Hindawi
Advances in Civil Engineering
Volume 2023, Article ID 3692090, 20 pages
https://doi.org/10.1155/2023/3692090

Research Article

Machine Learning-Based Prediction of Unconfined Compressive Strength of Sands Treated by Microbially-Induced Calcite Precipitation (MICP): A Gradient Boosting Approach and Correlation Analysis

Saeed Talamkhani

Department of Civil Engineering, Imam Khomeini International University, Qazvin, Iran

Correspondence should be addressed to Saeed Talamkhani; s.talamkhani@edu.ikiu.ac.ir

Received 14 March 2023; Revised 15 May 2023; Accepted 17 May 2023; Published 10 June 2023

Academic Editor: Tao Wen

The current study applies a soft-computing approach based on the gradient boosting method to predict the unconfined compressive strength (UCS) of sands treated with microbially-induced calcite precipitation (MICP). A 10-fold cross-validation method and hyperparameter tuning are performed to find the optimal architecture of the gradient boosting algorithm. A total of 402 data of unconfined compression tests performed on biocemented sands are utilized in this study. The dataset includes eight input parameters: median sand particle size, uniformity coefficient of sand, initial void ratio, calcium chloride concentration, urea concentration, urease activity, optical density of bacteria, and calcite content. The finding demonstrates that the gradient boosting method outperformed five commonly used machine learning algorithms (artificial neural networks, random forests, k-nearest neighbors, support vector regression, and decision trees) in predicting the UCS of biocemented sands. Using the gradient boosting, the predicted UCS has a strong correlation with the actual values ($R^2 = 0.95$). Moreover, a series of correlation and feature importance analyses are carried out over the dataset. The relationships between unconfined compressive strength, calcite content, and initial void ratio are discussed within the article. Furthermore, some guidelines are provided for assessing the effect of environmental factors on the UCS of biocemented sands. For further study, the limitations of this study regarding the insufficiency of data for correlation and environmental modification are addressed.

1. Introduction

The inexorable growth of the global population has led to utilizing every available piece of land for construction. Since not all soils have the compressive strength to support structural loads, loose sites are frequently stabilized to facilitate construction. Several approaches are being used to enhance the compressive behavior of sand, each of which has advantages and disadvantages in terms of economics, environment, and practicality. In traditional soil improvement techniques, adhesive materials such as cement, lime, or other chemicals are usually added to soil to improve its strength. Although these approaches are effective in enhancing the strength of soils, their negative environmental impacts far outweigh their mechanical benefits [1–3]. In order to minimize the environmental impact associated with traditional soil improvement methods, some ecofriendly approaches have been developed.

Microbially-induced calcite precipitation (MICP) is a sustainable, cost-effective, and novel approach for enhancing the compressive strength of sand [4, 5]. The biocementation process enhances the compressive strength of sand with a biological activity that produces calcite minerals within soil structure; thus, no cement or other chemical binders are included in the stabilization process, leading to an ecofriendly ground improvement approach. In most laboratory and in situ explorations, it is frequent to perform the unconfined compression test to acquire the compressive strength of cemented sands [6, 7]. The unconfined compressive strength (UCS) of biocemented sand depends on several factors, such as soil properties, details of the MICP process, and environmental conditions, so the UCS of treated sand covers a wide range from 0.15 to 34 MPa [4]. This wide range of compressive strength for soil can significantly influence the design and function of the overlaying structures. A reasonable prediction of the compressive strength of biocemented sands can enhance the reliability of the predesign of overlaying structures and clarify the suitability of the MICP method for a problematic site.

For predicting the unconfined compressive strength of biocemented sand, Wang and Yin [8] developed a multi-expression programming algorithm combined with the Monte–Carlo method (MEP-MC) that relies on an evolutionary algorithm for developing mathematical expressions. A database consisting of 351 data driven from previous studies was employed for developing MEP-MC algorithms. Several MEP-MC models were developed for predicting the UCS of biocemented sand which was found to be reliable and accurate. However, considering the superiority of advanced soft-computing techniques, such as machine learning techniques, the authors proposed that implementing these novel techniques can produce more reliable and straightforward models and algorithms for predicting the UCS of biocemented sand.

In the past few years, advances in analytical and computational studies have led to the development of soft-computing techniques, which are derived from mathematical and statistical algorithms. Machine learning is one of the novel techniques which can be used to identify linear and nonlinear relationships between variables. Machine learning is being utilized in geotechnical investigations as a means of solving problems, predicting disasters, or estimating soil characteristics [9–13]. Currently, several effective algorithms are commonly used for geotechnical issues, including artificial neural networks (ANNs), support vector machines (SVMs), k-nearest neighbors (KNNs), decision trees (DTs), and many others. In data-driven modeling, it is most common to construct only one strong prediction model. In an alternative approach, a group of models could be developed to address a particular learning objective. The ensemble learning method is a general application of several weak learners in which predictions from several models are combined to improve predictive performance. Consequently, combining more simple learners will result in a higher level of predictive accuracy than only an individual model. Furthermore, since ensemble machines contain several learners, the implementation of ensemble methods is highly effective for both linear and nonlinear data [14, 15].

Generally, ensemble methods can be classified into two groups based on their structure: parallel and sequential. Parallel algorithms run several learners simultaneously and then calculate the final prediction from all independent learners. Among the parallel ensemble methods, random forest is being used more frequently in engineering and geotechnical problems [16–18]. On the other hand, a sequential process (also known as boosting) builds base estimators sequentially and attempts to reduce the bias of the combined estimator at each iteration. Gradient boosting (GB) is an ensemble algorithm constructing additive regression models by sequentially fitting a weak learner at each iteration to current pseudo-residuals [19]. Since soil problems struggle with nonlinear behavior, the gradient boosting approach would be well-suited for solving geotechnical issues. Numerous studies have found gradient boosting to be a robust approach for predicting geotechnical problems, such as shear strength [20–22], slope stability [23–25], settlement [26–28], liquefaction [29–31], and other geotechnical concerns.

To develop functional and reliable models for predicting the UCS of sand treated with MICP, this study was carried out to predict the unconfined compressive strength of biocemented sands using machine learning techniques. Given that the gradient boosting is capable of analyzing datasets with nonlinear behavior, this method was employed as the main algorithm. Also, five frequently used and straight forward algorithms were utilized to compare the performance of the gradient boosting method. In this paper, the MICP method mechanism is delineated in Section 2. Afterward, gradient boosting fundamentals are discussed in Section 3. Section 4 describes the dataset and provides a correlation analysis of variables. This section also presents procedures of the k-fold cross-validation and hyperparameter tuning. In Section 5, the result and discussion are outlined. To modify the predicted UCS with environmental factors, section 6 provides guidelines for applying the effect of temperature and pH on the final UCS. Lastly, conclusions derived from this study and potential future works are presented.

2. MICP

Microbially-induced calcite precipitation (MICP) is an interdisciplinary approach for enhancing mechanical behavior of soil by microbial activity. The microorganisms produce calcite crystals ($CaCO_3$) within the soil pores that bind soil grains together and improve stiffness and strength. There are three steps involved in the MICP process: (1) bacteria cultivation; (2) treatment; (3) curing. Following the article, details regarding each step are presented separately. A schematic diagram of the MICP process is shown in Figure 1, depicting each step along with factors that contribute to the final strength of the treated sand.

2.1. Bacteria Cultivation. The primary function of the microorganisms in MICP is to break down urea and act as a catalyst for the formation of carbonate crystals between the sand grains. In MICP, *Sporosarcina pasteurii* (also known as Bacillus pasteurii) is most commonly used for ureolysis due to its high urease activity [33, 34]. *S. pasteurii*, which grows in alkaline media, requires urea and ammonium to grow: urea provides nitrogen and carbon to the bacteria, and ammonium regulates the pH and allows substrates to pass through the cell membrane [35–37]. Therefore, *S. pasteurii* is cultivated under an aerobic batch containing nutrients that help to grow the bacteria and increase urease activity. Most previous studies used ammonium-yeast extract [34–37] or trypticase soy agar [38, 39] media for S. *pasteurii* cultivation.

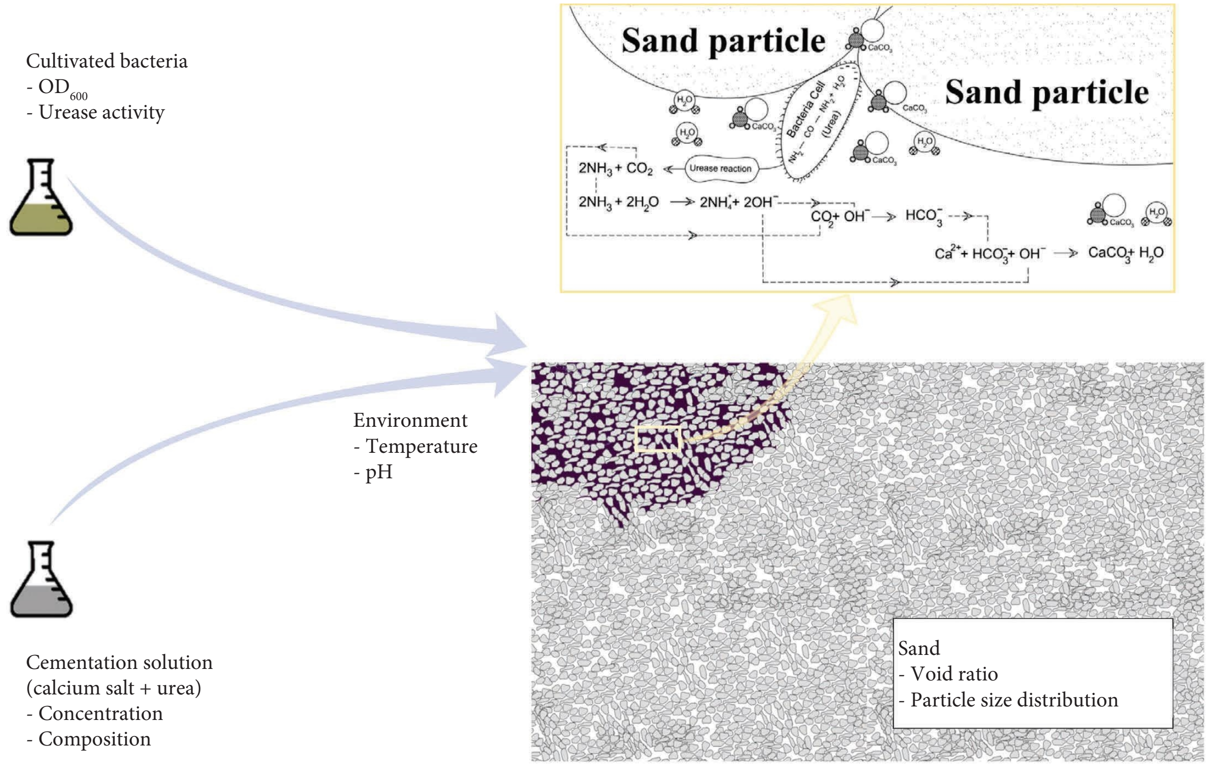


Figure 1: Schematic view of the MICP treatment (microscale simulation was depicted by Behzadipour et al. [32]).

