Geomechanics and Engineering, *Vol. 36, No. 4 (2024) 381-390*
https://doi.org/10.12989/gae.2024.36.4.381

Machine learning-based analysis and prediction model on the strengthening mechanism of biopolymer-based soil treatment

Haejin Lee†1a, Jaemin Lee†2b, Seunghwa Ryu**2 and Ilhan Chang*1

[1]*Department of Civil Systems Engineering, Ajou University, Republic of Korea, 16449*
[2]*Department of Mechanical Engineering, Korea Advance Institute of Science and Technology, Republic of Korea, 34141*

(*Received December 21, 2023, Revised January 15, 2024, Accepted January 16, 2024*)

Abstract. The introduction of bio-based materials has been recommended in the geotechnical engineering field to reduce environmental pollutants such as heavy metals and greenhouse gases. However, bio-treated soil methods face limitations in field application due to short research periods and insufficient verification of engineering performance, especially when compared to conventional materials like cement. Therefore, this study aimed to develop a machine learning model for predicting the unconfined compressive strength, a representative soil property, of biopolymer-based soil treatment (BPST). Four machine learning algorithms were compared to determine a suitable model, including linear regression (LR), support vector regression (SVR), random forest (RF), and neural network (NN). Except for LR, the SVR, RF, and NN algorithms exhibited high predictive performance with an R^2 value of 0.98 or higher. The permutation feature importance technique was used to identify the main factors affecting the strength enhancement of BPST. The results indicated that the unconfined compressive strength of BPST is affected by mean particle size, followed by biopolymer content and water content. With a reliable prediction model, the proposed model can present guidelines prior to laboratory testing and field application, thereby saving a significant amount of time and money.

Keywords: biopolymer-based soil treatment (BPST); machine learning; neural network; random forest; support vector regression; unconfined compressive strength

1. Introduction

Environmentally friendly biomaterials for soil improvement and stabilization have been actively studied and discovered in geotechnical engineering, recently (Almajed *et al.* 2021, Chang *et al.* 2020, Choi *et al.* 2020, Fatehi *et al.* 2021). Biological and bio-induced soil treatment techniques using microbial biopolymers (e.g., biopolymer-based soil treatment; BPST) (Chang *et al.* 2020, Kwon *et al.* 2023, Seo *et al.* 2021), inter-granular $CaCO_3$ mineralization using *in situ* microbes (e.g., microbially induced calcite precipitation; MICP), or enzymes (e.g., enzyme induced calcite precipitation; EICP) (Choi *et al.* 2020, Mekonnen *et al.* 2022) have been actively employed for sustainable ground improvement purposes. They compensate for environmental concerns, such as heavy metal leakage, ground alkalinization, and carbon dioxide emissions, which were known to be drawbacks of using cement, which is the most common material used in ground improvement practices (Chang *et al.* 2016, Worrell *et al.* 2001).

Although these novel biomaterials address environmental problems, they have not been as effective as cement, which ensures durability, compressive strength, and workability of a mixture. Moreover, unlike concrete, the soil improvement field does not use standardized aggregates; however, in most cases, it uses various types of soil generated onsite. Therefore, uncertainty with regard to biological ground reinforcement methods causes limitations in field application. Traditional statistical approaches have limitations in evaluating and predicting the characteristics of new composite materials. Therefore, machine learning techniques are currently being used as an alternate statistical tool, which can help save cost and time.

Machine learning algorithms are widely used in problems where analytic solutions do not exist. These algorithms use data to acquire models that can be employed for making decisions or predictions. For several decades, various algorithms have been developed, including linear regression (LR), random forest (RF) (Breiman 2001), support vector machine (SVR) ((Cortes and Vapnik 1995, Drucker *et al.* 1996), and artificial neural networks (Abraham 2005). These algorithms have effectively extracted patterns and features from complex and high-dimensional data and have been applied to a wide range of problems, which may be difficult for humans to solve. At present, the application of machine learning algorithms is one of the most interesting techniques, specifically in the

*Corresponding author, Associate Professor
E-mail: ilhanchang@ajou.ac.kr
**Co-corresponding author, Professor
E-mail: ryush@kaist.ac.kr
[a]M.S.
E-mail: gowls0605@ajou.ac.kr
[b]Ph.D.
E-mail: leexamine@kaist.ac.kr
†Equal contribution

ISSN: 2005-307X (Print), 2092-6219 (Online)

Table 1 Information of materials and properties used in retrieved literature

No.	Soil type (USCS)	BP type	Tested properties	Number of data	Reference
1	Silty sand (SM) from Jaffna, Sri Lanka	XG	UCS	12	(Lee *et al.* 2019)
2	Residual soil (MH) from Ha-dong, Korea	BG	UCS	24	(Chang and Cho 2012)
3	Standard sand (SP) from Jumunjin, Korea Sandy soil (SP-SM) from Cheonan, Korea Residual soil (CL) from Gochang, Korea	XG	UCS	41	(Chang *et al.* 2015a)
4	Sandy soil (SP-SM) from Cheonan, Korea Residual soil (CL) from Gochang, Korea	GG, AG	UCS	230	(Chang *et al.* 2015c)
5	Standard sand (SP) from Jumunjin, Korea	GG	UCS	152	(Chang *et al.* 2017)
6	Soft marine soil (CH) from Yeosu, Korea	XG	UCS, VS, FC, AL	4	(Kwon *et al.* 2019)
7	Residual soil (CL) from Gochang, Korea	GG, XG	UCS	16	(Chang *et al.* 2015b)
Total number of UCS data				479	

*USCS: Unified soil classification system; UCS: Unconfined compressive strength; VS: Vane shear strength; FC: Fall cone undrained shear strength; AL: Atterberg limits; BP type: xanthan gum (XG), beta-glucan (BG), gellan gum (GG), and agar gum (AG)

field of civil and geotechnical engineering, to predict soil behavior related to erosion (Kim and Gilley 2008), landslides (Ma *et al.* 2022), soil liquefaction (Goh and Goh 2007, Njock *et al.* 2020), soil nailing (Benayoun *et al.* 2020), soil classification (Bhattacharya and Solomatine 2006), and soil attributes (Gonos and Stathopulos 2005, Kiran *et al.* 2016, Pham *et al.* 2018).

