score, the greater the feature's contribution during model training and the more significant its impact on prediction results. For ease of comparative analysis, the importance scores of each feature are converted to percentages using the following formula.

$$\text{importance\_ratio} = \frac{\text{feature\_score}}{\sum_j \text{feature\_score}_j} \times 100\% \quad (10)$$

where, importance\_ratio represents the proportion of feature importance, feature\_score is the importance score of a single feature, and sum\_all\_feature\_scores is the sum of all feature importance scores [6].

### 3 Model Establishment and Result Analysis

#### 3.1 Data Preprocessing Results

After range value processing, missing value filling, and outlier filtering, 76 sets of UCS valid data and 77 sets of calcium carbonate content valid data were obtained. The comparison of key statistical features before and after data preprocessing is shown in Table 2.

**Table 2:** Comparison of key statistical features before and after data preprocessing.

| Feature                                | Before Preprocessing |                |           | After Preprocessing |                |           |
|----------------------------------------|----------------------|----------------|-----------|---------------------|----------------|-----------|
|                                        | Mean                 | Std. Deviation | Missing % | Mean                | Std. Deviation | Missing % |
| Soil particle size_mean (mm)           | 0.58                 | 0.32           | 5.19      | 0.59                | 0.31           | 0         |
| Urease activity_mean (mM/min)          | 12.85                | 4.67           | 3.90      | 12.91               | 4.58           | 0         |
| Urea concentration (mol/L)             | 1.02                 | 0.35           | 0         | 1.02                | 0.35           | 0         |
| Calcium chloride concentration (mol/L) | 0.65                 | 0.21           | 0         | 0.65                | 0.21           | 0         |
| Curing time (d)                        | 14.62                | 6.83           | 0         | 14.62               | 6.83           | 0         |
| UCS (kPa)                              | 5862.31              | 3215.78        | 1.29      | 5789.45             | 3089.62        | 0         |
| CCC (%)                                | 15.68                | 5.83           | 0         | 15.68               | 5.83           | 0         |

In terms of missing value handling, 5.19%, 3.90%, and 1.29% of values were missing for soil particle size, urease activity, and UCS, respectively, before preprocessing. After imputing missing values, the proportion of missing values across all features was reduced to zero, achieving sample set completeness and avoiding model training bias from data loss.

Regarding data distribution stability, the mean of UCS before preprocessing was 5862.31 kPa with a standard deviation of 3215.78 kPa. After preprocessing, the mean decreased to 5789.45 kPa, with the standard deviation reduced to 3089.62 kPa, indicating the effective removal of extreme outliers, reduced data dispersion, and a more concentrated distribution. The mean of soil particle size before preprocessing was 0.58 mm with a standard deviation of 0.32 mm. After preprocessing, the mean increased slightly to 0.59 mm, while the standard deviation decreased slightly to 0.31 mm, indicating a slight narrowing of data fluctuations. The mean urease activity before preprocessing was 12.85 mM/min with a standard deviation of 4.67 mM/min; after preprocessing, the mean increased slightly to 12.91 mM/min, with a slightly reduced standard deviation of 4.58 mM/min, reflecting a more stable data distribution. The means and standard deviations of the remaining features remained stable, reflecting the stability of the original data.

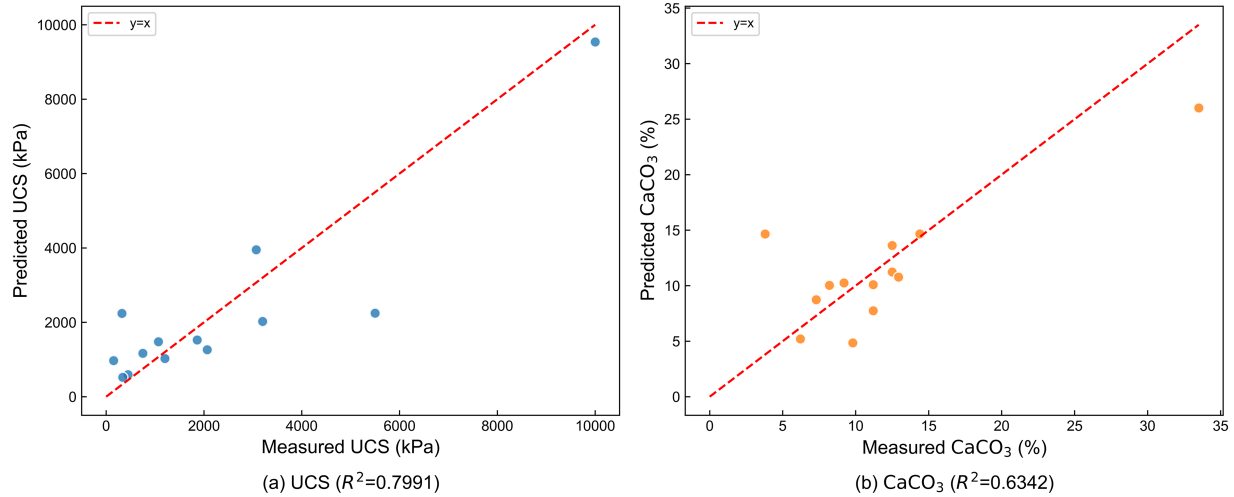
As shown in Table 2, the preprocessing pipeline effectively imputed missing values and removed extreme outliers while preserving the core distributional trends. It significantly improved the completeness and stability of the sample set, particularly by reducing the dispersion of UCS, providing a high-quality, reliable dataset foundation for subsequent machine learning model training [21].

#### 3.2 Model Prediction Results and Analysis

The performance metrics of the XGBoost dual-objective prediction model on the test set are shown in Table 3, and the comparison between predicted and actual values is shown in Fig. 2.

**Table 3:** Prediction performance indicators of the model test set.

| Target Prediction | R <sup>2</sup> | MAE    | RMSE    | MAPE (%) | MdAPE (%) |
|-------------------|----------------|--------|---------|----------|-----------|
| UCS (kPa)         | 0.7991         | 844.83 | 1193.03 | 116.60   | 38.32     |
| CCC (%)           | 0.6342         | 2.92   | 4.18    | 39.98    | 16.76     |

**Figure 2:** Comparison between predicted and measured values of the XGBoost model: (a) unconfined compressive strength (UCS, kPa); (b) calcium carbonate content (CCC, %). The dashed line represents the ideal fitting line ( $y = x$ ).

It can be seen from Table 3 and Fig. 2 that the performance of the XGBoost dual-objective prediction model varies significantly across different prediction targets, with the calcium carbonate content prediction model showing overall better performance.

In terms of UCS prediction, the  $R^2$  of the model is 0.7991, which can explain about 79.91% of UCS variation, with MAE of 844.83 kPa, RMSE of 1193.03 kPa, MAPE of 116.60%, and MdAPE of 38.32%. It is indicated that the UCS prediction accuracy is relatively limited, with a high degree of dispersion, which may be related to the complex factors affecting UCS, such as soil structure and cementation distribution. A large number of studies have shown that due to the strong coupling effects of factors such as microbial activity, uniformity of cementation fluid distribution, soil pore structure, and the micro morphology and spatial distribution of calcium carbonate precipitation, the UCS produced during the MICP process has a high degree of dispersion and non-linearity [8,22]. Usman et al. [6] also pointed out that when constructing the UCS prediction model, even with optimized algorithms, UCS prediction still faces the challenge of high data variability. Therefore, the prediction performance presented by this model with limited samples is essentially an objective reflection of the inherent uncertainty of this complex system, rather than a shortcoming of the model itself.

