Smart Structures and Systems, *Vol. 28, No. 2 (2021) 289-304*
https://doi.org/10.12989/sss.2021.28.2.289

Machine learning-based prediction and performance study of transparent soil properties

Bo Wang*[1,2], Hengjun Hou [1,2], Zhengwei Zhu**[1,2] and Wang Xiao [3]

[1] *School of Civil Engineering, Chongqing University, No. 83 Shabei Street, Shapingba District, Chongqing 400045, P.R. China*
[2] *Key Laboratory of New Technology for Construction of Cities in Mountain Area (Chongqing University), Ministry of Education, Chongqing 400045, P.R. China*
[3] *Shaoguan Construction Quality and Safety Center, No. 1, Wuzu Road, Wujiang District, Shaoguan City, Guangdong Province 512026, P.R. China*

(*Received November 7, 2020, Revised February 24, 2021, Accepted March 24, 2021*)

Abstract. An indispensable process of geotechnical modeling with transparent soils involves analyzing images and soil property simulations. This study proposes an objective framework for quantitative analysis of the influential mechanism of three key factors, namely, different aggregate proportions (DAP), solvent ratio (SR), and solute solution ratio (SSR) on transparent soils' transparency and shear strength. 125 groups of transparent soil samples considering these three factors were prepared to investigate their impact on transparency and shear strength through Elastic Net regression. Spearman correlation analysis was performed for transparency and shear strength. Furthermore, by comparing the performance of XGBoost, GBDT, Random Forest, and SVR after hyperparameter tuning in predicting transparency and shear strength, XGBoost proved to be the optimal machine learning model with the lowest MSE of 0.0048 and 0.0306 and was innovatively adopted to analyze how various factors affect the transparency and shear strength, thus enhancing the interpretability of machine learning. A ranking system, according to the importance scores of XGBoost, shows that SSR was the most important factor affecting both shear strength and transparency of transparent soils, with importance scores being 0.45 and 0.57, respectively. Our study may shed light on the preparation and performance study of transparent soils.

Keywords: transparent soil; properties prediction; transparency; shear strength; machine learning

1. Introduction

Transparent soils are used as a new material in the geotechnical model test (Iskander *et al*. 2002). Many scholars have applied it to model research such as compression deformation, soil element movement, shear strain, surface uplift, slope stability, etc. (Xu *et al*. 2020, Ye *et al*. 2020). However, existing research shows that transparency is one of the key factors affecting the success of model tests. Previously, many scholars described transparency qualitatively. For example, Iskander (2010) depicted the level of transparency by observing grid lines outside the transparent soil glass box with a width of 50 mm. Ezzein and Bathurst (2011) used Abbes refractometer to match the refractive index of pore-fluid with the refractive index of fused silica aggregate and did not carry out an optical evaluation on the mixed transparent soil samples. Ni *et al*. (2010) poured the prepared transparent soil into a Plexiglass box with dimensions of 110 × 35 × 110 mm and made a random digital test card with Times New Roman style. The font sizes ranged from 2 pt - 40 pt. The transparency of the soil was determined by observing the minimum readable font size recorded on the test card with naked eyes. Stanier (2012) used a letter-based test card to visually evaluate transparent soil samples. These methods are different in operation, and the light and temperature vary among batches prepared each time. Besides, the results are susceptible to visual errors and the subjectivity of testers. In other words, although visual observation is an intuitive and straightforward method to make a clear comparison, selection, and judgment of transparent soil samples, it is easy to be influenced by subjective factors such as interpretation and vision. Moreover, the achievable resolution is not sufficient, resulting in uneven observation results. By this token, introducing the concept of transparency to quantitatively describe transparent soils is necessary. However, studies on the soil's transparency are rare, while quantitative analysis and accurate predictions of factors influencing it are even less.