Herein, a machine learning model was applied to predict the unconfined compressive strength (UCS) of sand with biopolymer-based soil treatment (BPST). Biopolymer is produced by or derived from living organisms, e.g., plants and microbes. Biopolymers are commonly used in food and pharmaceuticals and have recently been emphasized as an eco-friendly plastic material (Bharti and Swetha 2016, Chang *et al.* 2020, Fatehi *et al.* 2023). Gel-type biopolymers, such as gellan gum, agar gum, xanthan gum, and beta-glucan, are used in geotechnical engineering. The effects of the surrounding environment are insignificant, and similar results can be obtained in repeated experiments. The unconfined compression test can be applied to adhesive materials that can be molded into samples. The test method is simple and thus produces a large number of laboratory results. Therefore, herein, accumulated experimental data were organized, and models that could predict the UCS of BPST were developed. The models enabled strength prediction and identified the influence of dominant variables on the behavior of BPST.

2. Methodology

2.1 Biopolymer-based soil treatment (BPST) UCS database configuration

For model training and testing purposes to predict the UCS of BPST, a database with 479 records was collected and compiled. These datasets are part of seven laboratory studies conducted at the Korea Advanced Institute of Science and Technology (KAIST) and the University of New South Wales (UNSW) Australia (Chang and Cho 2012, Chang *et al.* 2015a, c, d, 2017, Kwon *et al.* 2019, Lee *et al.* 2019) (Table 1). Although other properties (e.g., interparticle cohesion, friction angle, and hydraulic conductivity) of BPST were considered in the existing literature, only UCS-based data points were used for operating the models.

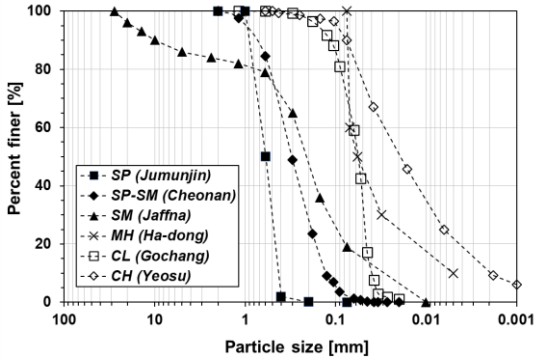

Fig. 1 Particle size distribution curves of soil types used for model training

Among the collected BPST UCS data, five different soil types (Jumunjin sand, Cheonan site soil, Gochang site soil, and Ha-dong site soil in Korea; Jaffna site soil from Sri Lanka) were used, which ranged from cohesive to cohesionless soils, exhibiting various geotechnical engineering properties (Fig. 1). The basic properties of each soil type, including sand–silt–clay composition, mean particle size (D_{50}), uniformity coefficient (C_u), curvature coefficient (C_c), liquid limit, and plastic index, were obtained from the literature. However, as the basic properties of soil are correlated to each other, appropriate feature selection becomes important, which governs the accuracy of machine learning–based behavior prediction.

For microbial-induced calcite precipitation (MICP), which is another bio-cementation technique, C_u and D_{50} are regarded as crucial variables in terms of strength enhancement (Konstantinou *et al.* 2023); thus, they were selected as features representing the soil type (Table 2).

Four types of biopolymers, i.e., xanthan gum (XG), gellan gum (GG), agar gum (AG), and beta-glucan (BG), were included in the configured database, whereas untreated

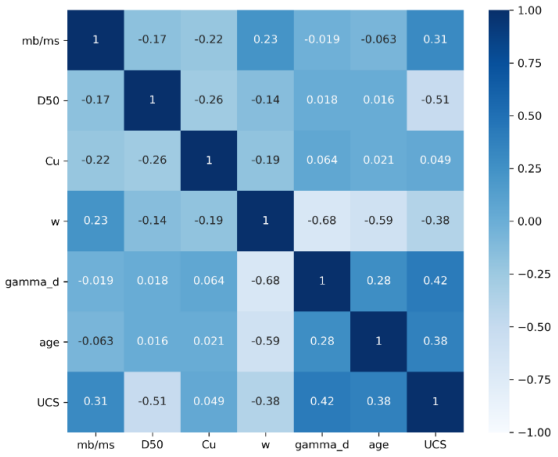

Fig. 2 Pearson correlation coefficients among selected features and UCS

Table 2 D_{50} and C_u of each soil type in the database

Soil sources	USCS	D50	Cu
Jumunjin	SP	0.595	1.56
Cheonan	SP-SM	0.309	3.06
Jaffna	SM	0.222	6.20
Ha-dong	MH	0.057	14.00
Gochang	CL	0.055	1.62
Yeosu	CH	0.019	13.73

Table 3 Selected features for the UCS prediction of BPST

No.	Feature	Abbreviation	Unit
1	BP type	XG/GG/AG/BG	-
2	Biopolymer content	m_b/m_s	%
3	Mean particle size	D_{50}	mm
4	Uniformity coefficient	C_u	-
5	Test water content	w	%
6	Dry density	ρ_d	g/cm^3
7	Curing time	Age	days

soil cases (nontreated, NT) were included for comparison. Meanwhile, the biopolymer type represents a piece of categorical data. For the purpose of building a machine learning model, this categorical data needs to be transformed into a numerical format. One-hot encoding is a standard technique for this purpose, which creates additional binary columns that indicate the presence of each value regardless of order and size (Potdar *et al.* 2017).

The UCS data collected from the literature were analyzed under three different moisture conditions: (1) initial (moist samples having the identical water content when the biopolymer–soil mixture was prepared), (2) dehydrated (biopolymer–soil mixture dried at room temperature (20ºC) for 1 to 63 days), and (3) saturated (submerged immediately after biopolymer–soil mixture preparation or re-submerged after dehydration) states. The mass and water content of all the specimens were recorded for the collected data.

Seven features were finally selected for data analysis: D_{50} and C_u (representing soil type), BP type and content (biopolymer to soil ratio in mass), water (moisture) content, dry density, and curing time (representing test environment). Details and statistical description of each feature are summarized in Tables 3 and 4.

Fig. 2 shows Pearson correlation coefficients between any two variables except the biopolymer type. The value of the Pearson correlation coefficient can be between -1 and +1. A value closer to 1 indicates a positive correlation, and a value closer to -1 indicates a negative correlation. A value of 0 indicates that there is no linear relationship (Cohen *et al.* 2009). The Pearson correlation coefficients among the seven main features were assessed to fit in the range between -0.7 and 0.7, indicating less correlation with each other.