In terms of CCC prediction, the model performance is significantly improved, with the coefficient of determination  $R^2$  reaching 0.6342, which can explain 63.42% of the CCC variation. The MAE is 2.92%, the RMSE is 4.18%, and the MAPE is 39.98%, while the MdAPE is 16.76%. The predicted values of CCC cluster around the 45° equal-value line, indicating good agreement with the actual values and smaller data dispersion, reflecting a more stable fit. An MdAPE of 16.76% means that the relative error of most samples is controlled within 17%, which is a precision level lower than the conventional error threshold for micro-parameter prediction in geotechnical engineering, typically not exceeding 20% [8]. The accuracy

meets the practical needs of rapid on-site evaluation of curing quality and guiding the optimization of construction parameters.

### 3.3 Comparison of Different Model Performances

To objectively evaluate the performance advantage of the XGBoost model constructed in this study for MICP small-sample prediction tasks, three representative machine learning models were selected as baselines for comparison, including Linear Regression, SVR, and random forest. Linear regression represents the most basic linear assumption model, used to test whether there is a significant linear relationship between MICP cementation performance and various factors. SVR is often used for comparison due to its potential in handling small samples and nonlinear problems. Random forest, like XGBoost, belongs to the ensemble learning algorithm, but it is based on the Bagging idea. Comparing the two can reveal the advantage of the Boosting framework in this task. Specific values are shown in [Tables 4 and 5](#).

**Table 4:** Comparison of UCS prediction performance.

| Model             | R <sup>2</sup> | MAE (kPa) | RMSE (kPa) | MAPE (%) | MdAPE (%) |
|-------------------|----------------|-----------|------------|----------|-----------|
| Linear Regression | 0.1815         | 1599.20   | 2408.16    | 262.01   | 73.56     |
| SVR (RBF)         | 0.3166         | 1499.65   | 2200.54    | 115.62   | 67.30     |
| Random Forest     | 0.5622         | 1275.49   | 1761.33    | 254.36   | 48.84     |
| XGBoost           | 0.7991         | 844.83    | 1193.03    | 116.60   | 38.32     |

**Table 5:** CCC prediction performance comparison.

| Model             | R <sup>2</sup> | MAE (%) | RMSE (%) | MAPE (%) | MdAPE (%) |
|-------------------|----------------|---------|----------|----------|-----------|
| Linear Regression | −0.3068        | 5.05    | 7.89     | 43.49    | 23.86     |
| SVR (RBF)         | −0.1330        | 4.31    | 7.35     | 40.18    | 18.87     |
| Random Forest     | 0.4991         | 3.51    | 4.89     | 37.73    | 17.82     |
| XGBoost           | 0.6342         | 2.92    | 4.18     | 39.98    | 16.76     |

As shown in [Table 4](#), the XGBoost model significantly outperforms traditional models such as linear regression, SVR, and random forest in UCS prediction. The R<sup>2</sup> of XGBoost reaches 0.7991, far exceeding the levels of other models. The MAE of the XGBoost model is 844.83 kPa, and the RMSE is 1193.03 kPa, both of which are the lowest among all models. However, the MAPE is notably high at 116.60%, primarily driven by samples with very low true strength values where relative errors are mathematically amplified [6]. In contrast, the MdAPE is only 38.32%, significantly lower than the range of 67.30% to 182.53% of other models. This indicates that in small-sample scenarios, the XGBoost model significantly mitigates the risk of overfitting through gradient boosting iteration optimization, built-in regularization (L1/L2), and row/column sampling mechanisms [9,14], thereby achieving significantly higher prediction accuracy and lower overall errors for the majority of samples.

As shown in [Table 5](#), the XGBoost model also outperforms the others in CCC prediction, achieving the highest R<sup>2</sup> (0.6342) and the lowest RMSE (4.18%) and MdAPE (16.76%). Random Forest has certain predictive capabilities, but all error metrics are higher than those of XGBoost. Linear Regression and SVR perform poorly, with SVR's R<sup>2</sup> even being negative, unable to effectively establish a correlation between features and calcium carbonate content. This indicates that XGBoost has higher prediction accuracy and

stability for calcium carbonate content, making it more suitable for rapid field assessment needs. The superior performance of XGBoost can be attributed to its ability to capture complex non-linear interactions, such as the synergistic effect between urease activity and curing time, which is critical for CCC prediction [23].

### 3.4 Feature Importance Analysis Results

#### 3.4.1 Feature Importance of the UCS Prediction Model

The feature importance analysis results of the UCS prediction model are shown in Fig. 3.

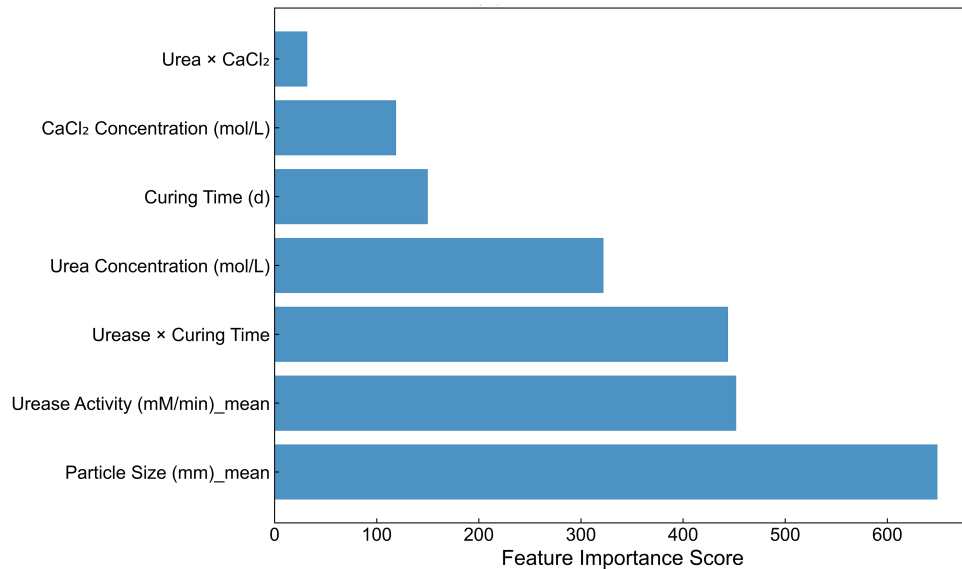


**Figure 3:** Feature importance ranking of the UCS prediction model. The vertical axis represents the feature name, and the horizontal axis represents the importance score.

The feature importance analysis results of the UCS prediction model are shown in Fig. 3. Particle Size (mm)\_mean is identified as the most critical factor (29.94%), which fundamentally regulates the micro-mechanical environment of cementation. It governs the mechanical response beyond macroscopic pore structure by controlling pore geometry and the permeability efficiency of cementation fluid [11]. The elastoplastic constitutive model established by Cui et al. [24] confirms that pore-scale distribution directly controls fluid penetration. Furthermore, based on studies focusing on the mechanical behaviors of fine/coarse soil mixtures [25,26], it can be inferred that the tighter confinement provided by fine particles increases inter-particle contact stress. Thus, for the same CCC value, a well-graded soil achieves higher UCS due to more efficient stress transfer through the dense granular matrix. This finding aligns with the state-of-the-art review by Wang et al. [4], emphasizing the dominance of physical soil properties in MICP performance.