The main characteristics of harvested bacteria measured before introducing bacteria to the soil are the optical density of biomass at 600 nm, OD_{600}, and the urease activity, *UA*. The optical density of biomass is correlated with bacterial concentration as well as bacterial size in a sample [38]. Urease activity is an indicator of bacteria's capability to hydrolyze urea, which is greatly influenced by environmental factors such as cultivation and storage conditions [34]. Urease activity is measured in units of $\mathrm{mM{\cdot}h^{-1}}$ or $\mathrm{U{\cdot}mL^{-1}}$, which can be converted as follows:

$$\begin{gathered} 1\,U = 1\,\mu\text{mol of urea hydrolyzed per minute}, \\ 1\,U.mL^{-1} = 1\,\mu\text{mol}.\text{min}^{-1}.mL^{-1} = 1\,mM.\text{min}^{-1} = 60\,mM.h^{-1}. \end{gathered} \tag{1}$$

Previous studies have shown that the performance of MICP and the final strength of treated soil are influenced by bacterial density and total activity [39–43]. A study by Hammad et al. [43] assessed the activity of *S. pasteurii* in an agar-urea medium and found that higher urease activity leads to faster crystallization of $CaCO_3$. Cheng et al. [41] evaluated the performance of biocementation with three different urease activities (5, 10, and 50 $\mathrm{U{\cdot}mL^{-1}}$) and found that specimens with a low level of urease activity exhibited a greater UCS for the same amount of $CaCO_3$ content, which was due to differences in nucleation sites affecting precipitation patterns. Zhao et al. [40] revealed that sands treated with high values of OD_{600} and urease activity could maintain more significant unconfined pressure. As bacteria concentrations increase, more $CaCO_3$ is precipitated. Based on the comparison of precipitation patterns of three bacteria with different OD_{600} (0.2, 1, and 3), Wang et al. [42] concluded that the density of bacteria greatly influences the stability of $CaCO_3$ crystals.

2.2. Treatment. Following the cultivation of bacteria with the desired density and activity, the bacteria and the cementation solutions are introduced to the soil. The addition of bacteria and cementation solutions to sandy soils can be carried out using three methods: injection, surface percolation, and premixing. Injections are more commonly used than either of the two other techniques, which are less efficient and practical due to certain limitations. In surface percolating, the main issue is related to restricted penetration depth. The treatment depth of surface percolation is limited to 2 m for coarse granular material and 1 m for fine sand [44]. Furthermore, the premixing method involves disturbing the soil mass in order to mix it with the solution.

Due to the intense mixing of cementing ingredients with soil, pseudo-stress emerges in the soil sample during this process [4]. On the other hand, injection is the most common method for MICP, which improves soils without disrupting the soil structure.

The injection of bacteria and cementation solutions takes place sequentially in several batches: first, bacteria suspensions are commonly injected into soil masses, followed by the injection of cementation solution. However, few drawbacks relating to the homogeneity of $CaCO_3$ within the soil were seen when solutions were injected into the soil [34, 45]. The uneven distribution of $CaCO_3$ is ascribed to the linear reduction of microbial concentration along the injection path [46]. It is possible to resolve the unequal distribution of $CaCO_3$ content by slowing down the injection rate of the bacterial suspension or considering a break between the injection of bacteria and cementation solutions [47, 48].

Cementation solutions are composed of urea and calcium salt, accompanied by injecting a small number of nutrients or ammonium chloride to maintain microorganism activity [49–51]. Calcium salt solution supplies calcium molecules for $CaCO_3$ crystallization, and its composition can affect the cemented sand formation and calcium content [52–54]. Among the calcium compositions, calcium chloride ($CaCl_2$) has been used most commonly for MICP [34, 39, 45, 49, 51, 55], which is due to the ability to produce a greater amount of $CaCO_3$ [54].

Moreover, the concentration of cementation solution influences the performance of MICP and the final strength of cemented soil. Al-Qabany and Soga [49] observed uniform distribution of $CaCO_3$ in sands treated with low-concentration solutions (0.25 and 0.5 M). A number of studies have also indicated that sandy soils treated with 1 M of cementation solution tolerate a lower UCS than soils treated with lower concentrations [39, 48, 56]. Aside from the concentration of each solution, the ratio between the concentration of urea and calcium salts can also influence the performance of the MICP. As the urea content exceeds the calcium salt content, the bacteria consume more urea and become more active; as a result, the calcium content and the shear strength of the biocemented sands will increase [57, 58]. However, Mahawish et al. [58] evaluated the behavior of soil treated with equimolar (similar molarity of urea and calcium chloride) and nonequimolar (the urea content was two times the calcium chloride content) cementation solutions and found that nonequimolar solutions produce more uniform distributions of $CaCO_3$ than that of equimolar solutions.

2.3. Curing. The involvement of bacteria and cementation solutions within the soil matrix triggered reactions that resulted in the formation of calcite crystals among sand pore space. The microscale simulation of chemical reactions is exhibited in Figure 1. The chemical reactions occur in the following order:

$$CO\left(NH_2\right)_2 + H_2O \xrightarrow{\text{Urease}} 2NH_3 + CO_2 \tag{2}$$

$$2NH_3 + 2H_2O \longrightarrow 2NH_4^+ + 2OH^- \tag{3}$$

$$CO_2 + OH^- \longrightarrow HCO_3^- \tag{4}$$

$$Ca^+ + HCO_3^- + OH^- \longrightarrow CaCO_3 + H_2O \tag{5}$$

As can be seen, the chemical reaction starts with the decomposition of urea ($CO(NH_2)_2 + H_2O$) by bacterial microorganisms, followed by producing calcite ($CaCO_3$) crystals and ammonium (NH_4^-) ions. Chemical reactions in MICP are influenced by environmental factors, such as the temperature and pH of the sand, which influence the $CaCO_3$ content and mechanical characteristics of the treated sands.

The temperature of the curing media can significantly affect the MICP performance. Increasing the setting temperature up to 50°C raises urease activity, leading to precipitation of a more considerable amount of $CaCO_3$ in the MICP process. However, sands treated at room temperature (20–25°C) show greater strength than those cured at 50°C, which indicates that $CaCO_3$ depositions produced at 50°C are less effective at strengthening biocemented sands than those produced at room temperature [4, 41, 58]. At a similar $CaCO_3$ content, cemented sands treated at room temperature show higher UCS than those treated at a colder or warmer temperature. Cheng et al. [41] ascribed this discrimination to the incompetency of $CaCO_3$ crystals to fill the gap between the sand grains, which stem from the faster nucleation rate of $CaCO_3$ precipitation at 50°C and the lower nucleation rate at 4°C [59, 60]. Mahawish et al. [58] attributed the ineffective precipitation to the formation of loose $CaCO_3$ crystals at elevated temperatures.

The initial pH level of the MICP environment has an impact on the activity of the microorganisms that affect the precipitation of $CaCO_3$ and the strength of treated sand [47, 61]. Soil media with high acidity and alkalinity are found to be in inhospitable environments for microorganisms to form $CaCO_3$ crystals [62, 63]. Liu et al. [62] observed no efficient $CaCO_3$ crystal among sand grains contact when the treated sand was immersed in an acidic medium with a pH value of 3.5. It was also reported that the $CaCO_3$ depositions were consumed through reacting proton ions (H^+) in the acidic solution. Overall, the optimum pH level for the MICP process was found to be around 7 or a neutral environment [63–65].

3. Gradient Boosting

Gradient boosting (GB) is a supervised machine learning algorithm that combines outputs of several weak learners sequentially to yield a robust model. A schematic representation of the GB mechanism is shown in Figure 2.

Boosting involves sequentially applying a weak learner, $f(x)$, to repeatedly modified versions of the data, resulting in a sequence of weak learners, $f_m(x)$, $m = 1, 2, \ldots, M$. The final

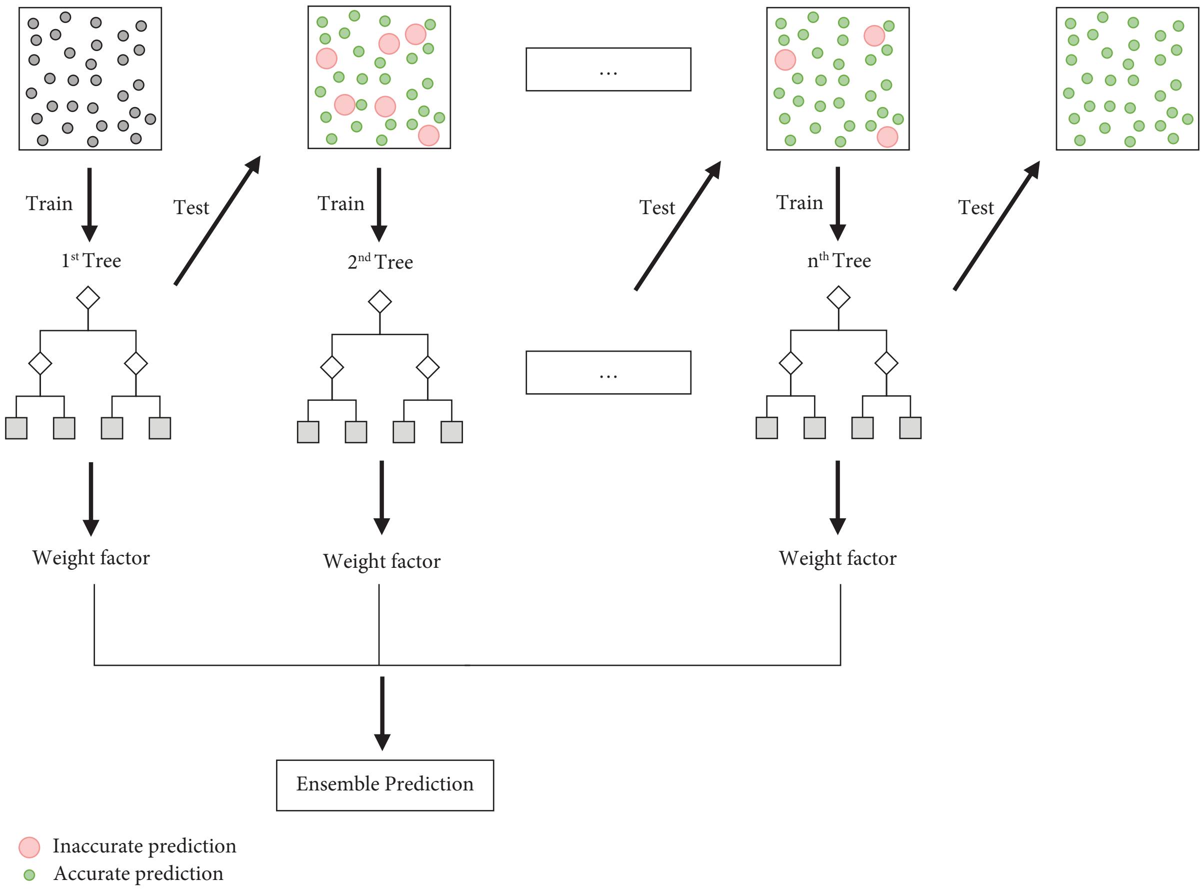


Figure 2: Schematic diagram of the gradient boosting method.

prediction is obtained by multiplying the predictions of all learners by a weight (α_m) [66]:

$$f(x) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m f_m(x)\right). \tag{6}$$

In order to fit these models, $f(x)$, the loss function, $L(y, f(x))$, is minimized over the training data:

$$\gamma(x) = \arg\min L(y, f(x)), \tag{7}$$

where x denotes input variables, and y is the target variable. The accuracy of the final prediction depends on the values of the weight factors obtained through the boosting algorithms. The weight of each learner is determined based on its accuracy, which is calculated by a loss function: the more precision is attained by a learner, the lower the weight factor is assigned. Therefore, by assigning unequal weight to the training set at each iteration, the learner knows how to focus on erroneous data at the next iteration.

In gradient boosting models, decision trees are harnessed as weak learners, which are relatively fast to construct and capable of performing robust predictions [67]. Decision trees split the training set into disjoint regions R_j, $j = 1, 2, \ldots, J$, according to the terminal nodes then assign a constant γ_j to each region, so the predictive rule, based on the inputs x, can be defined as follows:

$$x \in R_j \longrightarrow f(x) = \gamma_j. \tag{8}$$

Within the gradient boosting procedure, additive decision trees sequentially are constructed based; then, at each iteration with regard to each training data, the pseudo-residuals (gradient of the loss function) are minimized. The gradient boosting algorithm is written in Algorithm 1.