The entire dataset (consisting of 479 UCS cases) was divided into two sections. The training set was split for 90% of the dataset (431 cases) and the testing set was split for 10% (48 cases). Finally, to avoid the magnitude difference between features (e.g., water content: 0.04-70.32%; dry density: 0.907-2.037 g/cm^3), all data were normalized between 0 and 1 before running a model using the following formula

$$X_n = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

2.2 Machine learning algorithms

The choice of machine learning algorithms should be appropriate with respect to the characteristics and quantity of data being analyzed (Ağbulut *et al.* 2021, Bobbo *et al.* 2021, Chen *et al.* 2019, Nikou *et al.* 2019). Neural networks (NNs) can predict complex functional relationships by identifying nonlinear hidden characteristics through multiple layers of analysis. However, as the NN model requires many parameters, its performance may be dependent on the volume and characteristics of data (Shao and Lunetta 2012). In cases with a limited number of datasets, simpler models, such as support vector regression (SVR) or random forest (RF), may exhibit higher predictive power than the NN. For baseline comparison, a linear regression (LR) algorithm is also employed. Therefore, this study examines the predictability of UCS and compares the performance of four different models: LR, RF, SVR, and NN.

LR is used as a benchmark for comparison with other models. The LR model can be trained using the least square method, which results in a deterministic model that closely fits the data. However, complex algorithms require hyperparameters to be set, which can significantly affect the performance of the model (Breuel 2015, Zhang *et al.* 2017).

Herein, the hyperparameters of each model using grid search and 5-fold cross-validation were optimized. For SVR, we used a radial basis kernel and set the penalty parameter (C), kernel parameter (γ), and allowable error (ε) as hyperparameters. For RF, the number of trees (N_T) and the maximum depth (D_T) of each tree were set as hyperparameters. For the NN, we optimized the learning

Table 4 Statistical descriptions of all the parameters except the biopolymer type category

Feature	Unit	Min	Max	Median	Mean	Standard deviation	Coefficient of variation	Skewness	Kurtosis
m_b/m_s	%	0.0	3.0	1.0	1.423	0.895	0.629	0.815	-0.699
D_{50}	mm	0.019	0.595	0.309	0.292	0.235	0.805	0.584	-1.523
C_u	-	1.56	14.00	1.63	2.737	2.935	1.072	0.452	-1.564
w	%	0.04	70.21	3.78	15.547	19.462	1.252	0.220	-1.918
ρ_d	g/cm^3	0.907	2.037	1.420	1.418	0.218	0.154	0.143	-1.571
Age	days	0	63	7	12.451	14.915	1.198	1.303	0.756
UCS	kPa	7.707	13231.636	350.520	1722.713	2665.293	1.547	-0.081	0.740

rate (α), batch size (N_B), number of layers (N_L), and number of nodes in each layer (N_N) as hyperparameters. Finally, four models were trained with training data using the optimized hyperparameters, and the performance of each model was compared based on the evaluation of the test data.

2.3 Evaluation criteria

To evaluate the performance of the models, the root mean squared error (RMSE) and coefficient of determination (R^2) were considered. These parameters are defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1,p=1}^{n}\left(y_t - y_p\right)^2} \tag{2}$$

$$R^2 = \left[\frac{\sum\left(y_t - \overline{y_t}\right)\left(y_p - \overline{y_p}\right)}{\sqrt{\sum\left(y_t - \overline{y_t}\right)^2}\sqrt{\sum\left(y_p - \overline{y_p}\right)^2}}\right]^2 \tag{3}$$

where n denotes the total number of data samples, y_t denotes the true value, $\overline{y_t}$ denotes the mean of the true values, y_p denotes the predicted values, and $\overline{y_p}$ denotes the mean of the predicted values of the model.

RMSE indicates the average distance between the predicted values of the model and the true values (Willmott and Matsuura 2005). The smaller the value, the better the model's performance. R^2 is a statistical measure of the data's proximity to the fitted regression line. When R^2 is near 1, the differences between the observations and predicted points are minor and unbiased (Cameron and Windmeijer 1997).

2.4 Permutation feature importance

Machine learning models have gained immense popularity in solving complex problems; however, most prediction models are regarded as black-box models, which makes it challenging to justify their predictions. To address this limitation, the field of explainable AI has been established to interpret the black box (Altmann *et al.* 2010, Molnar 2020). One approach to address this issue is the feature importance method, which examines the contribution of each feature to prediction. Herein, we investigated critical features that influence the prediction of UCS using the permutation feature importance (PFI) method. Unlike the Gini index in the tree-based model, this model-agnostic method can be employed in any algorithm. Thus, we applied PFI to the LR, RF, SVR, and NN models to analyze the importance of features in UCS prediction. Herein, we conducted 100 random permutations on all the features and measured the average increase in RMSE to identify crucial factors in UCS prediction for each model.

3. Results and discussion

3.1 Modelling by algorithms

Four models, including LR, RF, SVR, and NN, were trained to predict the UCS of BPST using five different soil sources with ten dimensions as inputs. However, while LR requires no additional parameters, RF, SVR, and NN require hyperparameters that must be defined by the user. To optimize these hyperparameters, grid search and 5-fold cross-validation methods were employed.

As illustrated in Fig. 3(a), for RF, the RMSE value decreased as D_T and N_T increased. The optimal performance was achieved when the D_T value was 22 and the N_T value was 21. For the NN, increasing the number of layers and nodes resulted in a higher number of model parameters, leading to overfitting. Therefore, in this case, where the number of data points was limited to a few hundred, a relatively small number of nodes and layers were employed.

The optimal hyperparameters were found to be N_L = 3, N_N = 55, N_b = 32, and α = 0.01 (Figs. 3(b)-(c)). For SVR, the lowest RMSE value was obtained at C = 100,000, γ = 1, and ε = 10 (Figs. 3(d)-3(f)). However, the RMSE values exhibited a sharp transition in response to the variation in hyperparameters. To ensure optimal performance, grid search was repeated for the hyperparameters of SVR: C = $10^4 \sim 10^6$, γ = $10^{-1} \sim 10^1$, and ε = $10^0 \sim 10^2$. Thus, the optimal hyperparameters were found to be C = 200,000, γ = 3, and ε = 1 (Figs. 3(g)-3(i)). In summary, these results demonstrated the importance of hyperparameter optimization in enhancing the performance of machine learning models for predicting UCS. The optimized hyperparameters obtained through grid search and cross-validation were used to retrain the four different models on the entire training data, and their performance was evaluated using the test data (Table 5).