Urease Activity (mM/min)\_mean, ranking second (20.85%), is a key driving factor in urea hydrolysis, directly determining the rate of calcium carbonate generation [4]. Related experimental studies have confirmed that changes in urease activity significantly alter the kinetics of cementation reactions.

Curing Time (d), ranking third (6.92%), validates the engineering understanding that longer curing time leads to more calcium carbonate accumulation [22]. The deposition amount continues to increase with time, gradually enhancing inter-particle cementation strength.

The cross-feature Urease combined with Curing Time (20.48%) also exhibits high importance, indicating that the cumulative catalytic effect is vital for strength development.

In contrast, the cross-feature importance of Urea combined with  $\text{CaCl}_2$  is minimal (1.48%). This is related to the experimental design that primarily adopts equimolar ratios. Erdmann and Strieth [27] pointed out that using equimolar solutions achieves the best precipitation effect, a finding confirmed by Sun et al. [28] regarding the optimal urea concentration around 1.0 mol/L. Therefore, this ratio is widely used in MICP research [4,29], leading to a narrow variance in this interaction term within the dataset.

### 3.4.2 Feature Importance of the CCC Prediction Model

The feature importance analysis results of the CCC prediction model are shown in Fig. 4.

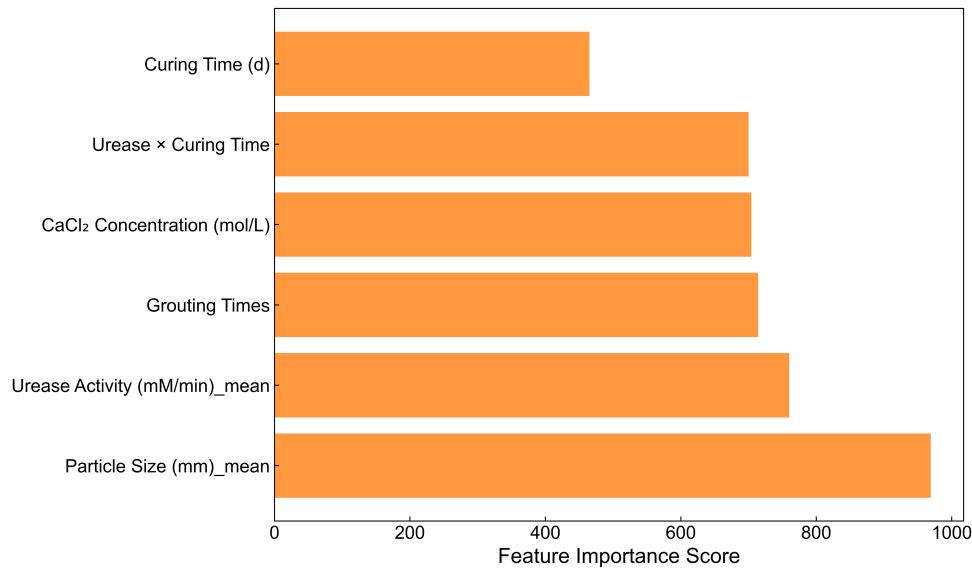


**Figure 4:** Feature importance ranking of the CCC prediction model. The vertical axis represents the feature name, and the horizontal axis represents the importance score.

As shown in Fig. 4, the most critical feature is soil particle size (mm)\_mean (22.47%), followed by urease activity (mM/min)\_mean (17.63%), grouting times (16.56%), calcium chloride concentration (mol/L) (16.33%), and the cross-feature urease activity combined with curing time (16.23%). The top five features collectively account for approximately 89.22% of the total importance, indicating that these physical and operational parameters dominate the prediction of calcium carbonate content.

Soil particle size (mm)\_mean is identified as the primary controlling factor, which aligns with findings from multiple studies. Erdmann et al. [30] indicated that fine sand yields significantly higher calcium carbonate precipitation due to its larger specific surface area and more suitable pore structure compared to coarse sand. Xie et al. [23] also emphasized that soil particle size is a key physical factor regulating the spatial distribution and crystal characteristics of calcium carbonate precipitation. Furthermore, Zhang et al. [31] discovered that particle size fundamentally affects bacterial distribution and the final cementation uniformity. These studies confirm that the physical matrix dictates the deposition efficiency of calcium carbonate.

Urease activity (mM/min)\_mean (17.63%) and the cross-feature urease activity combined with curing time (16.23%) rank second and fifth, respectively. Their combined importance highlights the critical role of biochemical kinetics. As a key catalytic factor in urea hydrolysis, urease activity directly determines the rate of

carbonate generation [27,29]. The cross-feature validates that the coupling effect between enzymatic activity and the time window for crystal growth jointly controls the cumulative amount of calcium carbonate [32].

Interestingly, grouting times (16.56%) emerge as the third most important feature, surpassing the pure biochemical interaction term. This underscores the significance of engineering operations; multiple injections are likely necessary to overcome gravitational segregation and ensure the uniform distribution of bacteria and cementation solution within the soil matrix [3]. This operational parameter proves to be as crucial as the chemical concentrations themselves.

Calcium chloride concentration (mol/L) (16.33%) also holds substantial importance. However, similar to the UCS model, the cross-feature importance of urea combined with calcium chloride is expected to be minimal (not listed in the top ranks). This is attributed to the prevalent use of equimolar ratios in the dataset, which standardizes the chemical driving force and reduces the variance contributed by their interaction [4,28].

### 3.4.3 SHAP Value Analysis

To address the lack of interpretability in traditional black-box machine learning models and further clarify the marginal influence of each feature on the prediction results, the SHAP method is introduced to analyze the proposed dual-objective XGBoost model. SHAP values quantify the contribution of each feature to individual predictions, enabling both global feature importance assessment and local interaction mechanism analysis.

The SHAP summary plot for the CCC shown in Fig. 5 reveals that Soil Particle Size mean is the dominant factor, followed by Grouting Times and Calcium Chloride Concentration. This ranking shifts the focus from purely biochemical kinetics to physical matrix constraints and operational parameters. The high importance of particle size aligns with Erdmann et al. [30], who noted that finer sands provide larger specific surface areas for bacterial adhesion and crystal nucleation.

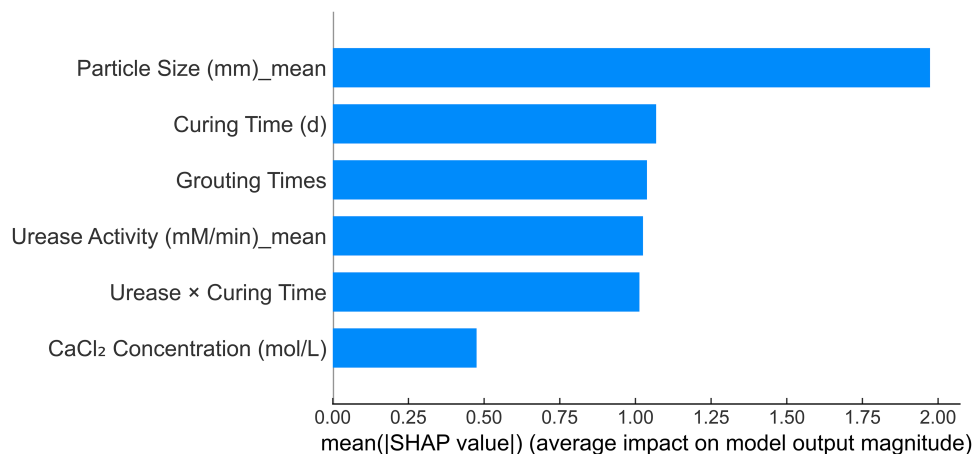


**Figure 5:** SHAP feature importance of CCC.

The dependence plot for CaCl<sub>2</sub> Concentration shown in Fig. 6 exhibits a distinct inverted U-shaped curve. The SHAP values peak at mid-range concentrations near 0.5 mol/L and decline at higher levels. This non-linearity suggests that while moderate calcium supply promotes precipitation, excessive concentrations may induce homogeneous nucleation or pore clogging, inhibiting effective cementation [27,28]. Notably, the

color gradient representing Soil Particle Size shows that fine particles in red achieve higher SHAP values at optimal concentrations, reinforcing the synergy between physical confinement and chemical supply.