In the first step, the model initializes with a single terminal node tree. Then, with boosting approach with $m = 1, 2, \ldots, M$, the best regression tree is fitted in 4 steps. First, the component of the negative gradient (pseudo-residual), r_{im}, for $i = 1, 2, \ldots, N$, is computed. Then, a regression tree partitions the training data into L-disjoint regions, $\{R_{jm}\}_1^L$ and assigns distinct constant values at each node. After that, the minimum value of the loss function within different regions is located. Consequently, the current approximation at each region is separately updated based on the previous iteration. The final GB model is obtained from the sum of all trees fitted at each iteration multiplied by its coefficient. In other words, the model constructed at the last iteration is equivalent to the final model, which involves all trees fitted at the previous iteration multiplied into the corresponding coefficient. The shrinkage parameter, ν, at Algorithm1 represents the learning rate of the additive procedure. An operation with a low shrinkage will have a higher degree of precision;

```
The overview of the GB algorithm for regression is summarized in the following order [67]:
(1) Initialize: $f_0(x) = \arg\min_\gamma \sum_{i-1}^{N} L(y_i, \gamma)$
(2) For $m = 1$ to $M$, do:
(a)     $r_{im} = -[\partial L(y_i, f(x_i))/\partial f(x_i)]_{f=f_{m-1}}, i = 1, 2, \ldots, N$
(b)     $\{R_{jm}\}_1^L = L -$ terminal node tree $(\{r_{im}, x_i\}_1^N)$
(c) $\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$
(d) $f_m(x) = f_{m-1}(x) + v.\gamma_{jm} I(\mathrm{x} \in R_{jm})$
(3) Output $\widehat{f}(x) = f_M(x)$
```

Algorithm 1: Gradient boosting.

accordingly, a substantial amount of computation time and CPU capacity are required [68].

3.1. GB Design. Gradient boosting algorithm can mainly be divided into three parts: (1) loss function, (2) additive model, and (3) decision tree. In this section, the configuration of each component is presented separately.

3.1.1. Loss Function. Within the boosting procedure, trees are fitted by a loss function over the training set. For regression problems with continuous data, three well-known loss functions (loss) are being harnessed: absolute error (L_{lad}), squared error (L_{ls}), and Huber ($L_{\mathrm{Huber},\delta}$). The first two functions are more renowned in the computation of regression algorithms. The equations of squared error and absolute error loss functions for target variable, y, and function estimate, f, can be formulated as follows:

$$\begin{aligned} L_{\mathrm{lad}} &= |y - f|, \\ L_{\mathrm{ls}} &= \frac{1}{2}(y - f)^2. \end{aligned} \tag{9}$$

The squared error is more convenient than the other loss functions because its derivative is equal to the residual of the current model at each iteration ($r_{\mathrm{im}} = y_i - f(x_i)$). Thus, for the squared-error loss function, the current residual is added to the expansion at each iteration, which facilitates the computation of the gradient boosting algorithm [19, 67, 69].

An alternative for the squared-error loss function is the Huber loss function [70], which is a combination of squared error and absolute error loss functions:

$$L_{\mathrm{Huber},\delta} = \begin{cases} \frac{1}{2}(y - f)^2 |y - f| \leq \delta, \\ \delta\left(|y - f| - \frac{\delta}{2}\right)|y - f| > \delta, \end{cases} \tag{10}$$

where δ denotes the threshold at which the loss function transitions from square error to absolute error. The optimum value of δ depends on the distribution of $(y-f)$. It is suggested to choose the α-quantile of the distribution of $|y-f|$ equivalent to the value of δ. In this case, $(1-\alpha)$ corresponds to the breakdown point in the procedure. The breakdown point refers to the fraction of observations capable of being arbitrarily modified without degrading the quality of the results [19].

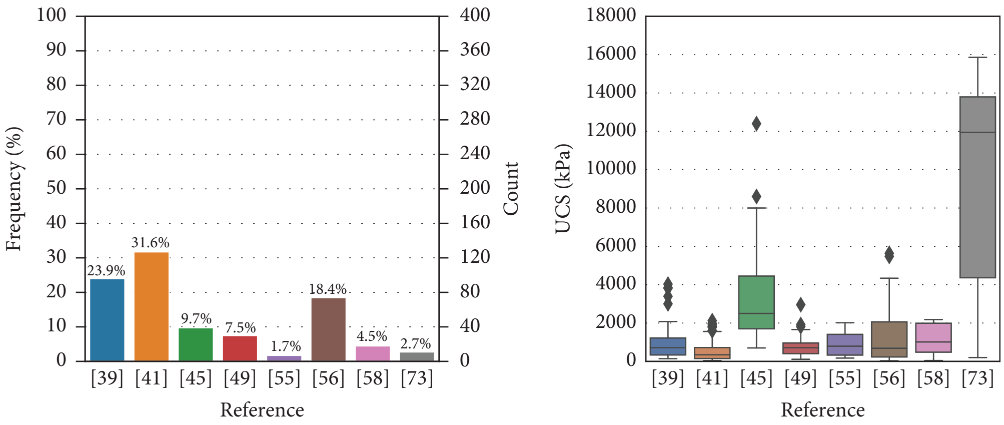
3.1.2. Additive Model. As mentioned before, the GB algorithm fits decision trees sequentially, and the accuracy of the model increases after each iteration. The number of iterations (*n_estimators*) can influence the final model and its accuracy. There are three common methods for determining the optimal number of iterations in the gradient boosting method: an independent test set, out-of-bag estimation, and k-fold cross-validation. As Ridgeway [71] demonstrated 5 or 10-fold cross-validation is more effective than the other approaches, although it may require more computing time.

Moreover, the shrinkage parameter (*Learning_rate*) significantly has an impact on the performance of the GB algorithm. Since shrinkage represents the learning rate of boosting procedure, its lower values result in models with better predictive performance. However, models with lower shrinkage demand far more storage and CPU time. A lower amount of learning rate requires more significant iterations to achieve the optimal algorithm [71].

3.1.3. Decision Tree. The configuration of decision trees fitted within the gradient boosting procedure can affect the final accuracy. For a decision, properties such as the maximum depth that limits the growth of trees (*max_depth*), the minimum number of samples required to split an internal node (*min_samples_split*), the minimum number of samples needed to be at an internal or external node (*min_samples_leaf*), and the number of features to consider when looking for the best split (*max_features*) determine the structure of the final decision tree [72].

4. Methodology

4.1. Dataset. The dataset consists of 402 unconfined compression test results conducted on sands treated with MICP, which were reported in previous studies [39, 41, 45, 49, 55, 56, 58, 73]. This literature-based database includes all research were conducted on biocemented sands that properly reported test procedures and results that could be relied upon. Figure 3 shows the contribution of references along with their UCS distribution. The barplot in Figure 3(a) demonstrates the frequency of each reference, and its portion is plotted above the column. As can be seen, the distribution of data is not equal among references; for



(a)

(b)

Figure 3: Statistical details of references: (a) frequency; (b) boxplot distribution of UCS.

instance, Cheng et al. [41], the most populated reference, constitutes 31.6% of the entire dataset, while another research by Cheng et al. [55] has only 1.7% contribution in the dataset. Furthermore, the box plot in Figure 3(b) illustrates UCS distribution for each individual referenced study. The bottom, middle, and top of each box are the first quartile, median, and third quartile of the UCS population, respectively. The lines extending from the top and bottom of each box indicate the minimum and maximum UCS. The outlier points for each reference are also shown in Figure 3(b). Similarly, the UCS distributions are not identical across all studies; however, it can be seen that most of the studies concentrate on UCS below 4000 kPa.

Eight parameters are considered as inputs in the dataset: median sand particle size, $(D)_{50}$, uniformity coefficient of sand, C_u, initial void ratio of sand, e_0, calcium chloride concentration, M_{ca}, urea concentration, M_u, optical density of bacteria, OD_{600}, urease activity of bacteria, UA, and calcite content, F_{CaCO3}. Apart from input parameters, some other variables are almost analogous throughout the dataset, so they are not included in the dataset. The source of calcium was calcium chloride in all studies. The treated sands were initially neutral (pH = 7) and were cured at room temperature (20–30°C).

Table 1 presents statistical information of the dataset, including mean, standard deviation (*std*), minimum, maximum, and quartiles for each variable. In addition, the distribution of unconfined compressive strength and each input parameter is exhibited in Figure 4. The description of variables can be summarized as follows:

(i) The sands are classified as fine to medium sands with median grain sizes ranging from 0.14 to 1.60 mm (Figure 4(a)); however, most of them can be categorized as fine-grained sands ($D_{50} < 0.425$ mm) [74]. Also, the majority of the sands have uniform particle size distribution with a coefficient of uniformity ranging from 1 to 2 (Figure 4(b)).

(ii) The initial void ratio varies from dense to loose sands ($0.43 < e_0 < 0.86$), while the majority of treated soils have an initial void ratio between 0.55 and 0.75 (Figure 4(c)).

(iii) The values of optical density and activity of bacteria are distributed uniformly throughout the dataset: bacteria have OD_{600} values between 0.1 and 4.46 and UA values between 1.7 and 50 $\mathrm{U{\cdot}ml^{-1}}$ (Figures 4(d) and 4(e)).

(iv) The concentration of urea and calcium chloride solutions were distributed from 0.1 to 2, mostly compounded with 1 $\mathrm{mol{\cdot}L^{-1}}$ (Figures 4(f) and 4(g)). Furthermore, the cementation solutions were mixed in both equimolar and nonequimolar proportions.

(v) The $CaCO_3$ content is distributed chiefly in values lower than 10%; however, almost 14% of the data have F_{CaCO3} exceeding 10% (Figure 4(h)).

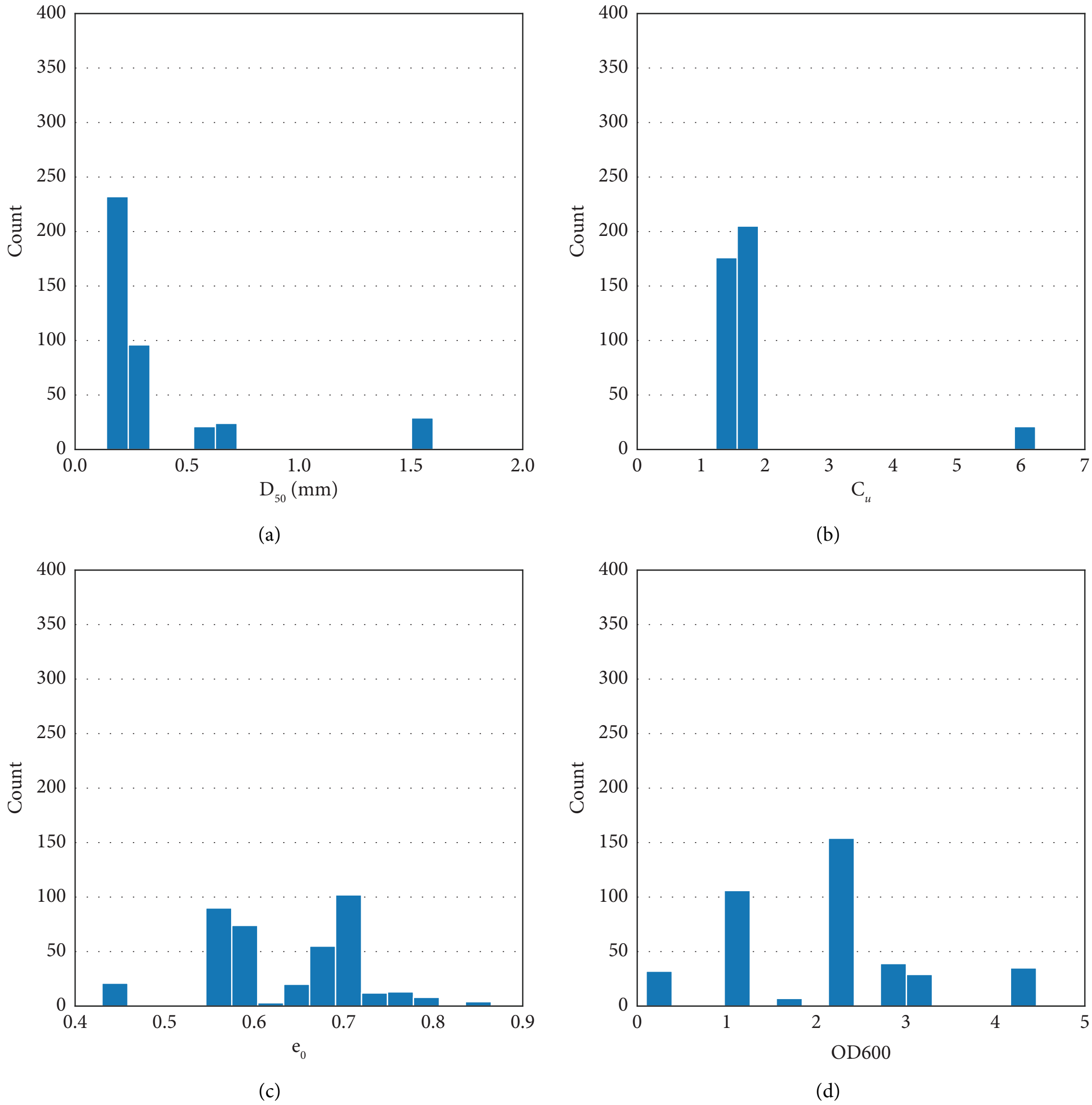
(vi) The UCS of biocemented sands fluctuates between 31 and 16000 kPa. The majority of the samples have UCS lower than 2000 kPa (Figure 4(i)).