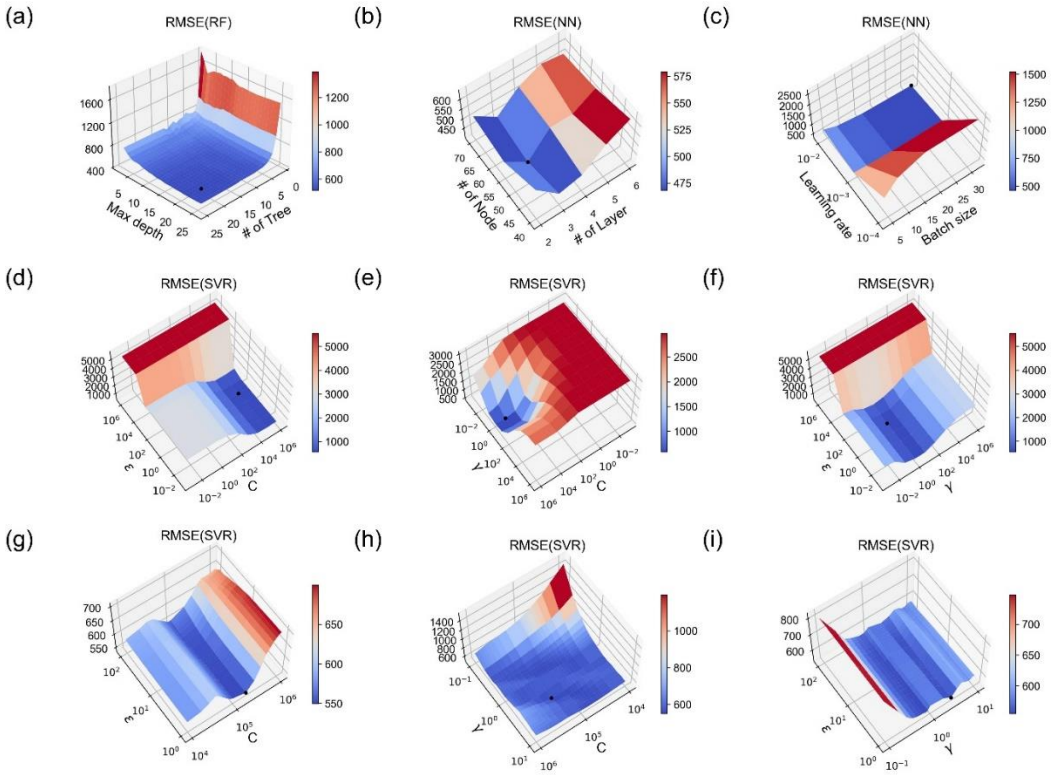

Fig. 3 Hyperparameter optimization results. RMSE values are plotted with respect to the hyperparameters with other hyperparameters fixed as the optimum values. The black dot indicates the lowest RMSE. (a) Random forest. (b) & (c) Neural network. (d)-(f) Support vector regression (coarse grid). (g)-(i) Support vector regression (fine grid)

Table 5 Comparison among the LR, RF, NN, and SVR models with optimized hyperparameters and prediction performance

Model	Hyperparameters	Training data		Test data	
		R^2	RMSE (kPa)	R^2	RMSE (kPa)
LR		0.493	1,536	0.496	1,622
RF	$D_T = 22$, $N_T = 21$	0.994	201	0.980	380
NN	$N_L = 3$, $N_N = 55$, $N_b = 32$, $\alpha = 0.01$	0.981	365	0.989	284
SVR	$C = 200{,}000$, $\gamma = 3$, $\varepsilon = 1$	0.987	300	0.990	273

The results of the test and training data prediction for each model are shown in Fig. 4, and their comparison is presented in Fig. 5. The linear regression model produced a low R^2 value of <0.5 and displayed relatively poor performance compared to other models. This suggests that the relationship between the UCS and input features is nonlinear, which is consistent with the low correlation coefficient observed in Fig. 2. In contrast, other algorithms successfully learned the nonlinear relationship between the features and UCS, displaying R^2 values close to 0.99 for the training and test data. Although SVR exhibited the highest prediction performance on the test data, differences among the three algorithms were minor. However, RF exhibited a higher RMSE value for the training data than that for the test data compared to other algorithms, suggesting some overfitting. Considering that an NN is more appropriate for larger datasets and requires a relatively longer training time, it can be concluded that SVR is the most efficient tool to solve problems.

3.2 Feature importance

Based on the performance evaluation of the four trained models, the importance of features in predicting UCS was further analyzed via PFI analysis. The results are illustrated in Fig. 6. The analysis revealed that some features are more important than others in predicting UCS. Herein, the reference values of RMSE and R^2 were set based on the model's performance on combined data from training and test sets, displayed as an orange horizontal line. By randomly permuting the sequence of each feature, the importance of each feature was identified based on the resulting performance degradation. The average importance of each feature was calculated through 100 independent runs and presented on the left based on importance.