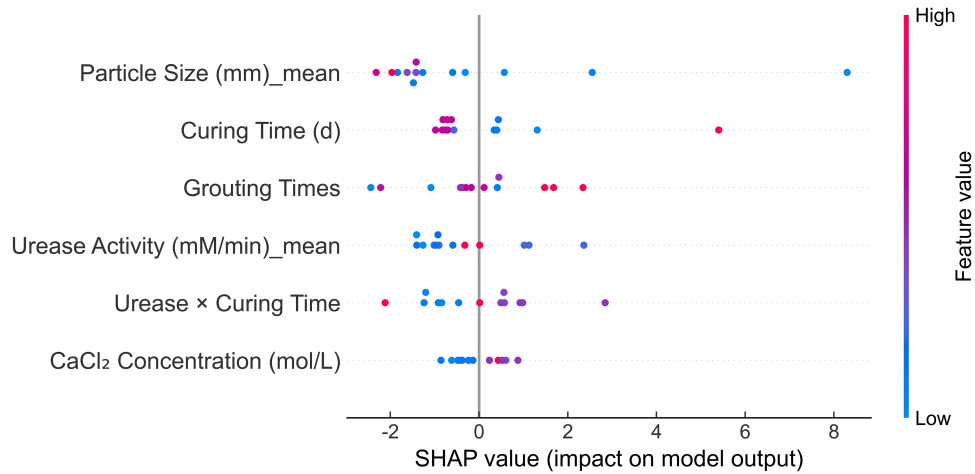


**Figure 6:** SHAP feature impact distribution of CCC.

Furthermore, the interaction between Grouting Times and CaCl<sub>2</sub> Concentration shown in Fig. 7 is critical. High-frequency grouting marked by red or purple points sustains positive SHAP values across a broader concentration range, whereas low-frequency grouting marked by blue points fails to capitalize on high concentrations. This underscores that operational uniformity, achieved through repeated injections, is as vital as chemical stoichiometry for maximizing CaCO<sub>3</sub> yield [3].

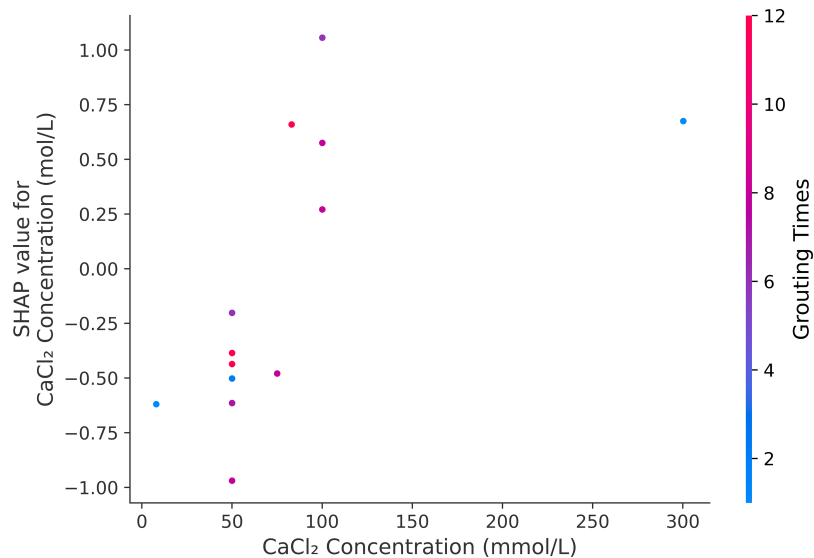


**Figure 7:** SHAP dependence plot of CaCl<sub>2</sub> concentration (mol/L) CCC.

For the UCS model shown in Fig. 8, Soil Particle Size mean again dominates, but Urease Activity mean and Urease combined with Curing Time emerge as significant contributors. This reflects the transition from chemical yield reflected by CCC to structural load-bearing capacity. The SHAP dependence plot for Curing

Time, shown in Fig. 9, shows a positive monotonic trend, confirming that prolonged curing enhances inter-particle bonding [22]. However, the interaction with particle size is evident: fine particles marked in red yield higher SHAP values at the same curing duration, indicating that a dense matrix amplifies the mechanical benefits of extended curing [24,25].