4.2. Correlation of Variables. Correlation analysis can efficiently reveal the relationship between variables in a dataset. In this study, the Pearson correlation coefficient approach is used to analyze the relationship between variables [75]. The Pearson correlation method determines the degree of the linear relationship between two variables. The Pearson correlation coefficient, $r_{p,}$ ranges from −1 to 1. The higher value of r_p represents the strong correlation between the two variables. In Figure 5, the heatmap of the Pearson correlation coefficients matrix of all features is depicted. It is evident that the UCS of biocemented sands strongly correlates with calcite content. By contrast, UCS is almost independent of the uniformity coefficient of sand.

To explore the relationship between UCS on F_{CaCO3}, Figure 6 displays the distribution of UCS with various F_{CaCO3}. A linear regression line with a positive slope is also

Table 1: Statistical description of dataset.

Parameter	Mean	std	Min	25%	50%	75%	Max
D_{50} (mm)	0.35	0.37	0.14	0.16	0.23	0.25	1.60
C_u	1.72	1.08	1.23	1.23	1.64	1.65	6.23
e_0	0.64	0.09	0.43	0.57	0.65	0.72	0.86
OD_{600}	2.09	1.12	0.10	1.00	2.25	2.88	4.46
UA (U/mL)	13.72	10.13	1.70	8.33	10.00	18.33	50.00
M_u (mol/L)	0.77	0.34	0.10	0.50	1.00	1.00	2.00
M_{Ca} (mol/L)	0.73	0.34	0.10	0.50	1.00	1.00	1.50
F_{CaCO3} (%)	6.54	5.32	1.09	3.22	4.86	7.16	27.30
UCS (kPa)	1328	2101	31	246	674	1599	15859



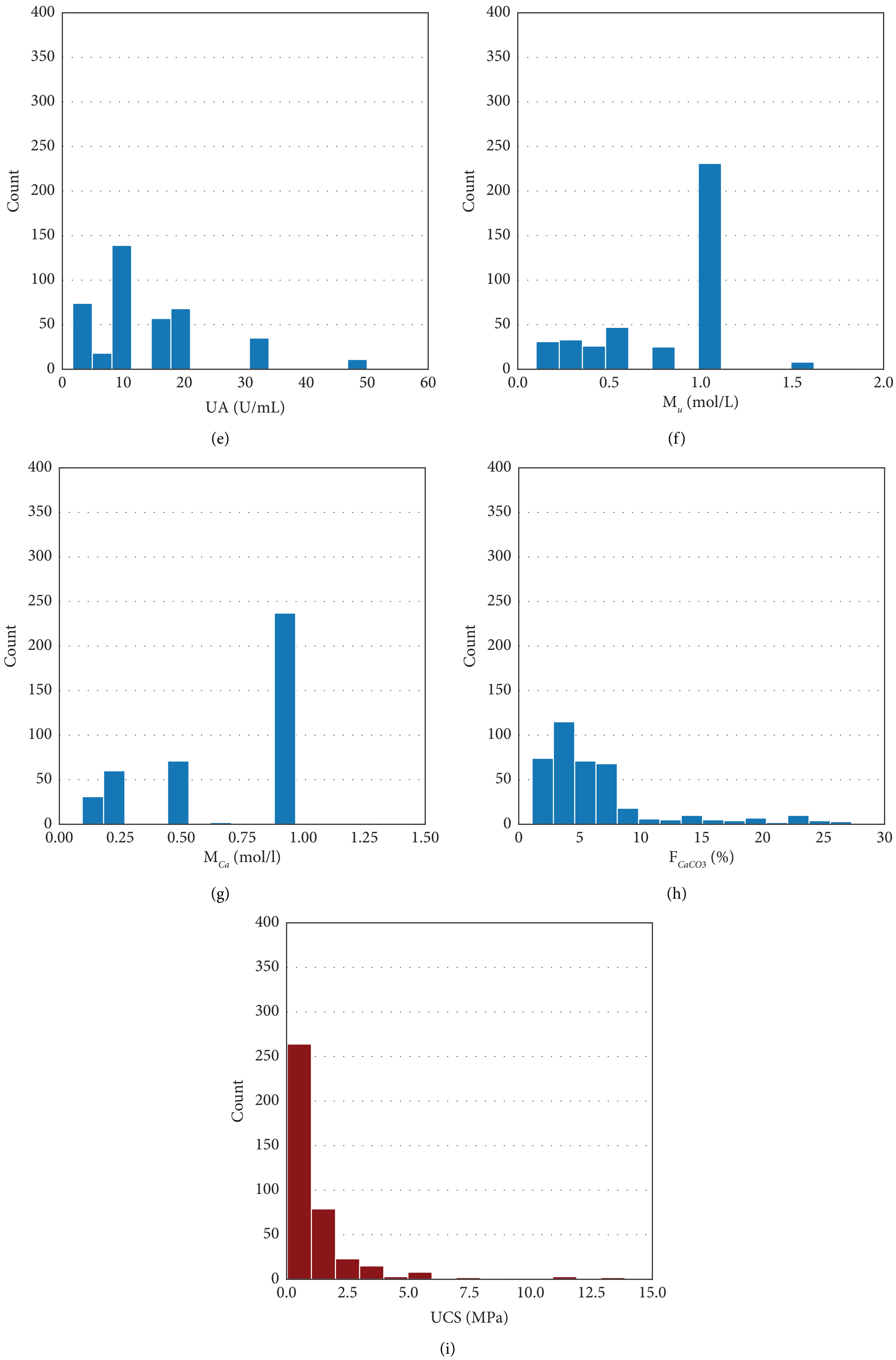
(a)

(b)

(c)

(d)

Figure 4: Continued.



(e)

(f)

(g)

(h)

(i)

Figure 4: Frequency of variables: (a) median particle size; (b) uniformity coefficient; (c) initial void ratio; (d) optical density of bacteria; (e) urease enzyme activity; (f) urea concentration; (g) calcium source concentration; (h) calcium carbonate content; (i) unconfined compressive strength.

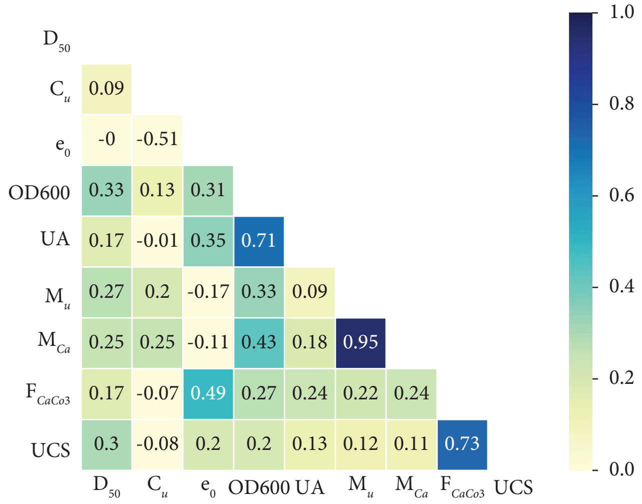


FIGURE 5: Correlation heatmap of input and output variables.

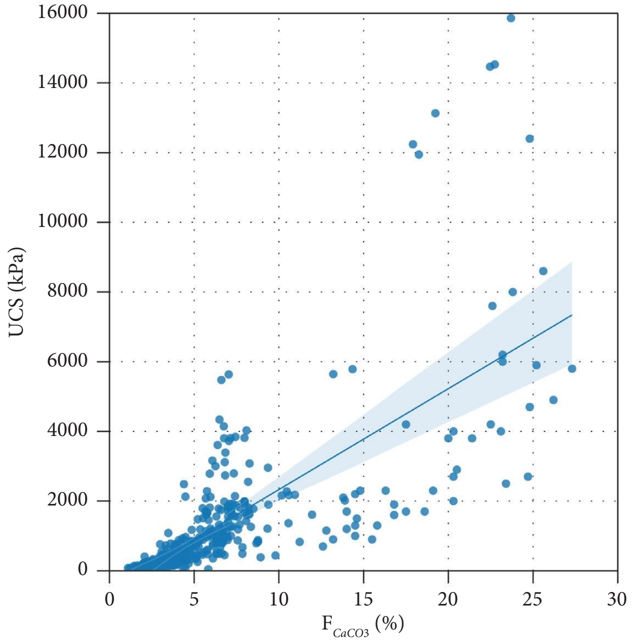


FIGURE 6: Distribution of unconfined compressive strength (UCS) with various $CaCO_3$ content (F_{CaCO3}) within dataset.

plotted in Figure 6, which establishes the direct relationship between UCS and F_{CaCO3}. In other words, for sands cemented with similar test properties, those samples with large amounts of $CaCO_3$ content would sustain higher compression. This strength enhancement mainly stems from the role of $CaCo_3$ crystals in the sand pores that binds sand grains together.

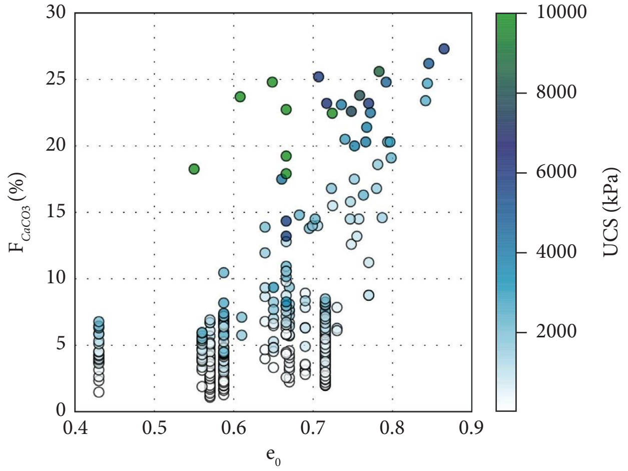


FIGURE 7: Distribution of $CaCO_3$ content (F_{CaCO3}) versus initial void ratio (e_0) colored based on UCS.

The initial void ratio of soil is a fundamental parameter for defining the density of soil. According to Figure 6, the void ratio shows the strongest correlation with $CaCO_3$ content within the dataset. In order to explore the relationship between void ratio, $CaCO_3$ content, and UCS, the gradient-colored scatterplot of e_0 and F_{CaCO3} is shown in Figure 7. The color bar displayed on the right side of Figure 7 gives the values of UCS for each data. The color bar represents the UCS ranges between 0 and 10 MPa in Figure 7, so data with UCS higher than 10 MPa are colored the data with a UCS of 10 MPa. In spite of the non-normal distribution of e_0, a correlation between e_0 and F_{CaCO3} can be derived from Figure 7: the F_{CaCO3} reaches to higher value for sand with greater e_0. In other words, sands with more void space have the potential to produce more amounts of calcite crystals among the sand particles. Furthermore, the color-mapped data with UCS demonstrate that sands with F_{CaCO3} higher than 10% mostly have e_0 between 0.6 and 0.9, and these treated samples have UCS higher than 2 MPa. Therefore, enhancing the compressive strength through MICP is more efficient in sands with more pore space ($0.6 < e_0 < 0.9$) than in dense ones. High-strength treated sands (UCS >10 MPa) are mainly found in sands with void ratios ranging from 0.6 to 0.8 in Figure 7.