Shear strength is the fundamental element for transparent soils to simulate natural ones (Zhao *et al*. 2010). The geotechnical properties of transparent soils prepared with different types of aggregates (pore liquid, particle size, etc.) vary greatly. Besides, the material's properties can be influenced by factors including the particle ratio, solution ratio, solute ratio, environment, temperature, and operator's experiences. Lo *et al*. (2010) used transparent soils made of water soluble beads to study the seepage, but the strength of

*Corresponding author, Ph.D.,
E-mail: bowang1313@outlook.com
**Co-corresponding author, Research Professor,
E-mail: zhuzw@cqu.edu.cn

ISSN: 1738-1584 (Print), 1738-1991 (Online)

the soils made is approximately zero, and its material properties differ substantially from natural soil. Subsequently, Ezzein and Bathurst (2011) discussed the basic physical and mechanical properties of transparent soils but their research lacks discussions on factors influencing these characteristics specifically. Therefore, it is necessary to establish a sound mathematical model with high prediction accuracy to analyze the influence of various factors on the shear strength of transparent soils. The study provides a reference for transparent soils preparation and lay the foundation for transparent soil model tests.

Recent advances in computational and apparatus have provided state of the art technology for complicated objectives (Ma and Yang 2015, Ma *et al*. 2019, Ding *et al*. 2020, Li *et al*. 2020a, Ma and Xu 2020). For example, in civil engineering, many studies have employed novel methodologies for geotechnical (Yang *et al*. 2015, Liu *et al*. 2018, Zhang *et al*. 2019a) and structural (Ju *et al*. 2020, Sun *et al*. 2020, Zhang *et al*. 2020a, b, Zhang and Wang 2020) investigations. Some of them are used specifically for analyzing the response (Zhang *et al*. 2020, Abedini and Zhang 2021, Zhang *et al*. 2021) and health monitoring (Fu and Yang 2020, Li *et al*. 2020b, Ma and Xu 2020, Yan *et al*. 2020). Another significant usage of these techniques could be optimization of an existing problem (Chen *et al*. 2019, Ma *et al*. 2020, Moayedi *et al*. 2020).

Machine learning is a widely used notion of these advanced techniques that has been extensively used (He *et al*. 2020, Lv and Qiao 2020, Yang *et al*. 2020b, Yue *et al*. 2020). For example, studies regarding material performance prediction have benefited from various intelligent tools (Chahnasir *et al*. 2018, Pourghasemi and Rahmati 2018, Bui *et al*. 2019, Liang *et al*. 2020). Deep learning models are a more sophisticated type of machine learning which have shown high competency for engineering modeling (Xu *et al*. 2018, Li *et al*. 2019, Lv and Qiao 2020, Yang *et al*. 2020a).

More specifically for our case, Samui (2008) used Support Vector Regression (SVR) to predict the friction resistance of pile driving into soils. Both Kuo *et al*. (2009) and Padmini *et al*. (2008) used Artificial Neural Networks (ANN) to predict the bearing capacity of shallow foundations. Chou *et al*. (2016) used data mining methods such as linear regression, classification and regression tree (CART) and used generalize linear (GENLIN) model, chi-squared automatic interaction detector (CHAID), ANN, and SVR to predict factors affecting the shear strength and the friction angle of fiber-reinforced soil (FRS). Das *et al*. (2011) used Support Vector Machine (SVM) and ANN to study the residual strength of soil. Kanungo *et al*. (2014) compared the predictive shear strength parameters of the ANN and CART methods. Kiran *et al*. (2016) used Probabilistic Neural Network (PNN) to predict soil shear strength parameters. These studies fully demonstrate the effectiveness of machine learning method in soil parameter prediction. But at present, research on the analysis and prediction of influencing factors of transparent soil parameters remains insufficient, and machine learning method has not been involved in relevant application research on transparent soils (Moavenian *et al*. 2016). As well as machine learning models, optimization techniques (e.g., whale optimization algorithm (Gadekallu *et al*. 2020), antlion optimization (Reddy *et al*. 2020b), genetic algorithm (Reddy *et al*. 2020a), firefly algorithm (Bhattacharya *et al*. 2020)) have played auxiliary roles in dealing with various problems.