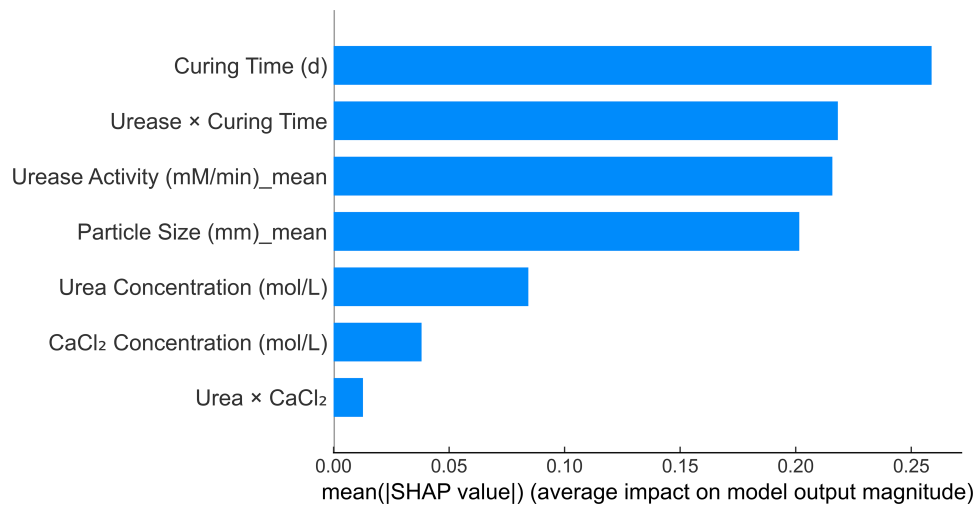


**Figure 8:** SHAP feature importance of UCS.

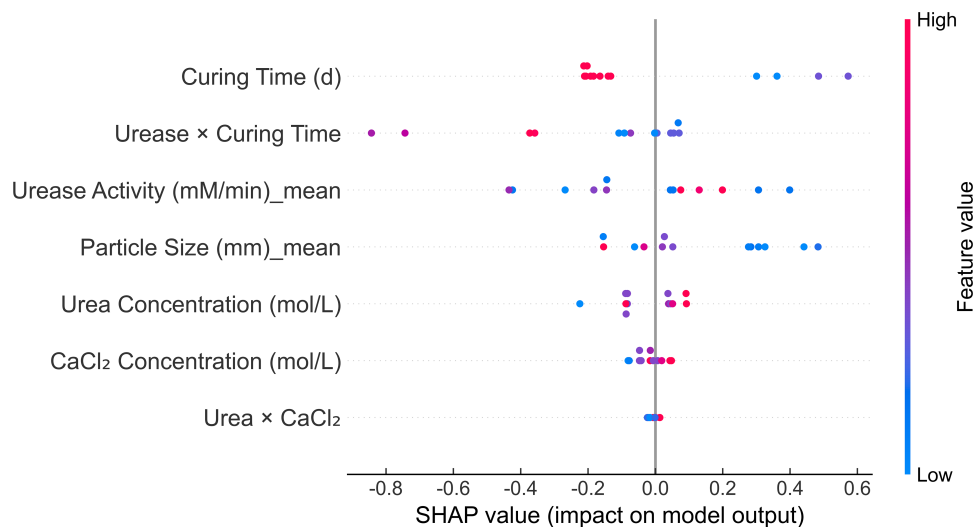


**Figure 9:** SHAP feature impact distribution of UCS.

The dependence plot for Urea Concentration, shown in Fig. 10, reveals a saturation effect. SHAP values increase sharply up to approximately 1.0 mol/L but plateau thereafter. This aligns with Sun et al. [28], who identified 1.0 mol/L as the optimum urea concentration for MICP. Beyond this threshold, additional substrate does not proportionally enhance strength, possibly due to osmotic stress on bacteria or excessive crystal growth at particle contacts rather than bridges [27].

The SHAP analysis validates the model's alignment with MICP mechanisms. For CCC, the physical constraint of particle size and the operational necessity of grouting times outweigh simple chemical inputs. For UCS, the synergy between enzymatic activity and curing time is paramount. Both analyses

highlight the non-linear thresholds, such as  $\text{CaCl}_2$  concentration and urea saturation, that must be navigated in engineering applications. These findings move beyond correlation, providing causal insights into the multi-factor coupling of MICP processes.

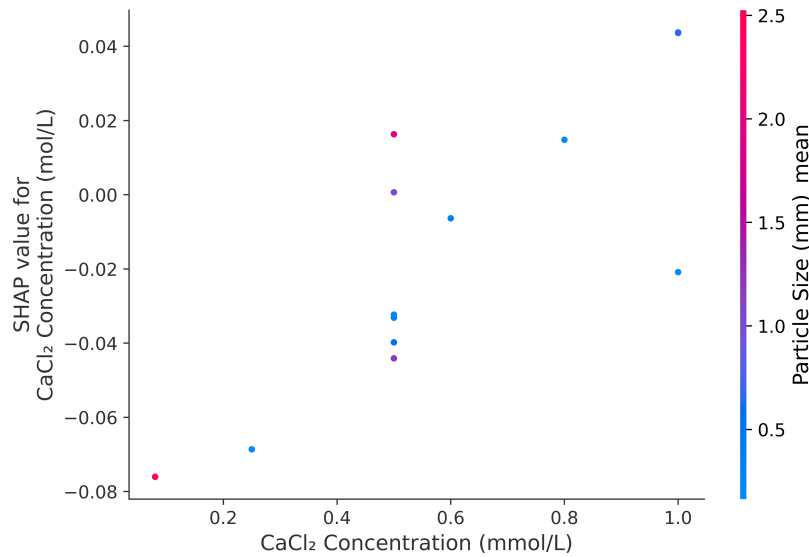


**Figure 10:** SHAP dependence plot of  $\text{CaCl}_2$  concentration (mol/L) UCS.

### 3.5 Error Analysis

#### 3.5.1 UCS Prediction Error Analysis

The histogram of UCS prediction errors is shown in Fig. 11. From the distribution pattern, the UCS prediction errors exhibit a clear right-skewed characteristic, with most absolute errors concentrated in the lower range, near 0 to 1500 kPa. The mean absolute error (MAE) is 844.83 kPa, indicating that the model's prediction accuracy for the majority of samples stays at a reasonable level. However, a long right tail extends close to 3000 kPa, signifying the presence of a few samples with extreme prediction deviations. These outliers significantly inflate the mean-based MAPE to 116.60%, while the more robust MdAPE remains at a considerably lower 38.32%.

The generation of these extreme errors can be attributed to the inherent complexity of the MICP process and the limitations of the dataset. Primarily, significant spatial heterogeneity exists in calcium carbonate precipitation within some specimens. Local variations in the micro-aggregation structure, such as uneven crystal growth at particle contacts vs. void fillings, are difficult to fully capture with only macroscopic features, such as soil particle size and curing time. This micro-scale heterogeneity creates a signal-to-noise issue in which macroscopic descriptors fail to resolve local bonding differences, leading to substantial deviations across specific samples [29]. Furthermore, the dataset contains only 76 groups of data, and samples with extremely low UCSs below 1000 kPa or extremely high UCSs above 10,000 kPa are sparsely represented. When such edge-case regimes are undersampled, the model cannot adequately learn their unique feature-response patterns, leading to larger prediction deviations in the tail region [33]. Finally, gradient-boosted tree models such as XGBoosts are inherently sensitive to noise and outliers under small-sample conditions. Unless sufficiently constrained by regularization techniques, including L1 and L2 penalties controlled by `reg_alpha` and `reg_lambda`, as well as stochastic sampling defined by `subsample` and `colsample_bytree`, the algorithm tends to overfit local perturbations in the training set rather than learning robust, generalizable patterns [9].

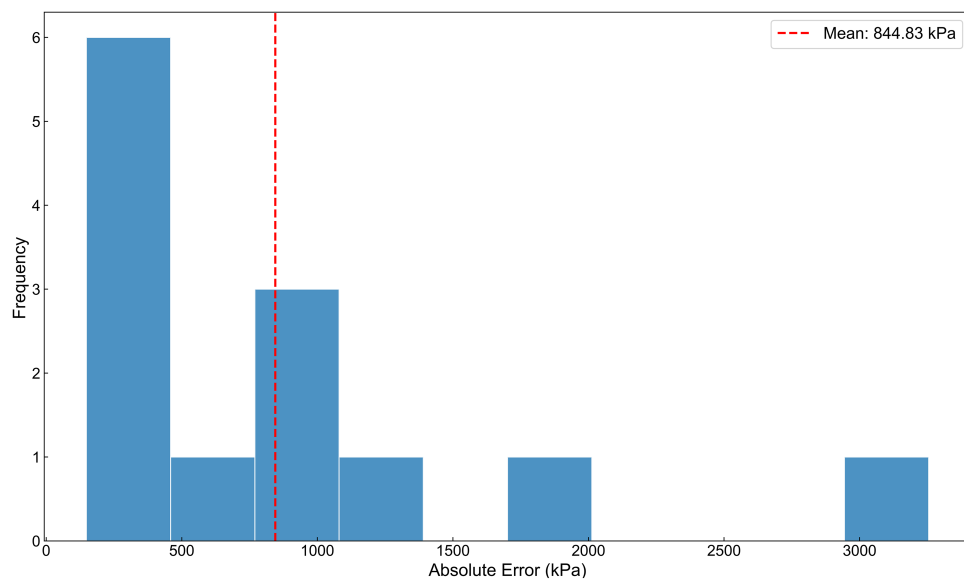


**Figure 11:** Distribution histogram of absolute prediction errors for UCS. The horizontal axis represents the error interval (kPa), and the vertical axis represents the number of samples.