4.3. *K-fold Cross-Validation.*

Validation of models was carried out through a *k*-fold cross-validation approach, which produces reliable models obtained from *k* times validation. In *k*-fold cross-validation, the dataset is divided into two sets: a training set and a test set. The test set is held out for the final evaluation of the model. The training set is divided into *k* subsamples with similar sizes. Then, a model is fitted based on the $(k-1)$ folds of the training data, and the remaining fold validates the constructed model. This procedure is repeated for *k* time, and each fold is harnessed as a cross-validation set for one time. In *k*-fold cross-validation, the evaluation of the model is obtained from the average of all models. This study uses 10-fold cross-validation by holding out 20% of the dataset as the test set for model development. The test set is selected randomly over the whole dataset. Figure 8 illustrates the schematic procedure of the 10-fold cross-validation used in this study.

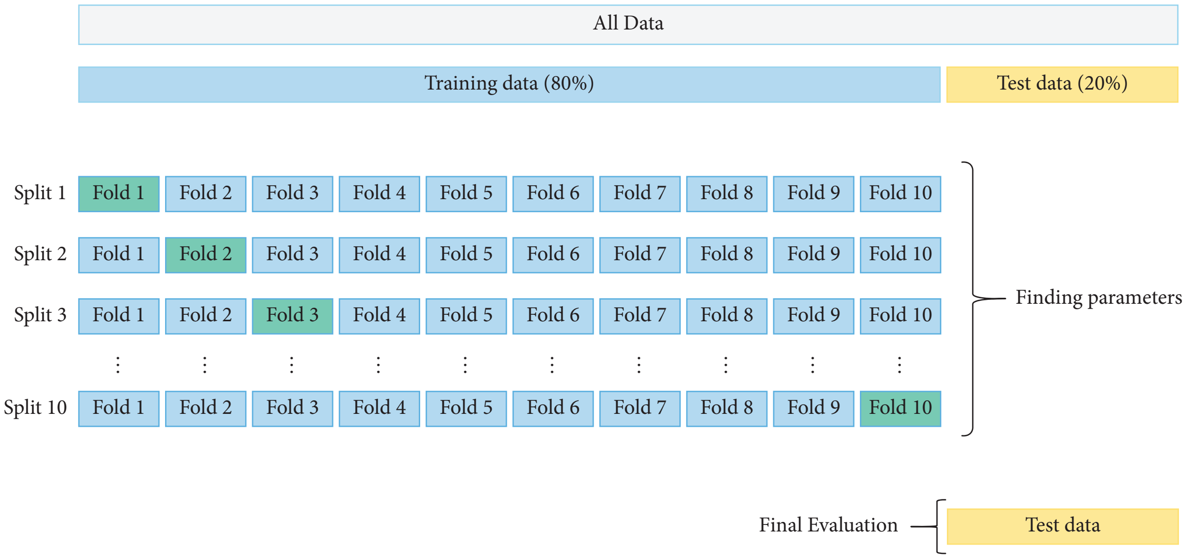


Figure 8: Overview of the 10-fold cross-validation procedure.

4.4. Hyperparameter Tuning. As stated previously, the gradient boosting algorithm incorporates three parts, including loss function, additive boosting, and decision tree, each of which has its own configuration. The performance of a gradient boosting model for a dataset can significantly fluctuate by changing the model architecture. Therefore, finding the optimal model is a key step for precise prediction. Calibrating models with different configurations to find the optimal model is commonly known as hyperparameter tuning, and the parameters are called hyperparameters. In this study, hyperparameter tuning is carried out using the *RandomizedSearchCV* module in the Scikit-learn package [76]. The *RandomizedSearchCV* randomly runs a set of hyperparameters and computes the scores and then returns the best set of parameters which yields the highest score as an output. The optimized model determined by this module is relied on the root mean squared error for the cross-validation score; therefore, the optimized model is not overfitted at all. The hyperparameters and the optimal model of the GB model are described in Table 2.

4.5. Accuracy Assessment. The performance of models was evaluated with standard statistical measures of MAE, RMSE, MAPE, and R^2. For a dataset containing N data with a target of y_i and prediction of f_i for i^{th} datum, these accuracy measurements can be expressed as follows:

(i) The MAE stands for mean absolute error, indicating the average absolute error for all predictions. The lower value of MAE reveals the lower error in a model. It can be measured as the following equation:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - f_i\right|. \tag{11}$$

(ii) The RMSE stands for the root mean squared error, a measurement of error produced in the model prediction. Therefore, the lower RMSE, the higher accuracy is attained. The RMSE parameter can be calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}\left(y_i - f_i\right)^2\right)}. \tag{12}$$

(iii) MAPE introduces the mean absolute percentage error, which is a relatively intuitive measure. Model performance improves as MAPE approaches 0. MAPE can be computed as follows:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - f_i}{y_i}\right| \times 100. \tag{13}$$

(iv) The R^2 is the coefficient of determination in regression problems that measures how well a model predicts targets. The R^2 ranges from 0 to 1, and the higher value represents the better performance of a model. The R^2 relates to the ratio of the residual sum of squares, SS_{res}, to the total sum of squares, SS_{tot}, and can be computed as follows:

$$\begin{gathered} R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - f_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}, \\ \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \end{gathered} \tag{14}$$

where $\overline{y}$ is the average of targets.

4.6. Comparison Models. In order to evaluate the performance of the gradient boosting model in predicting UCS of biocemented sands, its error metrics are compared with five commonly used machine learning techniques, including artificial neural network (ANN) [77], random forest (RF)

Table 2: Hyperparameter of gradient boosting algorithm according to Scikit-learn library [76].

	Parameter	Description	Range	Optimum value
Loss function	*loss*	Loss function to be optimized	[“*ls*,” “*lad*,” “*huber*”]	*lad*
Additive model	*learning_rate*	Shrinks the contribution of each tree by learning_rate	[0.01, 0.05, 0.1]	0.01
	n_estimators	The number of boosting stages to perform	[500, 1000, 1500]	1000
Decision tree	*max_depth*	Maximum depth of the individual regression estimators	[5, 15, 25]	15
	min_samples_leaf	The minimum number of samples required to be at a leaf node	[1,10, 25]	1
	min_samples_split	The minimum number of samples required to split an internal node	[5, 10, 15]	5
	max_features	The number of features to consider when looking for the best split	[“*sqrt*,” “log2,”, None]*	log2

*“ls” refers to the squared error of regression, “*lad*” refers to the absolute error of regression, and “*huber*” is a combination of the two. *The value of *max_features* is considered as follows (*n_features* refers to the number of features): (i) If “*sqrt*,” then *max_features = sqrt (n_features).* (ii) If “log2,” then *max_features* = log_2 *(n_features).* (iii) If *None*, then *max_features = n_features.*

[78], k-nearest neighbor (KNN) [79], support vector regression (SVR) [80], and decision tree (DT) [81].

Moreover, the results of this study are compared with those of Wang and Yin [8], who predicted the UCS of biocemented sands. They employed a multiexpression programming method combined with the Monte–Carlo method (MEP-MC) that relies on an evolutionary algorithm for developing mathematical expressions [82]. In the MEP-MC, five groups were constructed based on a database, and then a model was fitted for each group. The database used in their study was smaller than this study, contained 351 UCS test results. Wang and Yin [8], in contrast with this study, did not consider the urease activity of bacteria as an input variable.

5. Results and Discussion

5.1. Models Performance. Table 3 summarizes the error metrics of gradient boosting methods and other models for training and testing sets. It is evident that gradient boosting (GB) outperforms other algorithms in predicting the unconfined compressive strength of biocemented sands. Predictions made by GB produced MAE equal to 34 kPa for the training set and 229 kPa for the testing set. In other words, when a test datum is introduced to the optimal GB model with parameters presented in Table 2, its UCS can be predicted with an average error of 229 kPa. In the dataset, the mean value of UCS is 1328 kPa (Table 1); thus, it can be stated that the mean absolute error produced by GB is 17 percent of the mean value of UCS over the entire dataset. Furthermore, the RMSE of the GB shows a similar trend which is equal to 404 kPa for the test set. The parameter of MAPE can better explore the superiority of GB to other algorithms, which is a scale-independent and interpretable error parameter. The UCS values estimated through GB show an MAPE equal to 25% for the test set, while other algorithms have MAPE in a range of 36 to 54%. Therefore, it can be stated that the GB algorithm is capable of predicting the UCS of biocemented sand with an average error of 25%.

As stated in the literature review, random forest (RF) is an ensemble algorithm consisting of several parallel learners; in contrast, gradient boosting consists of several sequential learners. It can be seen from Table 3 that the GB technique is far more robust than RF in predicting the UCS of sands treated with MICP. The RF algorithm makes predictions with MAE and RMSE that are 62 and 44% higher than GB, respectively. Moreover, MAPE obtained with RF for the test set is equal to 44.8%, which is almost 20% greater than GB. According to these observations, the sequential harnessing of weak learners is far more efficient than parallel ones for predicting the unconfined compressive strength of sands treated with MICP.

Moreover, the performance of the multiexpression programming method (MEP-MC) performed by Wang and Yin [8] is presented in Table 3. Gradient boosting is clearly superior to MEP-MC in all aspects of error metrics. The MAE and RMSE of predictions obtained from MEP-MC were 409 and 652 kPa, respectively, which are 78 and 61 percent greater than those obtained from the GB model.

The distribution of predicted UCS versus actual UCS for the training and testing sets are exhibited in Figures 9 and 10, respectively. It can be seen that the predictions made for the training set are mostly close to or equal to the targets, and the majority of points in Figure 9(a) lie along the line of equality. The error distribution in Figure 9(b) shows that more than 200 of the training data have no error in their estimation. The distribution of the test set, shown in Figure 10(a), corroborates the reliability of the GB model. The test set predictions are well concentrated around the line of equality, demonstrating the strong correlation between predicted and actual UCS. According to Table 3, the coefficient of determination (R^2) for the test set of the GB model is equal to 0.95. Additionally, the produced errors for the test set are distributed normally in Figure 10, with the majority being lower than 500 kPa.

5.2. Reliability Analysis. In order to establish the effectiveness and dependability of the algorithms, a reliability analysis based on the Friedman analysis is performed [17]. According to this method, the models are ranked according to their errors in their predictions, from 1 indicating the least error to z indicating the highest error, for z models. For a database containing N data, the average ranking for model j ($\overline{r}_j$) can be calculated using the following formula:

$$\overline{r}_j = \frac{1}{N}\sum_{i=1}^{N} r_j^i, \tag{15}$$

where r_j^i denotes the ranking of the i^{th} data for model j.

Using equation (15), the average ranking ($\overline{r}_j$) of all utilized models are computed and plotted on Figure 11. This plot illustrates the superiority of the gradient boosting method, which has the lowest average ranking in comparison to the other models. This point endorses the outperformance of GB over five other frequently used machine learning techniques in predicting the UCS of biocemented sands. To find out whether this variation in models' performances is significant or not, the chi-Square (χ_r^2) of the average ranking throughout the test set is computed as follows:

$$\chi_r^2 = \frac{12N}{z(z+1)}\left[\sum_{j=1}^{z}\overline{r}_j^{\,2} - \frac{z(z+1)^2}{4}\right], \tag{16}$$

where N is the number of test data, and z is the number of algorithms which is equal to 6 in this study. The chi-square test relies on null hypothesis with (z − 1) degrees of freedom, which would be rejected if the computed chi-square value is equal to or greater than the critical one at a prespecified level of significance [83]. The critical chi-square for a distribution similar to this study, with 5 degrees of freedom and considering 0.95 significance, is equal to 11.07. Using equation (16), chi-square is equal to 38.65 for this study; thus, it can be

Table 3: Performance of the gradient boosting method compared to common machine learning models.