As the PFI method is based on a trained model, it is closely related to the performance of the model. Hence, the PFI result of the LR model, which exhibited low performance, was deemed irrelevant. However, considering

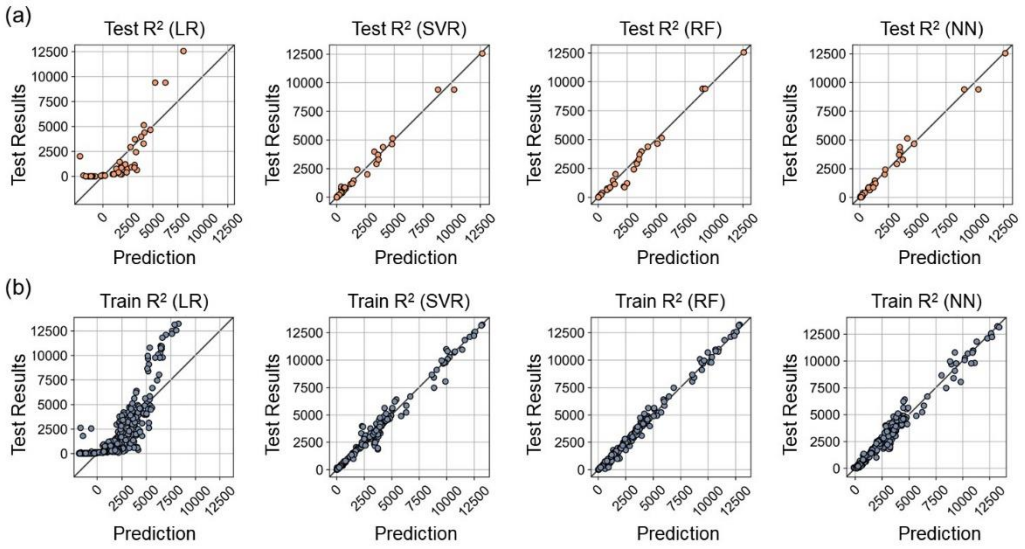

Fig. 4 R^2 values of UCS prediction with four different machine learning algorithms. (a) Test dataset and (b) Training dataset

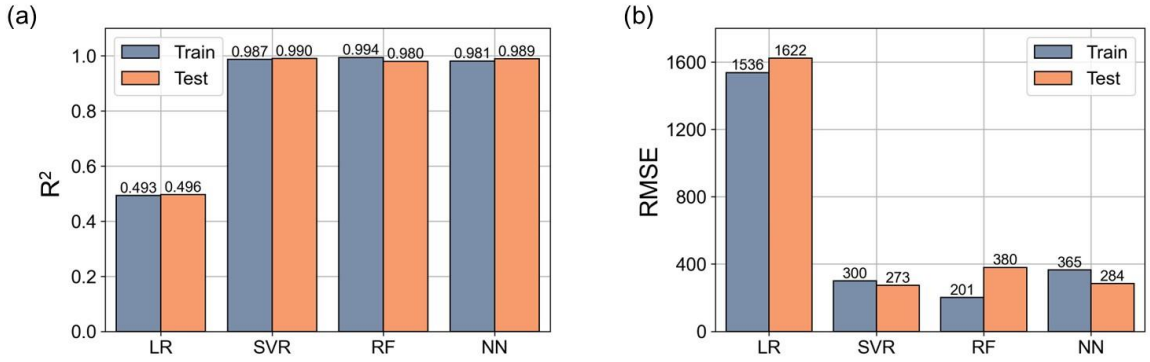

Fig. 5 Performance error comparison of each machine learning algorithm. (a) R^2 and (b) RMSE

the PFI of the three models with good predictive performance, D_{50} was found to be the most important feature. Additionally, m_b/m_s and w_{test} were identified as the main features, while age and C_u were found to be relatively less important than other features. The importance of the features varied slightly depending on the model and evaluation criteria used.

As the PFI method transforms the feature values of the data, the permutated data cannot be explored and are unrealistic. Moreover, importance is determined by how wrongly the imaginary data is predicted, which can be measured differently depending on the model's extrapolation ability. Therefore, it was concluded that an extreme quantitative comparison is not appropriate, and the three models with good predictive performance had common importance in the order of D_{50}, (m_b/m_s & w_{test}), (ρ_d & BP type), and (age & C_u).

3.3 Dominant variables of strength enhancement

According to PFI, the mean particle size (D_{50}) is the most dominant factor in predicting the UCS of BPST. As shown in Fig. 7, the high UCS values are distributed in the soil with D_{50} = 0.055 mm, which is classified as CL soil (Table 2). Chang and Cho (2019) found that BPST strengthening significantly depended on the biopolymer-to-clay ratio in mas (m_b/m_c), which was optimized at around m_b/m_c = 4% for gellan gum–kaolinite clay mixtures, regardless of the soil composition and water content. The direct hydrogen bonding between gellan gum monomers and clay particles generated gellan gum-clay matrices. Herein, the strength of the soils was increased by enhancing conglomeration, resulting in an increase in the friction angle of the soils (Chang and Cho 2019). In other words, the clay content predominantly affected the enhancement of BPST strength, and the four models well reflected this characteristic.

Following D_{50}, biopolymer content (m_b/m_s) and water content (w_{test}) were assessed as prominent variables in SVM, RF, and NN (the PFI result of LR is not discussed because its prediction accuracy is unreliable at R^2 = 0.49).

The UCS values of BPST tend to increase with an increase in the m_b/m_s due to biopolymer hydrogel condensation and subsequent biofilm formation (Cabalar *et al.* 2017, Chang *et al.* 2015a, c, Latifi *et al.* 2017,

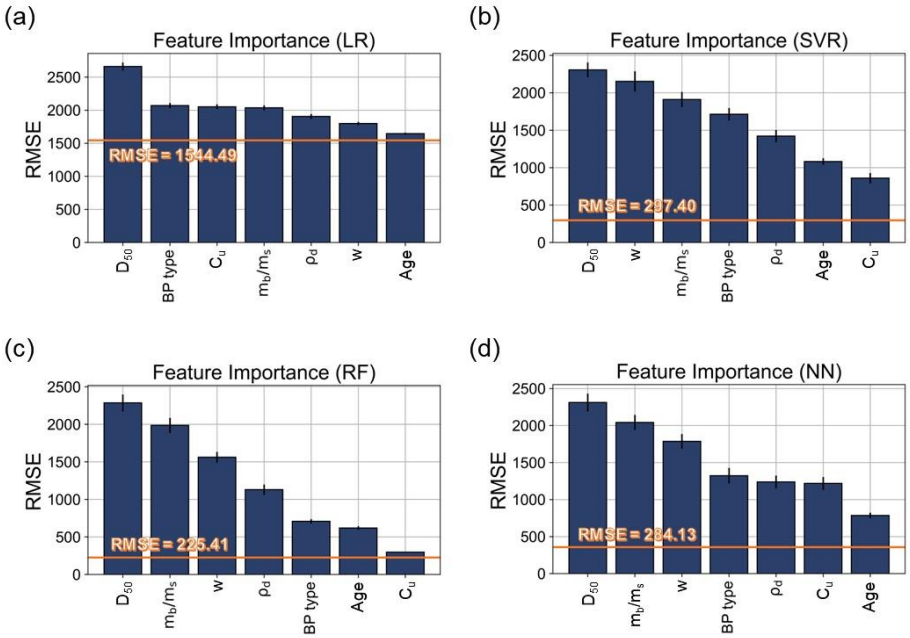

Fig. 6 Feature importance for predicting UCS is analyzed for four different models. Each bar indicates the decreased performance when the specific feature is replaced with the value of randomly permutated sequence. The yellow horizontal line represents the reference performance value that every feature normally uses (a) LR, (b) SVR, (c) RF and (d) NN