### 3.5.2 CCC Prediction Error Analysis

The histogram of the CCC prediction error distribution is shown in Fig. 12. From the distribution pattern, the prediction errors are highly concentrated overall, exhibiting a strong left-skewed characteristic. The vast majority of samples are clustered within the 0% to 3% error range, with the mean absolute error recorded at 2.92%. This tight clustering indicates that the model demonstrates high precision and robustness in predicting the chemical curing outcomes for most specimens.

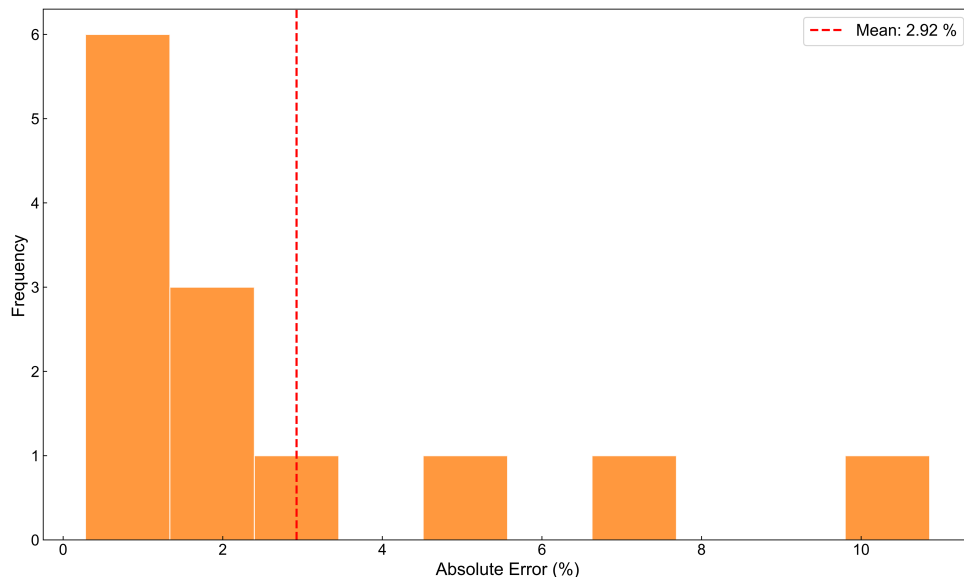


**Figure 12:** Distribution histogram of absolute prediction errors for CCC. The horizontal axis represents the error interval (%), and the vertical axis represents the number of samples.

The low MdAPE value of 16.76% further corroborates the model's reliability. However, a few scattered samples extending beyond 5% indicate minor prediction uncertainties. These discrepancies primarily stem from the inherent challenges in quantifying multi-parameter coupling effects within heterogeneous porous media. Specifically, while the model captures the macroscopic features well, the localized efficiency of microbial attachment and subsequent calcium carbonate precipitation in fine sand samples can fluctuate due to microscale heterogeneities. Furthermore, the non-linear interaction between soil particle size and urease activity, along with the complex transport kinetics of reactants, presents a significant challenge for purely data-driven models to fully encapsulate without residual error [4,34].

#### 4 Suggestions for Model Engineering Application

Based on the aforementioned XGBoost prediction model and its feature importance analysis results, the following suggestions are proposed.

- (1) As the most influential factors on the cementation performance, the soil particle size (0.1–0.6 mm) and urease activity (8–15 mM/min) are suggested to be controlled within the optimal range. Precise parameter control can significantly improve prediction accuracy and practical solidification performance [10], which is consistent with the experimental results of key parameter optimization [4].
- (2) The concentration and ratio of the cementation solution are suggested to be 1.0–1.2 mol/L equimolar urea and calcium chloride. This ratio presents optimal prediction stability and cementation performance in the model. Furthermore, it avoids the inhibition of microbial activity by high salinity, as confirmed by experimental investigations [29].
- (3) The curing time is suggested to be 14–21 days, balancing calcium carbonate accumulation with engineering cost and schedule.
- (4) For extreme scenarios with UCS below 1000 kPa or above 10,000 kPa, the field test data are needed to adjust the modeling prediction. This can be done, for example, by supplementing one or two sets of targeted test data to generate virtual samples and balance data distribution [33], thereby improving prediction accuracy [7].
- (5) With the accumulation of engineering data, it is suggested to regularly supplement new data for model retraining and to expand sample coverage. In scenarios with extremely limited data, multi-task deep transfer learning methods could be introduced to utilize data from other similar projects to improve prediction accuracy [35].

#### 5 Limitations and Recommendations for Future Studies

This study also has limitations. First, the dataset used in this study contains only 77 groups of samples collected from published literature. Although data preprocessing, 5-fold cross-validation, and regularization strategies have been adopted to improve reliability, the relatively limited sample size still restricts the universal generalization ability of the model [7,8]. The prediction accuracy may decrease when applied to new soil types, microbial strains, or unconventional experimental conditions, indicating that the model is currently more suitable for engineering reference and preliminary parameter analysis rather than high-precision field design [3,4,36].

Second, the input features do not include key injection strategies. Common injection methods such as direct injection, pre-mixing, and percolation, as well as grouting rate and cycle times, have been proven to dominate the spatial distribution uniformity of calcium carbonate precipitation and structural integrity [1,3,27]. Different injection methods lead to completely different cementation morphologies and macroscopic mechanical behaviors [27,30], which may cause obvious prediction deviations when the model is extended from laboratory conditions to field-scale engineering [3,4].

Third, important geotechnical parameters such as degree of saturation and relative density are not considered in the model. These parameters significantly affect the penetration, bacterial adhesion, and reaction efficiency during MICP treatment [1,30], and their absence limits the model's ability to capture coupled hydro-mechanical-biochemical effects, especially under variable environmental and placement conditions [34].

Finally, in practical engineering, MICP-treated soils are often subjected to complex environmental actions. Beyond drying-wetting cycles, factors such as freeze-thaw cycles, wind, and rainfall erosion may weaken the bonding strength of calcium carbonate crystals, deteriorate the internal microstructure, and reduce the stability of long-term mechanical performance [1,27,37,38]. This study does not consider the performance degradation caused by these environmental actions, which may lead to inconsistencies between predicted values and long-term field performance [1,30].

Based on the above limitations, future research will focus on establishing a larger and more standardized MICP database covering diverse soil conditions, microbial strains, and construction parameters to further enhance model generalization ability [3,4,8]. Future studies will also incorporate injection modes, grouting parameters, saturation, relative density, and environmental cycle conditions into feature engineering to build a more comprehensive and mechanism-consistent prediction model [35]. In addition, long-term laboratory durability tests and field monitoring data can be combined to quantify performance degradation laws and optimize the model accordingly [1,30]. Finally, a multi-scale and multi-physical field-coupled prediction framework will be developed to promote the transformation from laboratory prediction to practical engineering application [7,34,39].