	MAE		RMSE		MAPE		R^2	
	Training	Test	Training	Test	Training	Test	Training	Test
GB	34	229	142	404	2.7	25.0	0.99	0.95
RF	379	370	665	585	56.7	44.8	0.91	0.89
DT	324	353	622	561	34.2	36.1	0.92	0.90
ANN	350	399	690	617	52.8	54.7	0.90	0.88
SVR	220	319	600	549	19.7	36.8	0.92	0.91
KNN	2	349	13	601	0.5	37.7	0.99	0.89
MEP-MC [8]	378	409	593	652	—	—	0.91	0.86

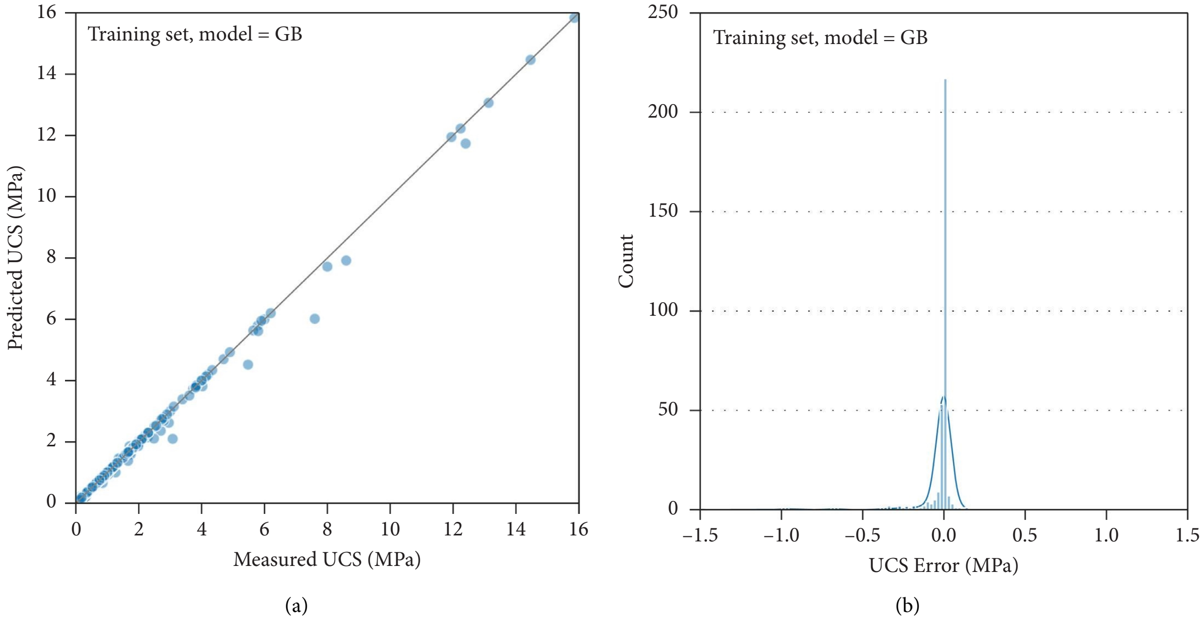


Figure 9: Results of gradient boosting model for training set: (a) relation of predicted UCS with actual values; (b) error distribution.

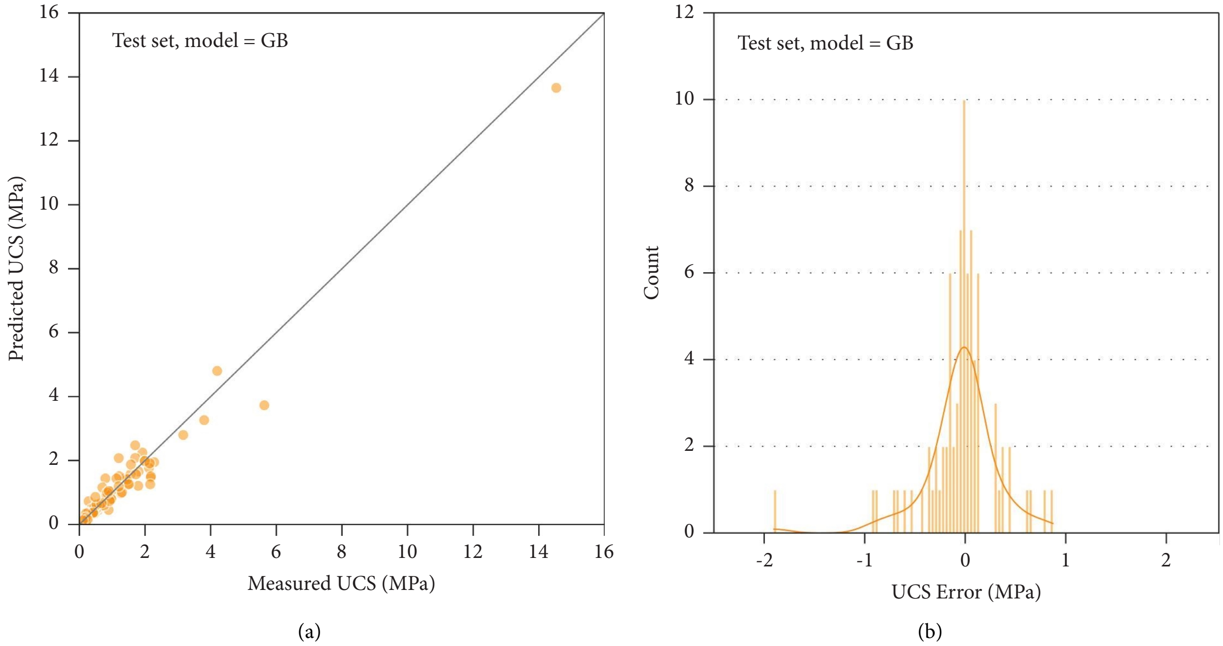


Figure 10: Results of the gradient boosting model for test set: (a) relation of predicted UCS with actual values; (b) error distribution.

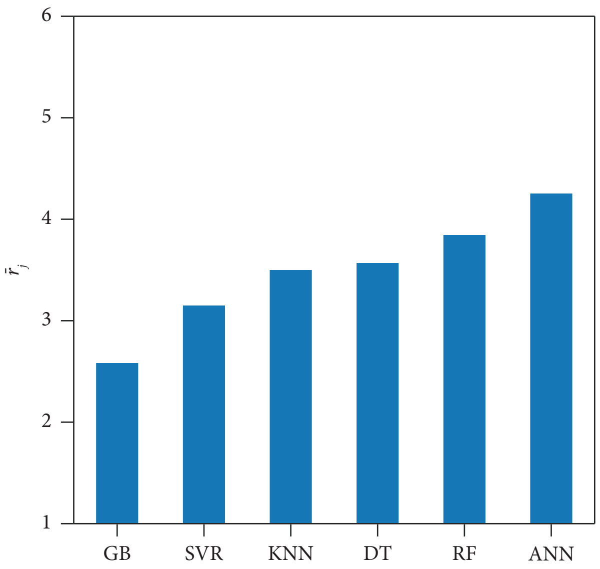


Figure 11: Average ranking of employed algorithms for test set.

concluded that the null hypothesis is rejected, so the distribution of models' performances is found to be significant.

5.3. Feature Importance. Generally, the gradient boosting technique is also capable of providing an importance score for each variable to recognize how valuable each feature is in the construction of the boosted decision trees. The feature importance score fluctuates within a range of 0 to 100, and the higher values for a variable demonstrate the greater importance. The results of the feature importance analysis of this study are presented in Figure 12. Similar to the heatmap outlined in Figure 5, calcite content (F_{CaCO3}) is found to be the most influential feature for the gradient boosting algorithm. The second most important feature is the initial void ratio (e_0), which has a 10% feature importance. The other features of the sands, bacteria, and cementation solutions have the lowest influence on the final UCS in the gradient boosting algorithm.

6. Environmental Modification

According to the literature review, the UCS of biocemented sand is influenced by the surrounding temperature and initial pH of the soil. However, given that all the test results included in the dataset were obtained from unconfined compression tests performed on neutral sand (pH = 7) at room temperature (20–30°C), these two variables are not included in training the models. It should be noted that the available data that focused on the effect of temperature and pH are too small that extracting a model based on these variables is almost impossible. As a solution for this limitation, this section provides guidelines for applying the effect of temperature and pH on the UCS of biocemented sands based on those small set of data. Given that these findings are based on a limited number of tests, the results from such analyses should be treated with considerable caution.

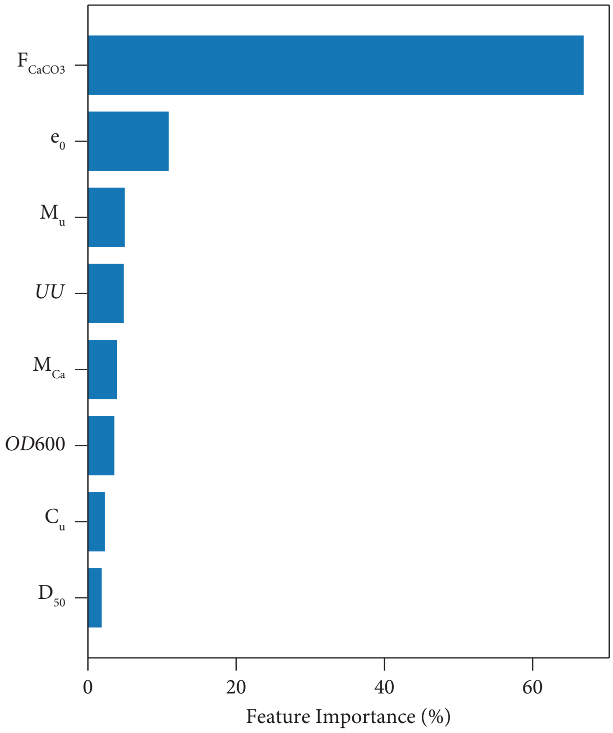


Figure 12: Results of feature importance analysis.

6.1. Temperature. Throughout the dataset used in this study, the temperature of the curing environment is close to room temperature (20–30°C). Research conducted by Cheng et al. [41] can present guidelines for modifying the predicted UCS to other temperatures. Cheng et al. [41] conducted a series of unconfined compression tests on sands treated with an identical treatment program but cured at three different curing temperatures (4, 25, and 50°C). It was reported that the strongest biocemented sands were cured under the temperature of 25°C. Since all samples were treated with similar properties, the UCS values of the samples corresponding to the temperatures of 50 and 4°C can be normalized with the temperature of 25°C. Therefore, the parameter of temperature coefficient, $r_{t,25^\circ C}$, is defined as follows:

$$r_{t,25^\circ C} = \frac{\mathrm{UCS}_T}{\mathrm{UCS}_{25^\circ C}}, \tag{17}$$

where UCS_T and $\mathrm{UCS}_{25^\circ C}$ are the value of unconfined compressive strength for specimens at a temperature of T and 25°C. It should be mentioned that the calcite content is equal for both samples. The parameter of $r_{t,25^\circ C}$ introduces the ratio of UCS of sands treated at a temperature of T to 25°C. The distribution of $r_{t,25^\circ C}$ in the study of Cheng et al. [41] is illustrated in Figure 13. It can be observed that specimens treated under 4°C have $r_{t,25^\circ C}$ values in a range of 0.55 to 0.85 with an increasing trend line. In contrast, the values of $r_{t,25^\circ C}$ for sands treated at 50°C decline as F_{CaCO3} values increase, ranging from 0.3 to 0.5.