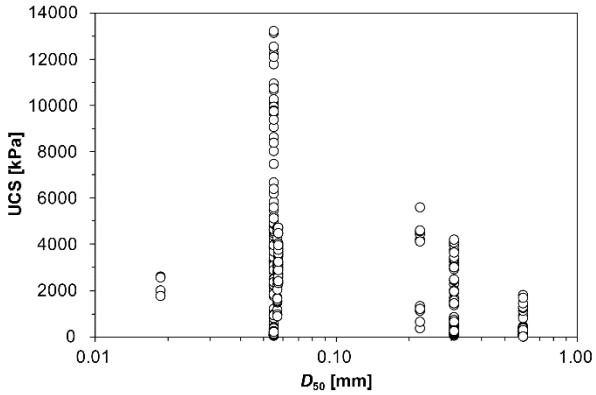

Fig. 7 UCS distribution according to the mean particle size (D_{50}) in the database

Wiszniewski *et al.* 2017). Higher biopolymer content is expected to form thicker and more complex intergranular biopolymer networks in coarse soils and facilitate more clay-biopolymer–clay interactions in fine soils.

The UCS, however, increases nonlinearly, and the margin of increase tends to decrease or level off at higher biopolymer concentrations (Chang *et al.* 2015a). As biopolymers are hydrophilic, water content plays an important role in BPST fabrication. The thickening characteristic of a biopolymer solution is related to its viscosity, which mainly depends on water content and is reduced by dilution owing to the hydrophilicity of biopolymers (García-Ochoa *et al.* 2000). Therefore, the hydrogels of BPST grow stronger and stiffer as they dehydrate, thereby increasing the inter-particle stress between soil particles (Chang *et al.* 2015d).

The PFI results were consistent with the findings of the previous BPST study. It was confirmed that the mean particle size (i.e., clay content) significantly impacted the UCS prediction of BPST. Therefore, it is believed that appropriate soil type selection, biopolymer content, and water content are crucial to attaining the desired strength of BPST.

Meanwhile, the most insensitive variables for strength enhancement were found to be age and C_u. Generally, cement-based materials harden through the hydration reaction of cement compounds, i.e., C_3S. Therefore, changes in strength over time have been critically evaluated (Bullard *et al.* 2011). On the other hand, BPST exhibited strength through the hydrogen and ion bonding between biopolymer particles and between biopolymer and soil particles (Chang *et al.* 2016). Therefore, the thickness of the hydrogel and the absolute water content significantly impacted strength enhancement. Moreover, C_u was evaluated as a less important variable. In recent MICP research, the influence of C_u on strength expression was investigated (Reed and Montoya 2023). However, this study revealed that in BPST, D_{50} was more critical than C_u.

4. Conclusions

This study focused on the construction of a prediction model to measure the UCS of BPST. The main objectives were to predict UCS and analyze the influence parameters of strength enhancement. Four different machine learning algorithms, viz., LR, SVR, RF, and NN, were proposed and compared. Based on verified and consistent high-quality

experimental data, the value of R^2 was found to be greater than 0.98 for the test and training datasets related to SVR, RF, and NN (except for LR). Since the ten features demonstrated a nonlinear relationship with the UCS, it was observed that the SVR, RF, and NN models, known for their effectiveness in nonlinear analysis, were apt for predicting the UCS of BPST. Nevertheless, the NN model required substantial time for optimization and was more appropriate for big data applications; hence, it offered poor cost-performance in analyzing the respective UCS data for BPST. While both RF and SVR offer a favorable balance of cost and performance, SVR was ultimately selected as the preferred model since RF has a propensity to overfit.

The PFI technique was used to derive the feature importance of the black box models. As the prediction performance of LR was significantly worse than other models with $R^2 = 0.493$ and RMSE = 1536, it was excluded from the feature importance evaluation. As a result, the mean particle size (D_{50}) in common among the three models caused the greatest variability in model prediction. This indicates that the UCS varies considerably depending on the type of soil in the biopolymer-mixed soil. As the biopolymer forms direct hydrogen bonds with clay, soils containing clay cause a large strength increase. This mechanism was evident in the SVR, RF, and NN models, and the results were revealed via PFI analysis. Following D_{50}, biopolymer content (m_b/m_s) and water content (w_{test}) were found to be important features for strength enhancement. These variables were mainly controlled in the existing BPST research field. Through PFI analysis, it was possible to quantitatively derive parameters that affect the strength enhancement of BPST, which is expected to provide insight into future BPST laboratory tests and field applications.

Herein, it was confirmed that the strength of biopolymer, a biological material with uncertainty, could be predicted using machine learning models (SVR, RF, and NN). Specifically, given the size and characteristics of the available dataset, the SVR model was identified as the most appropriate for predicting the strength of biopolymers. If this SVR prediction model is used, strength can be predicted without experimentation in the design stage and the mixing ratio can be designed accordingly for the target strength. We can save time and money when applying biopolymers for soil improvement and field stabilization.

Acknowledgments

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2C2091517).

References

Abraham, A. (2005), "Artificial neural networks", *Handbook of Measuring System Design.*

Ağbulut, Ü,, Gürel, A.E. and Biçen, Y. (2021), "Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison", *Renew. Sustain. Energ. Rev.*, **135**, 110114. https://doi.org/10.1016/j.rser.2020.110114.

Almajed, A., Lateef, M.A., Moghal, A.A.B. and Lemboye, K. (2021), "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*, **11**(4), 370. https://doi.org/10.3390/cryst11040370.