## 6 Conclusions

This study establishes a dual-objective XGBoost prediction framework specifically adapted for small-sample scenarios to rapidly estimate the unconfined compressive strength (UCS) and calcium carbonate content (CCC) of MICP-treated sandy soil. Leveraging 77 sets of laboratory data, this study adopts a mechanism-guided feature engineering strategy to construct interpretable cross-features such as urease activity combined with curing time. These features effectively capture the nonlinear synergies inherent in biogeochemical mineralization processes.

The model demonstrates differentiated predictive capabilities for the two prediction targets. The CCC model achieves a test set  $R^2$  of 0.6342 and an MAE of 2.92%, delivering reliable accuracy for cementation yield prediction. The UCS model obtains a higher test set  $R^2$  of 0.7991 and an MAE of 844.83 kPa. A notable discrepancy exists between a high MAPE of 116.60% and a moderate MdAPE of 38.32%. This phenomenon is scientifically explained by the mathematical amplification of relative errors from a small number of low-strength samples rather than model prediction failure, which well reflects the inherent complexity and high dispersion of strength evolution in MICP solidification systems.

SHAP interpretability analysis verifies the excellent consistency between model prediction and physical mechanisms. For the CCC model, soil particle size with a contribution of 22.47% and urease activity with a contribution of 17.63% serve as the dominant controlling factors. Meanwhile, grouting times with a contribution of 16.56% emerges as a critical variable, highlighting the essential engineering value of uniform construction operation. For the UCS model, soil particle size with a contribution of 29.94% and the interactive feature of urease activity combined with curing time with a contribution of 20.48% dominate the prediction performance. This finding confirms that macroscopic mechanical strength is jointly governed by biochemical reaction kinetics and microstructural mechanical conditions. Furthermore, SHAP dependence results reveal critical nonlinear threshold effects, including the typical inverted U-shaped response of calcium chloride concentration, providing quantitative optimization guidance for practical MICP parameter design.

Comparative validation with baseline models, including Linear Regression, Support Vector Regression, and Random Forest, demonstrates the superiority of the proposed framework. The integration of XGBoost regularization strategies and five-fold cross-validation effectively alleviates overfitting risks under data-limited conditions and ensures stable generalization performance. Overall, this study develops a practical, interpretable, and sample-adaptive tool for rapid and preliminary evaluation. Future work will expand the experimental database, incorporate more construction strategies and geotechnical indices such as relative density, and consider environmental durability factors. These efforts can bridge the gap between laboratory prediction and field application, further improving the universality and practicability of the prediction model.

**Acknowledgement:** Not applicable.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** Conceptualization, Chen Zeng and Pingan Tang; methodology, Pingan Tang and Guang Zhu; software, Pingan Tang and Jun Sun; validation, Pingan Tang, Guang Zhu and Junjun Xu; formal analysis, Pingan Tang; investigation, Pingan Tang; resources, Pingan Tang, Chaojian Yue and Wei He; data curation, Pingan Tang; writing—original draft preparation, Pingan Tang; writing—review and editing, Pingan Tang, Guang Zhu, Chen Zeng, Junjun Xu, Chaojian Yue and Wei He; visualization, Pingan Tang and Jun Sun; supervision, Chen Zeng; project administration, Chen Zeng. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** The data used in this study were compiled from domestic and international published literature, consisting of 77 groups of laboratory MICP soil solidification experiments.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|                |                                             |
|----------------|---------------------------------------------|
| MICP           | Microbially Induced Carbonate Precipitation |
| UCS            | Unconfined Compressive Strength             |
| CCC            | Calcium Carbonate Content                   |
| XGBoost        | Extreme Gradient Boosting                   |
| MAE            | Mean Absolute Error                         |
| RMSE           | Root Mean Square Error                      |
| MAPE           | Mean Absolute Percentage Error              |
| MdAPE          | Median Absolute Percentage Error            |
| R <sup>2</sup> | Coefficient of Determination                |
| BP             | Back Propagation                            |
| Std.           | Standard                                    |
| SVR            | Support Vector Regression                   |
| SHAP           | SHapley Additive exPlanations               |

## References

1. Behzadipour H, Sadrekarimi A. Effects of microbially induced calcite precipitation on static liquefaction behavior of a gold tailings sand. *Biogeotechnics*. 2024;2(4):100097. doi:10.1016/j.bgtech.2024.100097.
2. Almajed A, Lateef MA, Ali Baig Moghal A, Lemboye K. State-of-the-art review of the applicability and challenges of microbial-induced calcite precipitation (MICP) and enzyme-induced calcite precipitation (EICP) techniques for geotechnical and geoenvironmental applications. *Crystals*. 2021;11(4):370. doi:10.3390/cryst11040370.

3. Omoregie AI, Muda K, Ong DEL, Ojuri OO, Bin Bakri MK, Rahman MR, et al. Soil bio-cementation treatment strategies: state-of-the-art review. *Geotech Res.* 2024;11(1):3–27. doi:10.1680/jgere.22.00051.
4. Wang Y, Sun X, Miao L, Wang H, Wu L, Shi W, et al. State-of-the-art review of soil erosion control by MICP and EICP techniques: problems, applications, and prospects. *Sci Total Environ.* 2024;912(3):169016. doi:10.1016/j.scitotenv.2023.169016.
5. Li J, Chen M, Jin H, Sheng N, Tian L. Green and sustainable stabilization and solidification of coal gangue wastes via synergistic high-temperature activation and microbial-induced carbonate precipitation. *J Environ Chem Eng.* 2026;14(2):122041. doi:10.1016/j.jece.2026.122041.
6. Usman H, Khan M, Khan WA. Prediction of unconfined compressive strength of sands treated with microbially-induced calcite precipitation using a hybrid machine learning model. *Transp Infrastruct Geotechnol.* 2025;12(7):228. doi:10.1007/s40515-025-00685-7.
7. Zhang P, Yin ZY, Jin YF. State-of-the-art review of machine learning applications in constitutive modeling of soils. *Arch Comput Method Eng.* 2021;28(5):3661–86. doi:10.1007/s11831-020-09524-z.
8. Talamkhani S. Machine learning-based prediction of unconfined compressive strength of sands treated by microbially-induced calcite precipitation (MICP): a gradient boosting approach and correlation analysis. *Adv Civ Eng.* 2023;2023(1):3692090. doi:10.1155/2023/3692090.
9. 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.* 2021;12(1):469–77. doi:10.1016/j.gsf.2020.03.007.
10. Ly HB, Thai Pham B. Soil unconfined compressive strength prediction using random forest (RF) machine learning model. *Open Constr Build Technol J.* 2020;14(1):278–85. doi:10.2174/1874836802014010278.
11. Qiu Y, Yao S, Qi H, Zhou J, Khandelwal M. Interpretable ensemble learning with lévy flight-enhanced heuristic technique for strength prediction of MICP-treated sands. *Appl Sci.* 2025;15(14):7972. doi:10.3390/app15147972.
12. Tan WK, Hu NY, Ye YC, Wu ML, Huang ZY, Wang XH. Rockburst intensity classification prediction based on four ensemble learning. *Chin J Rock Mech Eng.* 2022;41(Suppl 2):3250–9. (In Chinese). doi:10.13722/j.cnki.jrme.2022.0026.
13. Wang JM, Ye YR, Rao CM, Zhuo RJ, Liu JZ. Prediction on composite interface bonding strength between ceramsite lightweight aggregate concrete and normal concrete based on GBDT algorithm. *J Build Mater.* 2023;26(2):150–5. (In Chinese). doi:10.3969/j.issn.1007-9629.2023.02.006.
14. Li C, Wang L, Li J, Chen Y. Application of multi-algorithm ensemble methods in high-dimensional and small-sample data of geotechnical engineering: a case study of swelling pressure of expansive soils. *J Rock Mech Geotech Eng.* 2024;16(5):1896–917. doi:10.1016/j.jrmge.2023.10.015.
15. Zhuo X, Fan L, Hu D, Zhu H. Multifactor optimization of MICP base on BP model. *J Phys Conf Ser.* 2022;2200(1):012003. doi:10.1088/1742-6596/2200/1/012003.
16. Yang B, Heagy LJ, Morgenroth J, Elmo D. Algorithmic geology: tackling methodological challenges in applying machine learning to rock engineering. *Geosciences.* 2024;14(3):67. doi:10.3390/geosciences14030067.
17. Atangana Njock PG, Shen SL, Zhou A, Lyu HM. Evaluation of soil liquefaction using AI technology incorporating a coupled ENN/t-SNE model. *Soil Dyn Earthq Eng.* 2020;130(3):105988. doi:10.1016/j.soildyn.2019.105988.
18. Sun XH, Miao LC, Tong TZ, Wu LY, Wang HX. Cementation effect evaluation of MICP sand solidification via electrical resistivity. *Chin J Geotech Eng.* 2021;43(3):579–85. doi:10.11779/CJGE202103022.
19. Alnmr A, Hosamo HH, Lyu C, Ray RP, Alzawi MO. Novel insights in soil mechanics: integrating experimental investigation with machine learning for unconfined compression parameter prediction of expansive soil. *Appl Sci.* 2024;14(11):4819. doi:10.3390/app14114819.
20. Tang Y, Cai SX, Cai ZY. A prediction model of dynamic pore water pressure in microbially reinforced silty sand. *Rock Soil Mech.* 2025;46(10):3187–96. (In Chinese). doi:10.16285/j.rsm.2024.1438.
21. Sah AK, Hong YM. Performance comparison of machine learning models for concrete compressive strength prediction. *Materials.* 2024;17(9):2075. doi:10.3390/ma17092075.
22. Wang HL, Pathak B, Yin ZY. Investigation on the microstructure, unconfined compressive strength, and thermal conductivity of compacted CDG soil by MICP treatment during curing. *J Mater Civ Eng.* 2023;35(6):04023131. doi:10.1061/jmcee7.mteng-14901.