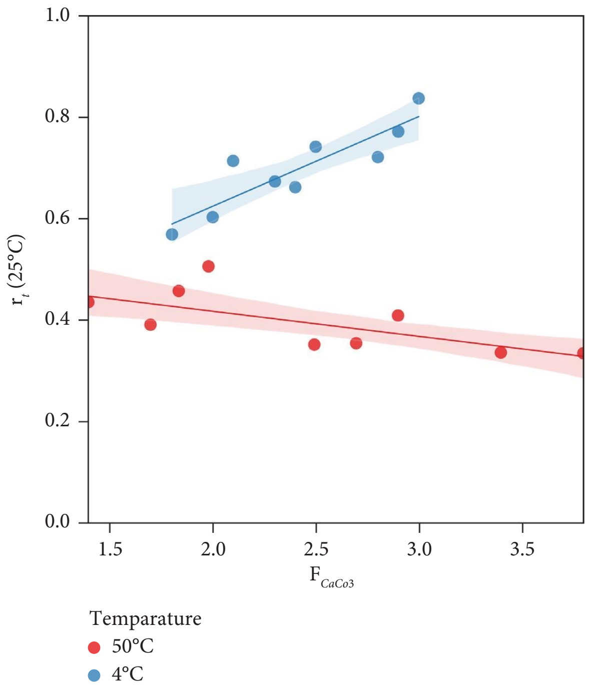


Figure 13: Variation of temperature coefficient ($r_{t,25°C}$) for various $CaCO_3$ content (F_{CaCO3}) at different temperatures [41].

When it is aimed to estimate the UCS of bio-cemented sands cured in a hotter or colder environment, the trend lines in Figure 13 can be used to adopt the UCS values estimated using GB models. The predicted UCS should be multiplied with the corresponding $r_{t,25°C}$ with regards to the temperature and $CaCO_3$ content. Although the lack of experimental data related to different temperatures restricts temperature modification, these results can be conducive to providing insight into other temperatures.

6.2. pH. As stated previously, both acidity ($pH < 7$) and alkalinity ($pH > 7$) negatively impact the UCS of sands treated by MICP. The degree of UCS reduction cannot be accurately estimated due to the lack of high-quality literature with extensive datasets; however, the results of Cheng et al. [63] could provide an initial guideline. Cheng et al. [63] demonstrated that sands with pH levels equal to 9.5 and 3.5 endure lower UCS than neutral sand, even with high levels of $CaCO_3$ crystals. Acidic sand showed higher drop rates than alkaline ones: the UCS of acidic sand was approximately 25% of neutral sand, whereas the UCS of alkaline sands was 50% of neutral sand. Therefore, when estimating the UCS of acidic or alkaline sands treated by MICP, the final UCS of acidic and alkaline sands can be considered to be 25 and 50% of neutral sand, respectively.

7. Conclusions and Future Works

Given the environmental benefits and wide application of microbially-induced calcite precipitation of sandy soils, the unconfined compressive strength of sands treated with MICP was predicted using a gradient boosting technique in this study. Based on a dataset consisting of 402 data extracted from previous studies, the findings can be summarized as follows:

(i) An acceptable performance of the gradient boosting algorithm was achieved in predicting the UCS of biocemented sands in neutral condition ($pH = 7$) and room temperature (20–30°C). For the test set, predictions made by the gradient boosting had MAE and RMSE equal to 229 and 404 kPa, respectively. Also, MAPE and R^2 were 25% and 0.95, respectively. The comparison of error metrics with five other frequently used machine learning techniques (ANN, SVR, KNN, RF, and DT) demonstrated the outperformance of the gradient boosting in all aspect of error metrics.

(ii) The correlation analysis revealed that the UCS of biocemented sands mostly depends on the calcite content. Furthermore, a correlation was found between the void ratio and calcite content suggesting that high levels of CaCO3 precipitation could occur in soils with a void ratio between 0.6 and 0.9.

(iii) Using existing literature on the UCS of biocemented sands in harsh environments, the guidelines were developed for modifying predicted values. These analyses revealed a trend for low calcite samples in cold (4°C) and hot (50°C) weather. Furthermore, biocemented sands treated in alkaline and acidic environments showed lower UCS than neutral ones. These modifications were limited to a specific range of temperature and pH level because few data are available for performing analysis.

Overall, this study provides valuable insights into the application of machine learning algorithms in predicting the UCS of biocemented sands treated with MICP, which can be useful for civil engineering applications. However, further experimental studies with clear and detailed treatment procedures (particularly injection details) can be reinforce the database for developing our models and study. MICP treatment of sands with varying void ratios can provide valuable insight into determining the optimal initial condition for the MICP treatment. Also, further research at a variety of temperatures and pH levels is needed to enhance the accuracy and feasibility of the environmental modifications.

Notation

$r_{t,25°C}$:	Temperature coefficient for 25°C
C_u:	Uniformity coefficient
D_{50}:	Median sand particle size
e_0:	Initial void ratio
f:	Function estimate
$\widehat{f}$ ( ):	Final boosted function at iteration M
f_m ( ):	Function for the m^{th} iteration
F_{CaCO3}:	Calcite content
k:	Number of folds in cross-validation

$L(\ )$:	Loss function
L_{Huber}:	Huber function
L_{lad}:	Absolute error loss function
L_{ls}:	Squared error loss function
Learning_rate:	Shrinkage parameter
MAE:	Mean absolute error
MAPE:	Mean absolute percentage error
max_depth:	The maximum depth that limits the growth of trees
max_features:	Number of features to consider when looking for the best split
M_{ca}:	Calcium chloride concentration
min_samples_leaf:	Minimum number of samples required to be at an internal or external node
min_samples_split:	Minimum number of samples required to split an internal node
M_u:	Urea concentration
n_estimators:	Number of iterations
OD_{600}:	Optical density of biomass at 600 nm
R^2:	Coefficient of determination
R_j:	Region j in a tree
R_{jm}:	Region j in a tree for m^{th} iteration
r_{im}:	Negative gradient at m^{th} iteration
r_j^i:	Ranking of the i^{th} data for model j
$\overline{r}_j$:	Average ranking for model j
RMSE:	Root mean squared error
UA:	Urease activity
UCS:	Unconfined compressive strength
$\text{UCS}_{25^\circ\text{C}}$:	Unconfined compressive strength at a temperature of 25°C
UCS_T:	Unconfined compressive strength at a temperature of T
v:	Learning rate (shrinkage)
x_i:	Input variables of the i^{th} sample
$\overline{y}$:	Average of targets
y_i:	Target variable of the i^{th} sample
z:	Number of algorithms
α:	Breakdown point parameter in Huber loss function
α_m:	Weight factor for the m^{th} sample
γ_j:	Constant for terminal j
γ_{jm}:	Optimal constants in each region at m^{th} iteration
δ:	Threshold of Huber loss function
χ_r^2:	Chi-square in Friedman analysis.

Data Availability

The dataset used in this study is included within the supplementary materials.

Conflicts of Interest

The author declares that he has no conflicts of interest that could have appeared to influence the work reported in this paper.

Supplementary Materials

A supplementary material file is provided that contains the dataset used in this study. The dataset includes the values for input and target parameters sorted by referenced sources and are provided in a text file. The notations and references numbers of the dataset are according to the main article. (*Supplementary Materials*)

References

[1] C. Chen, G. Habert, Y. Bouzidi, and A. Jullien, "Environmental impact of cement production: detail of the different processes and cement plant variability evaluation," *Journal of Cleaner Production*, vol. 18, no. 5, pp. 478–485, 2010.

[2] A. Akbarpour, M. Mahdikhani, and R. Ziaie Moayed, "Mechanical behavior and permeability of plastic concrete containing natural zeolite under triaxial and uniaxial compression," *Journal of Materials in Civil Engineering*, vol. 34, no. 2, Article ID 4021453, 2022.

[3] A. Akbarpour, M. Mahdikhani, and R. Z. Moayed, "Effects of natural zeolite and sulfate ions on the mechanical properties and microstructure of plastic concrete," *Frontiers of Structural and Civil Engineering*, vol. 16, no. 1, pp. 86–98, 2022.

[4] D. Mujah, M. A. Shahin, and L. Cheng, "State-of-the-Art review of biocementation by microbially induced calcite precipitation (MICP) for soil stabilization," *Geomicrobiology Journal*, vol. 34, no. 6, pp. 524–537, 2017.

[5] L. Chen, Y. Song, H. Fang, Q. Feng, C. Lai, and X. Song, "Systematic optimization of a novel, cost-effective fermentation medium of Sporosarcina pasteurii for microbially induced calcite precipitation (MICP)," *Construction and Building Materials*, vol. 348, Article ID 128632, 2022.

[6] K. M. N. S. Wani and B. A. Mir, "Microbial geo-technology in ground improvement techniques: a comprehensive review," *Innov. Infrastruct. Solut*, vol. 5, no. 3, p. 82, 2020.

[7] J. Liu, G. Li, and X. Li, "Geotechnical engineering properties of soils solidified by microbially induced CaCO3 precipitation (MICP)," *Advances in Civil Engineering*, vol. 2021, Article ID 6683930, 21 pages, 2021.

[8] H. L. Wang and Z. Y. Yin, "Unconfined compressive strength of bio-cemented sand: state-of-the-art review and MEP-MC-based model development," *Journal of Cleaner Production*, vol. 315, Article ID 128205, 2021.

[9] P. Zhang, Z. Y. Yin, and Y. F. Jin, "Machine learning-based modelling of soil properties for geotechnical design: review, tool development and comparison," *Archives of Computational Methods in Engineering*, vol. 29, no. 2, pp. 1229–1245, 2022.

[10] M. I. Shah, M. F. Javed, A. Alqahtani, and A. Aldrees, "Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data," *Process Safety and Environmental Protection*, vol. 151, pp. 324–340, 2021.

[11] A. Alqahtani, M. I. Shah, A. Aldrees, and M. F. Javed, "Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality," *Sustainability*, vol. 14, no. 3, p. 1183, 2022.

[12] M. Kamran, "A state of the art catboost-based T-distributed stochastic neighbor embedding technique to predict back-

break at dewan cement limestone quarry," *International Journal of Mining, Reclamation and Environment*, vol. 12, pp. 679–691, 2021.

[13] M. F. Javed, F. Farooq, S. A. Memon et al., "New prediction model for the ultimate axial capacity of concrete-filled steel tubes: an evolutionary approach," *Crystals*, vol. 10, no. 9, Article ID 741, 2020.

[14] N. García-Pedrajas, C. García-Osorio, and C. Fyfe, "Nonlinear boosting projections for ensemble construction," *Journal of Machine Learning Research*, vol. 8, pp. 1–33, 2007.

[15] M. Kamran, R. K. Wattimena, D. J. Armaghani, P. G. Asteris, I. M. Jiskani, and E. T. Mohamad, "Intelligent based decision-making strategy to predict fire intensity in subsurface engineering environments," *Process Safety and Environmental Protection*, vol. 171, pp. 374–384, 2023.

[16] H.-B. Ly, T.-A. Nguyen, and B. T. Pham, "Estimation of soil cohesion using machine learning method: a random forest approach," *Advances in Civil Engineering*, vol. 2021, Article ID 8873993, 14 pages, 2021.

[17] S. Talamkhani, S. A. Naeini, and A. Ardakani, "Prediction of static liquefaction susceptibility of sands containing plastic fines using machine learning techniques," *Geotechnical & Geological Engineering*, vol. 41, no. 5, pp. 3057–3074, 2023.

[18] M. I. Shah, M. F. Javed, F. Aslam, and H. Alabduljabbar, "Machine learning modeling integrating experimental analysis for predicting the properties of sugarcane bagasse ash concrete," *Construction and Building Materials*, vol. 314, Article ID 125634, 2022.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[20] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geoscience Frontiers*, vol. 12, no. 1, pp. 469–477, 2021.

[21] M. Rezaee, S. F. F. Mojtahedi, E. Taherabadi, K. Soleymani, and M. Pejman, "Prediction of shear strength parameters of hydrocarbon contaminated sand based on machine learning methods," *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, vol. 15, no. 4, pp. 317–335, 2021.

[22] N. M. Shahani, M. Kamran, X. Zheng, C. Liu, and X. Guo, "Application of gradient boosting machine learning algorithms to predict uniaxial compressive strength of soft sedimentary rocks at Thar coalfield," *Advances in Civil Engineering*, vol. 2021, Article ID 2565488, 19 pages, 2021.

[23] D. Karir, A. Ray, A. Kumar Bharati, U. Chaturvedi, R. Rai, and M. Khandelwal, "Stability prediction of a natural and man-made slope using various machine learning algorithms," *Transportation Geotechnics*, vol. 34, Article ID 100745, 2022.