Altmann, A., Toloşi, L., Sander, O. and Lengauer, T. (2010), "Permutation importance: a corrected feature importance measure", *Bioinformatics*, **26**(10), 1340-1347. https://doi.org/10.1093/bioinformatics/btq134.

Benayoun, F., Boumezerane, D., Bekkouche, S.R.and Bendada, L. (2020), "Application of genetic algorithm method for soil nailing parameters optimization", *Proceedings of the IOP Conference Series: Materials Science and Engineering.*

Bharti, S.N. and Swetha, G. (2016), "Need for bioplastics and role of biopolymer PHB: a short review", *J. Pet. Environ. Biotechnol.*, **7**(272), 2. https://doi.org/10.4172/2157-7463.1000272.

Bhattacharya, B. and Solomatine, D.P. (2006), "Machine learning in soil classification", *Neural Networks*, **19**(2), 186-195. https://doi.org/10.1016/j.neunet.2006.01.005.

Bobbo, T., Biffani, S., Taccioli, C., Penasa, M. and Cassandro, M. (2021), "Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows", *Scientific Reports,* **11**(1), 1-10. https://doi.org/10.1038/s41598-021-93056-4.

Breiman, L. (2001), "Random forests", *Mach.Learning,* **45**, 5-32. https://doi.org/10.1023/A:1010933404324.

Breuel, T.M. (2015), "The effects of hyperparameters on SGD training of neural networks", *arXiv Preprint arXiv:1508.02788.*

Bullard, J.W., Jennings, H.M., Livingston, R.A., Nonat, A., Scherer, G.W., Schweitzer, J.S., Scrivener, K.L. and Thomas, J.J. (2011), "Mechanisms of cement hydration", *Cem.Concr.Res.*, **41**(12), 1208-1223. https://doi.org/10.1016/j.cemconres.2010.09.011.

Cabalar, A.F., Wiszniewski, M. and Skutnik, Z. (2017), "Effects of xanthan gum biopolymer on the permeability, odometer, unconfined compressive and triaxial shear behavior of a sand", *Soil Mech. Found. Eng.*, **54**(5), 356-361. https://doi.org/10.1007/s11204-017-9481-1.

Cameron, A.C. and Windmeijer, F.A. (1997), "An R-squared measure of goodness of fit for some common nonlinear regression models", *J. Econ.*, **77**(2), 329-342. https://doi.org/10.1016/S0304-4076(96)01818-0.

Chang, I. and Cho, G. (2012), "Strengthening of Korean residual soil with -l,3/1,6-glucan biopolymer", *Constr. Build. Mater.*, **30**(1), 30. https://doi.org/10.1016/j.conbuildmat.2011.11.030.

Chang, I., Im, J., Prasidhi, A.K. and Cho, G. (2015a), "Effects of Xanthan gum biopolymer on soil strengthening", *Constr. Build. Mater.*, **74**, 65-72. https://doi.org/10.1016/j.conbuildmat.2014.10.026.

Chang, I., Jeon, M. and Cho, G. (2015b), "Application of microbial biopolymers as an alternative construction binder for earth buildings in underdeveloped countries", *Int. J. Polymer Sci.*, https://doi.org/10.1155/2015/326745.

Chang, I., Prasidhi, A.K., Im, J. and Cho, G. (2015c), "Soil strengthening using thermo-gelation biopolymers", *Constr. Build. Mater.*, **77**, 430-438. https://doi.org/10.1016/j.conbuildmat.2014.12.116.

Chang, I., Prasidhi, A.K., Im, J., Shin, H. and Cho, G. (2015d), "Soil treatment using microbial biopolymers for anti-desertification purposes", *Geoderma*, **253-254**, 39-47. https://doi.org/10.1016/j.geoderma.2015.04.006.

Chang, I., Im, J. and Cho, G. (2016), "Introduction of microbial biopolymers in soil treatment for future environmentally-

friendly and sustainable geotechnical engineering", *Sustainability*, **8**(3), 251. https://doi.org/10.3390/su8030251.

Chang, I., Im, J., Lee, S. and Cho, G. (2017), "Strength durability of gellan gum biopolymer-treated Korean sand with cyclic wetting and drying", *Constr. Build. Mater*., **143**, 210-221. https://doi.org/10.1016/j.conbuildmat.2017.02.061.

Chang, I. and Cho, G. (2019), "Shear strength behavior and parameters of microbial gellan gum-treated soils: From sand to clay", *Acta Geotechnica*, **14**(2), 361-375. https://doi.org/10.1007/s11440-018-0641-x.

Chang, I., Lee, M., Tran, A.T.P., Lee, S., Kwon, Y., Im, J. and Cho, G. (2020), "Review on biopolymer-based soil treatment (BPST) technology in geotechnical engineering practices", *Transport. Geotech.*, **24**, 100385. https://doi.org/10.1016/j.trgeo.2020.100385.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., Van Donkelaar, A., Hvidtfeldt, U.A. and Katsouyanni, K. (2019), "A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide", *Environ. Int.*, **130**, 104934. https://doi.org/10.1016/j.envint.2019.104934.

Choi, S., Chang, I., Lee, M., Lee, J., Han, J. and Kwon, T. (2020), "Review on geotechnical engineering properties of sands treated by microbially induced calcium carbonate precipitation (MICP) and biopolymers", *Constr. Build. Mater*., **246**, 118415. https://doi.org/10.1016/j.conbuildmat.2020.118415.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y. and Cohen, I. (2009), "Pearson correlation coefficient", *Noise Reduction in Speech Processing*, 1-4. https://doi.org/10.1007/978-3-642-00296-0_5.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks", *Mach. Learning*, **20**(3), 273-297. https://doi.org/10.1007/BF00994018.

Drucker, H., Burges, C.J., Kaufman, L., Smola, A. and Vapnik, V. (1996), "Support vector regression machines", *Adv. Neural Inform. Process. Syst.*, **9**.

Fatehi, H., Ong, D.E., Yu, J. and Chang, I. (2021), "Biopolymers as green binders for soil improvement in geotechnical applications: A review", *Geosciences*, **11**(7), 291. https://doi.org/10.3390/geosciences11070291.

Fatehi, H., Ong, D.E., Yu, J. and Chang, I. (2023), "The effects of particle size distribution and moisture variation on mechanical strength of biopolymer-treated soil", *Polymers*, **15**(6), 1549. https://doi.org/10.3390/polym15061549.