Compared with traditional prediction methods, machine learning method can better deal with high dimensional and non-linear problems (Nehdi *et al*. 2006, Mohammadhassani *et al*. 2014). It is widely applied because it can effectively utilize big data to exploit information and improve high prediction accuracy (Ye *et al*. 2019). However, due to the 'black box' nature of machine learning prediction, using model training solely cannot quantify and explain how influencing factors affect the transparency and shear strength of transparent soils. Therefore, this paper adopted the Elastic Net regression and Spearman correlation analysis together with machine learning to study the influence as well as the correlation between transparency and shear strength. Four well-known notions, namely eXtreme Gradient Boosting (XGBoost), Gradient Boosting Decision Tree (GBDT), Random Forest, and SVR are selected, due to their excellent performance in many complicated engineering issues (Zhang *et al*. 2019b, Pham *et al*. 2020, Rong *et al*. 2020, Sharma *et al*. 2021). Moreover, to the best knowledge of the authors, these models have not been previously applied to this specific problem. Thus, employing and comparing them is of interest for obtaining an efficient and applicable evaluative tool.

2. Materials and methods

2.1 Setting of test parameters

2.1.1 Transparency

Transparency is one of the key problems in visual model tests. According to the Lambert-Beer law, transparency in this article is described as the following: as a beam of monochromatic light passes through a medium with a certain thickness, the intensity of transmitted light will be weakened because the medium absorbs part of light energy. The greater the concentration of the absorbing medium and the thicker the medium, the more obvious the attenuation of the light intensity is. The relation is

$$\mathrm{A} = lg_{10}\frac{I_0}{I_t} = lg_{10}\frac{1}{T} = K * l * c, \tag{1}$$

where A is the absorbance; I_0 and I_t represent the incident light intensity and the transmitted light intensity, respectively. T is the transmission ratio; K, l, c refer to the coefficient, the thickness of the absorption medium (cm), the concentration of the light-absorbing substance (g/l), respectively.

Transparency is collected by putting transparent soils into a cuvette with external dimensions of 12.5 × 12.5 × 45 mm, and an inner diameter of 10 mm. Then, it is put into UV-2100 ultraviolet-visible spectrophotometer to collect

data. The wavelength used in this experiment is 800 nm.

2.1.2 Shear strength

Simulating natural soils based on transparent soils can effectively and accurately study soil's internal problems. One of the key problems for simulation is soil properties similarity between the two kinds of soils. Shear strength serves as the reference basis for the simulated soils to have similar properties to natural ones. The direct shear test is commonly used to test soil samples and to measure the soil shear strength index. Previous studies on the compressive properties of transparent soils have been carried out with different preparation methods. This paper mainly studies the influence of DAP, SR, and SSR on the shear strength of transparent soils.

(a) Different aggregate proportions

The transparency and strength of transparent soil samples vary under different states due to different proportions of aggregates with various particle sizes, thus the costs are different. Aggregates with different particle sizes should be matched to ensure transparency and strength and economic factors should be considered to save cost as much as possible. Therefore, the solute ratio is taken as one of the influencing factors in the preparation of transparent soil in this study. In this paper, silica powders with particle sizes of 15 nm and 30 nm respectively, and purity of 99.8% is selected and mixed according to the proportion shown in Table 1.

(b) Solvent ratio

When preparing transparent soils, two solutions are usually mixed in a certain ratio. Besides, the solvents must be colorless, tasteless, safe, and stable, and they need to have similar physical properties, stable chemical properties, good transparency, and the same or similar refractive index as the selected aggregate. According to the ratio experiment of white oil and tridecane, we obtained the curve of relations between refractive index and solution ratio, as shown in Fig. 1. The refractive indexes in previous experiments are generally between 1.442 and 1.450. We calculated the ratio of white oil to tridecane, as shown in Table 2.

The solvent selected should be matched with the refractive index range of the solute. According to factors including solid particle refractive index (less than 1.450), solubility (greater than 60 g), molecular weight (greater than 90), innocuity, low toxicity, and low price, etc., the two chemical substances of tridecane and white oil were finally selected as solutions (Table 3 Technical indexes).

Therefore, both the influence of solute and the proportion of pore fluid solution should be considered during the preparation process. The refractive index of the solution is not significantly related to that of the solute's solid particles. Instead, it is related to the solubility, concentration (Ezzein and Bathurst 2014).