[24] L. Wang, C. Wu, L. Tang et al., "Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method," *Acta Geotech*, vol. 15, no. 11, pp. 3135–3150, 2020.

[25] B. Ullah, M. Kamran, and Y. Rui, "Predictive modeling of short-term rockburst for the stability of subsurface structures using machine learning approaches: T-SNE, K-means clustering and XGBoost," *Mathematics*, vol. 10, no. 3, p. 449, 2022.

[26] W. G. Zhang, H. R. Li, C. Z. Wu, Y. Q. Li, Z. Q. Liu, and H. L. Liu, "Soft computing approach for prediction of surface settlement induced by earth pressure balance shield tunneling," *Underground Space*, vol. 6, no. 4, pp. 353–363, 2021.

[27] D. M. Zhang, J. Z. Zhang, H. W. Huang, C. C. Qi, and C. Y. Chang, "Machine learning-based prediction of soil compression modulus with application of 1D settlement," *Journal of Zhejiang University Science A*.vol. 21, no. 6, pp. 430–444, 2020.

[28] R. Zhang, Y. Li, A. T. C. Goh, W. Zhang, and Z. Chen, "Analysis of ground surface settlement in anisotropic clays using extreme gradient boosting and random forest regression models," *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 13, no. 6, pp. 1478–1484, 2021.

[29] J. Zhou, X. Chen, M. Wang, E. Li, H. Chen, and X. Shi, "Classification of seismic-liquefaction potential using friedman's stochastic gradient boosting based on the cone penetration test data," in *Sustainable Civil Infrastructures*, pp. 67–78, Springer Science and Business Media B.V Berlin, Germany, 2019.

[30] Z. Chen, H. Li, A. T. C. Goh, C. Wu, and W. Zhang, "Soil liquefaction assessment using soft computing approaches based on capacity energy concept," *Geosciences*, vol. 10, no. 9, pp. 330-331, 2020.

[31] J. Zhou, E. Li, M. Wang, X. Chen, X. Shi, and L. Jiang, "Feasibility of stochastic gradient boosting approach for evaluating seismic liquefaction potential based on SPT and CPT case histories," *Journal of Performance of Constructed Facilities*, vol. 33, no. 3, Article ID 04019024, 2019.

[32] H. Behzadipour, M. S. Pakbaz, and G. R. Ghezelbash, "Effects of biocementation on strength parameters of silty and clayey sands," *Bioinspired, Biomimetic and Nanobiomaterials*, vol. 9, no. 1, pp. 24–32, 2020.

[33] L. Cheng and M. A. Shahin, "Microbially induced calcite precipitation (MICP) for soil stabilization," *Ecological Wisdom Inspired Restoration Engineering*, Springer, Singapore, pp. 47–68, 2019.

[34] V. S. Whiffin, L. A. van Paassen, and M. P. Harkes, "Microbial carbonate precipitation as a soil improvement technique," *Geomicrobiology Journal*, vol. 24, no. 5, pp. 417–423, 2007.

[35] T. Jahns, "Ammonium/urea-dependent generation of a proton electrochemical potential and synthesis of ATP in Bacillus pasteurii," *Journal of Bacteriology*, vol. 178, no. 2, pp. 403–409, 1996.

[36] S. Stocks-Fischer, J. K. Galinat, and S. S. Bang, "Microbiological precipitation of CaCO3," *Soil Biology and Biochemistry*, vol. 31, no. 11, pp. 1563–1571, 1999.

[37] S.-G. Choi, I. Chang, M. Lee, J.-H. Lee, J.-T. Han, and T.-H. Kwon, "Review on geotechnical engineering properties of sands treated by microbially induced calcium carbonate precipitation (MICP) and biopolymers," *Construction and Building Materials*, vol. 246, Article ID 118415, 2020.

[38] A. Zapata and S. Ramirez-Arcos, "A comparative study of McFarland turbidity standards and the densimat photometer to determine bacterial cell density," *Current Microbiology*, vol. 70, no. 6, pp. 907–909, 2015.

[39] D. Mujah, L. Cheng, and M. A. Shahin, "Microstructural and geomechanical study on biocemented sand for optimization of MICP process," *Journal of Materials in Civil Engineering*, vol. 31, no. 4, Article ID 04019025, 2019.

[40] Q. Zhao, L. Li, C. Li, M. Li, F. Amini, and H. Zhang, "Factors affecting improvement of engineering properties of MICP-treated soil catalyzed by bacteria and urease," *Journal of Materials in Civil Engineering*, vol. 26, no. 12, Article ID 04014094, 2014.

[41] L. Cheng, M. A. Shahin, and D. Mujah, "Influence of key environmental conditions on microbially induced cementation for soil stabilization," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 143, no. 1, Article ID 04016083, 2017.

[42] Y. Wang, K. Soga, J. T. DeJong, and A. J. Kabla, "Effects of bacterial density on growth rate and characteristics of microbial-induced CaCO3 precipitates: particle-scale experimental study," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 147, no. 6, 2021.

[43] I. A. Hammad, F. N. Talkhan, and A. E. Zoheir, "Urease activity and induction of calcium carbonate precipitation by Sporosarcina pasteurii NCIMB 8841," *Journal of Applied Sciences Research*, vol. 9, pp. 1525–1533, 2013.

[44] L. Cheng and R. Cord-Ruwisch, "Upscaling effects of soil improvement by microbially induced calcite precipitation by surface percolation," *Geomicrobiology Journal*, vol. 31, pp. 396–406, 2014.

[45] L. A. van Paassen, R. Ghose, T. J. M. van der Linden, W. R. L. van der Star, and M. C. M. van Loosdrecht, "Quantifying biomediated ground improvement by ureolysis: large-scale biogrout experiment," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 136, no. 12, pp. 1721–1728, 2010.

[46] T. R. Ginn, E. M. Murphy, A. Chilakapati, and U. Seeboonruang, "Stochastic-convective transport with nonlinear reaction and mixing: application to intermediate-scale experiments in aerobic biodegradation in saturated porous media," *Journal of Contaminant Hydrology*, vol. 48, no. 1-2, pp. 121–149, 2001.

[47] M. P. Harkes, L. A. van Paassen, J. L. Booster, V. S. Whiffin, and M. C. M. van Loosdrecht, "Fixation and distribution of bacterial activity in sand to induce carbonate precipitation for ground reinforcement," *Ecological Engineering*, vol. 36, no. 2, pp. 112–117, 2010.

[48] A. Al Qabany, K. Soga, and C. Santamarina, "Factors affecting efficiency of microbially induced calcite precipitation," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 138, no. 8, pp. 992–1001, 2012.

[49] A. Al Qabany and K. Soga, "Effect of chemical treatment used in MICP on engineering properties of cemented soils, Bio-Chemo-Mechanical Process," *Geotechnical Engineering*, vol. 107–115, 2014.

[50] H. Behzadipour and A. Sadrekarimi, "Biochar-assisted biocementation of a sand using native bacteria," *Bulletin of Engineering Geology and the Environment*, vol. 80, no. 6, pp. 4967–4984, 2021.

[51] A. Nafisi, Q. Liu, and B. M. Montoya, "Effect of stress path on the shear response of bio-cemented sands," *Acta Geotech*, vol. 16, no. 10, pp. 3239–3251, 2021.

[52] W. De Muynck, N. De Belie, and W. Verstraete, "Microbial carbonate precipitation in construction materials: a review," *Ecological Engineering*, vol. 36, no. 2, pp. 118–136, 2010.

[53] Y. Zhang, H. X. Guo, and X. H. Cheng, "Influences of calcium sources on microbially induced carbonate precipitation in porous media," *Materials Research Innovations*, vol. 18, no. 2, pp. S2–S79, 2014.

[54] V. Achal and X. Pan, "Influence of calcium sources on microbially induced calcium carbonate precipitation by Bacillus sp. CR2," *Applied Biochemistry and Biotechnology*, vol. 173, no. 1, pp. 307–317, 2014.

[55] L. Cheng, R. Cord-Ruwisch, and M. A. Shahin, "Cementation of sand soil by microbially induced calcite precipitation at various degrees of saturation," *Canadian Geotechnical Journal*, vol. 50, no. 1, pp. 81–90, 2013.

[56] Y. Wang, C. Konstantinou, K. Soga, J. T. DeJong, G. Biscontin, and A. J. Kabla, "Enhancing strength of MICP-treated sandy soils: from micro to macro scale," 2020, https://arxiv.org/abs/2006.15760.

[57] G. D. O. Okwadha and J. Li, "Optimum conditions for microbial carbonate precipitation," *Chemosphere*, vol. 81, no. 9, pp. 1143–1148, 2010.

[58] A. Mahawish, A. Bouazza, and W. P. Gates, "Factors affecting the bio-cementing process of coarse sand," *Proceedings of the Institution of Civil Engineers Ground Improvement*, vol. 172, no. 1, pp. 25–36, 2019.

[59] J. Wojtowicz, "Factors affecting precipitation of calcium carbonate," *Journal of the Swimming Pool and Spa Industry*, vol. 3, pp. 18–23, 1998.

[60] K. L. Sahrawat, "Effects of temperature and moisture on urease activity in semi-arid tropical soils," *Plant and Soil*, vol. 78, no. 3, pp. 401–408, 1984.

[61] M. J. McWhirter, A. J. McQuillan, and P. J. Bremer, "Influence of ionic strength and pH on the first 60 min of *Pseudomonas aeruginosa* attachment to ZnSe and to TiO2 monitored by ATR-IR spectroscopy," *Colloids and Surfaces B: Biointerfaces*, vol. 26, no. 4, pp. 365–372, 2002.

[62] S. Liu, K. Wen, C. Armwood et al., "Enhancement of MICP-treated sandy soils against environmental deterioration," *Journal of Materials in Civil Engineering*, vol. 31, no. 12, Article ID 04019294, 2019.

[63] L. Cheng, M. A. Shahin, and R. Cord-Ruwisch, "Soil stabilisation by microbial-induced calcite precipitation (MICP): investigation into some physical and environmental aspects," in *Proceedings of the 7th Int. Congr. Environ. Geotech. ICEG 2014*, vol. 64, pp. 1105–1112, Melbourne, Australia, November, 2014.

[64] A. I. Omoregie, G. Khoshdelnezamiha, N. Senian, D. E. L. Ong, and P. M. Nissom, "Experimental optimisation of various cultural conditions on urease activity for isolated Sporosarcina pasteurii strains and evaluation of their biocement potentials," *Ecological Engineering*, vol. 109, pp. 65–75, 2017.

[65] G. Kim, J. Kim, and H. Youn, "Effect of temperature, pH, and reaction duration on microbially induced calcite precipitation," *Applied Sciences*, vol. 8, p. 1277, 2018.

[66] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[67] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2010.

[68] L. Breiman, "Using adaptive bagging to debias regressions," Statistics Dept. UCB, Technical Report 547, 1999.

[69] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013.

[70] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, pp. 73–101, 1964.

[71] G. Ridgeway, "Generalized Boosted Models: a guide to the gbm package," *Update*, vol. 1, pp. 1–12, 2007.

[72] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021.

[73] A. Mahawish, A. Bouazza, and W. P. Gates, "Unconfined compressive strength and visualization of the microstructure of coarse sand subjected to different biocementation levels," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 145, no. 8, Article ID 04019033, 2019.

[74] Astm D422, "Standard test method for particle-size analysis of soils," Astm. D422-63, 2007, https://www.astm.org/standards/d422.

[75] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Springer Top Signal Process*, pp. 1–4,

Springer Science and Business Media B.V Berlin, Germany, 2009.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[77] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[78] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[79] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood Components Analysis," in *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, USA, 2005.

[80] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[81] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, CRC Press, Boca Raton, FL, USA, 2017.

[82] M. Oltean and D. Dumitrescu, "Multi Expression Programming," J Genet Program Evolvable Mach, 01, 2002.

[83] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, FL, USA, Fifth edition, 2011.