García-Ochoa, F., Santos, V.E., Casas, J.A. and Gómez, E. (2000), "Xanthan gum: production, recovery, and properties", *Biotechnol. Adv*., **18**(7), 549-579. https://doi.org/10.1016/S0734-9750(00)00050-1.

Goh, A.T. and Goh, S.H. (2007), "Support vector machines: their use in geotechnical engineering as illustrated using seismic liquefaction data", *Comput. Geotech*., **34**(5), 410-421. https://doi.org/10.1016/j.compgeo.2007.06.001.

Gonos, I.F. and Stathopulos, I.A. (2005), "Estimation of multilayer soil parameters using genetic algorithms", *IEEE Trans.Power Del.*, **20**(1), 100-106. https://doi.org/10.1109/TPWRD.2004.836833.

Kim, M. and Gilley, J.E. (2008), "Artificial Neural Network estimation of soil erosion and nutrient concentrations in runoff from land application areas", *Comput. Electron. Agric*., **64**(2), 268-275. https://doi.org/10.1016/j.compag.2008.05.021.

Kiran, S., Lal, B. and Tripathy, S.S. (2016), "Shear strength prediction of soil based on probabilistic neural network", *Indian J. Sci. Technol*., **9**(41), 1-6. https://doi.org/10.17485/ijst/2016/v9i41/124740.

Konstantinou, C., Wang, Y. and Biscontin, G. (2023), "A systematic study on the influence of grain characteristics on hydraulic and mechanical performance of MICP-treated porous media", *Transport. Porous Media*, **147**(2), 305-330. https://doi.org/10.1007/s11242-023-01909-5.

Kwon, Y., Chang, I., Lee, M. and Cho, G. (2019), "Geotechnical engineering behavior of biopolymer-treated soft marine soil", *Geomech. Eng*., **17**(5), 453-464. https://doi.org/10.12989/gae.2019.17.5.453.

Kwon, Y., Moon, J., Cho, G., Kim, Y. and Chang, I. (2023), "Xanthan gum biopolymer-based soil treatment as a construction material to mitigate internal erosion of earthen embankment: A field-scale", *Constr. Build. Mater.*, **389**, 131716. https://doi.org/10.1016/j.conbuildmat.2023.131716.

Latifi, N., Horpibulsuk, S., Meehan, C.L., Abd Majid, M.Z., Tahir, M.M. and Mohamad, E.T. (2017), "Improvement of problematic soils with biopolymer—an environmentally friendly soil stabilizer", *J. Mater. Civ. Eng*., **29**(2), 04016204. https://doi.org/10.1061/(ASCE)MT.1943-5533.0001706.

Lee, S., Chung, M., Park, H.M., Song, K. and Chang, I. (2019), "Xanthan Gum Biopolymer as Soil-Stabilization Binder for Road Construction Using Local Soil in Sri Lanka", *J. Mater. Civ. Eng*., **31**(11), 06019012. https://doi.org/10.1061/(ASCE)MT.1943-5533.0002909.

Ma, J., Xia, D., Guo, H., Wang, Y., Niu, X., Liu, Z. and Jiang, S. (2022), "Metaheuristic-based support vector regression for landslide displacement prediction: A comparative study", *Landslides*, **19**(10), 2489-2511. https://doi.org/10.1007/s10346-022-01923-6.

Mekonnen, E., Amdie, Y., Etefa, H., Tefera, N. and Tafesse, M. (2022), "Stabilization of expansive black cotton soil using bioenzymes produced by ureolytic bacteria", *Int. J. Geo-Eng*., **13**(1), 10. https://doi.org/10.1186/s40703-022-00175-6.

Molnar, C. (2020), *Interpretable machine learning*, Lulu. com.

Nikou, M., Mansourfar, G. and Bagherzadeh, J. (2019), "Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms", *Intelligent Systems in Accounting, Finance and Management*, **26**(4), 164-174.

Njock, P.G.A., Shen, S., Zhou, A. and Lyu, H. (2020), "Evaluation of soil liquefaction using AI technology incorporating a coupled ENN/t-SNE model", *Soil Dyn. Earthq. Eng.*, **130**, 105988.

Pham, B.T., Hoang, T., Nguyen, D. and Bui, D.T. (2018), "Prediction of shear strength of soft soil using machine learning methods", *Catena*, **166**, 181-191.

Potdar, K., Pardawala, T.S. and Pai, C.D. (2017), "A comparative study of categorical variable encoding techniques for neural network classifiers", *Int. J. Comput. Appl.*, **175**(4), 7-9.

Reed, M. and Montoya, B.M. (2023), "Influence of the coefficient of uniformity on bio-cemented sands: a microscale investigation", *Proceedings of the 8th International Symposium on DEFORMATION CHARACTERISTICS OF GEOMATERIALS, .*

Seo, S., Lee, M., Im, J., Kwon, Y., Chung, M., Cho, G. and Chang, I. (2021), "Site application of biopolymer-based soil treatment (BPST) for slope surface protection: in-situ wet-spraying method and strengthening effect verification", *Constr. Build. Mater*., **307**, 124983. https://doi.org/10.1016/j.conbuildmat.2021.124983.

Shao, Y. and Lunetta, R.S. (2012), "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points", *ISPRS J. Photogramm. Remote Sens*., **70**, 78-87. https://doi.org/10.1016/j.isprsjprs.2012.04.001.

Willmott, C.J. and Matsuura, K. (2005), "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", *Climate Res.*, **30**(1), 79-82.

Wiszniewski, M., Skutnik, Z., Biliniak, M. and Cabalar, A.F.

(2017), "Some geomechanical properties of a biopolymer treated medium sand", *Annals of Warsaw University of Life Sciences-SGGW Land Reclamation*, **49**(3), 201-212.

Worrell, E., Price, L., Martin, N., Hendriks, C. and Meida, L.O. (2001), "Carbon dioxide emissions from the global cement industry", *Annu. Rev. Energ. Environ.*, **26**(1), 303-329.

Zhang, X., Yao, L., Huang, C., Sheng, Q.Z.and Wang, X. (2017), "Intent recognition in smart living through deep recurrent neural networks", *Proceedings of the Neural Information Processing: 24th International Conference, ICONIP 2017*, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24.

IC