23. Xie M, Zhang J, Guan W, Shi G, Ji F, Zhao X, et al. Exploring sustainable mineralization pathways: multi-factor impacts on microbial-induced carbonate precipitation and crystals characteristics. *Processes*. 2025;13(3):711. doi:10.3390/pr13030711.
24. Cui H, Xiao Y, Sun ZC, Wang CG, Liang F, Liu HL. Elastoplastic constitutive model for biocemented sands. *Chin J Geotech Eng*. 2022;44(3):474–82. doi:10.11779/CJGE202203009.
25. Su Y, Cui YJ, Dupla JC, Canou J. Experimental investigation and modelling of the mechanical behaviors of fine/coarse soil mixture. *Transp Geotech*. 2024;49(12):101382. doi:10.1016/j.trgeo.2024.101382.
26. Su Y, Zhang Y, Yao C, Duan J, Zhao H, Zeng P, et al. Investigation on deterioration of mechanical properties of Jiangxi lateritic soil subjected to drying-wetting cycles. *J Rock Mech Geotech Eng*. 2025;857(10):188. doi:10.1016/j.jrmge.2025.10.012.
27. Erdmann N, Strieth D. Influencing factors on ureolytic microbiologically induced calcium carbonate precipitation for biocementation. *World J Microbiol Biotechnol*. 2023;39(2):61. doi:10.1007/s11274-022-03499-8.
28. Sun Y, Liu K, Sun D, Jiang N, Xu W, Wang A. Evaluation of urea hydrolysis for MICP technique applied in recycled aggregate: concentration of urea and bacterial spores. *Constr Build Mater*. 2024;419(11):135366. doi:10.1016/j.conbuildmat.2024.135366.
29. Sreekala AGV, Nair S, Nathan VK. Microbially induced calcium carbonate precipitation using *Lysinibacillus* sp.: a ureolytic bacterium from uttarakhand for soil stabilization. *Curr Microbiol*. 2024;81(11):387. doi:10.1007/s00284-024-03899-z.
30. Erdmann N, Schaefer S, Simon T, Becker A, Bröckel U, Strieth D. MICP treated sand: insights into the impact of particle size on mechanical parameters and pore network after biocementation. *Discover Mater*. 2024;4(1):45. doi:10.1007/s43939-024-00108-3.
31. Zhang Y, Xu X, Liu S, Wang Y, Du J, Jiang N. Bacterial activity and cementation pattern in biostimulated MICP-treated sand-bentonite mixtures. *J Rock Mech Geotech Eng*. 2024;16(12):5121–34. doi:10.1016/j.jrmge.2024.07.005.
32. Kolawole O, Assaad RH, Adams MP, Ngoma MC, Anya A, Assaf G. Coupled experimental assessment and machine learning prediction of mechanical integrity of MICP and cement paste as underground plugging materials. *Biogeotechnics*. 2023;1(2):100020. doi:10.1016/j.bgtech.2023.100020.
33. Song C, Zhao T, Xu L, Huang X. Probabilistic prediction of uniaxial compressive strength for rocks from sparse data using Bayesian Gaussian process regression with Synthetic Minority Oversampling Technique (SMOTE). *Comput Geotech*. 2024;165(2):105850. doi:10.1016/j.compgeo.2023.105850.
34. Bhukya PK, Arnepalli DN. Coupled modelling framework for evaluating the influence of bacterial attachment models on the bio-chemo-hydro-mechanical behaviour during microbially induced calcite precipitation treatment in soils. *Comput Geotech*. 2024;173(8):106561. doi:10.1016/j.compgeo.2024.106561.
35. Li HB, Li X, Wang HJ, Zhang LM, Chen ZY. Tunnel boring machine performance prediction using knowledge-driven transfer learning. *Acta Geotech*. 2025;20(10):4921–39. doi:10.1007/s11440-025-02656-1.
36. Nobari AZ, Jabbarzadeh M, Bertola N, Sengupta A, Alimardani Lavasan A. Sustainable bio-based construction: a review of MICP technology and bio-cementation across scales and materials. *Case Stud Constr Mater*. 2026;24(7):e06100. doi:10.1016/j.cscm.2026.e06100.
37. Gowthaman S, Nakashima K, Kawasaki S. Effect of wetting and drying cycles on the durability of bio-cemented soil of expressway slope. *Int J Environ Sci Technol*. 2022;19(4):2309–22. doi:10.1007/s13762-021-03306-1.
38. Ahenkorah I, Rahman MM, Karim MR, Beecham S. Unconfined compressive strength of MICP and EICP treated sands subjected to cycles of wetting-drying, freezing-thawing and elevated temperature: experimental and EPR modelling. *J Rock Mech Geotech Eng*. 2023;15(5):1226–47. doi:10.1016/j.jrmge.2022.08.007.
39. Li D, Chen S, Gao X. Numerical analysis of MICP treated sand based on bio-chemo-hydro model. *Front Bioeng Biotechnol*. 2024;12:1380213. doi:10.3389/fbioe.2024.1380213.