(c) Solute solution ratio

In this paper, the SSR is calculated by Eq. (2), which is expressed as

$$\begin{aligned} \mathrm{S} &= \frac{m_{30}+m_{15}}{m_s}\times 100\% \\ &= \frac{m_{30}+m_{15}}{m_{30}+m_{15}+m_w+m_t}\times 100\% \end{aligned} \tag{2}$$

S is the solute-solution ratio, $m_{30,}$ m_{15} refer to the mass of 30 nm and 15 nm silica powders respectively,

Table 1 Proportion table of silicon powder with different particle sizes

Serial number	Mesh	Mass ratio		15 nm/30 nm ratio	Price ($/100g)
		15 nm	30 nm		
1	Percentage	50%	50%	1	110
2	Percentage	60%	40%	1.5	114
3	Percentage	40%	60%	0.67	106
4	Percentage	70%	30%	2.33	118
5	Percentage	30%	70%	0.43	102

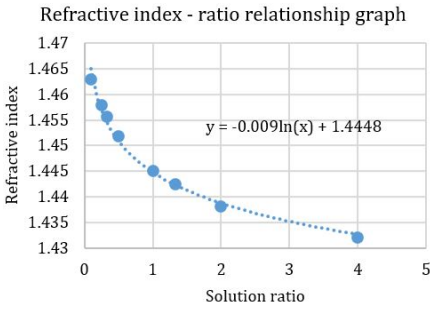


Fig. 1 Graph of relationship between refractive index and solution ratio

Table 2 Proportion table on refractive index calculation solution

Serial number	Refractive index	Solvent ratio (white oil: tridecane)
1	1.442	0.72
2	1.444	0.91
3	1.446	1.15
4	1.448	1.44
5	1.450	1.82

Table 3 Technical Indexes of White Oil and Tridecane

Name	Tridecane	White oil
Molecular weight	184.36	98.15
Molecular formula	C13H28	C8H16C
Boiling point	110~112℃12 m Hg	92.8℃
Density	0.756 /ml (25℃)	3.439 kg/m³
Storage conditions	30℃	Normal temperature
Color	Colorless	Colorless
Manufacturer	Dongguan Walixi Chemical Co., Ltd.	Shanghai Eastern Biz Co., Ltd

Table 4 Setting of various factors

Factor level	A solute ratio	B pore-fluid ratio	C solute solution ratio
Level 1	1	0.72	5%
Level 2	1.5	0.91	6%
Level 3	0.67	1.15	7%
Level 4	2.33	1.44	8%
Level 5	0.43	1.82	9%

m_w and m_t represent the mass of white oil and tridecane, respectively.

It is found that when the solute ratio changes from 5% to 9%, the transparent soil samples undergo state transitions from liquid to semi-solid and finally to plastic. When the ratio reaches 10%, the transparency of the transparent soil samples is poor by naked-eye observation. Therefore, the SSR adopted in this experiment are 5%, 6%, 7%, 8% and 9%.

2.2 Test plan

2.2.1 Transparency

DAP, SR, SSR, and the modes of three-factors and five-levels of the test scheme are shown in Table 4. According to the three factors in Table 4, and there are a total of 125 groups of experiments.

2.2.2 Test precautions

Insufficient attention to details when making transparent soil samples can result in a loss of overall transparency. Therefore, special attention should be paid to the following matters during preparation:

(a) Before the sample is prepared, spray a small amount of water into the room, and clean air-dry beakers, straws, and other instruments with water and detergent to prevent dust and water from affecting transparency.

(b) The samples should be stirred constantly to prevent uneven mixing or segregation of materials. Before vacuuming, the transparent material must be stirred again before each pour.

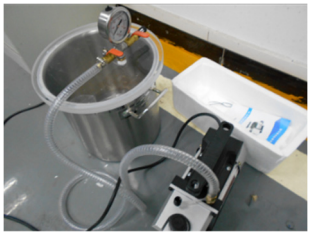
(c) When the vacuum pump and vacuum box are used for air extraction, the sample can be continuously pumped, exhausted, and slightly shaken, which is essential for the bubble to be discharged easily.

(d) When filling the sample into the cuvette, a large number of bubbles are easily generated, and the gas should be continuously shaken slightly or sucked out by the needle tube slowly.

2.3 Test process

The test in this paper includes the following steps: material ratio, extraction of vacuum, transparency test, direct shear test, and other steps to determine the transparency and shear strength of transparent soils. The precision of the electronic scale used for preparation was 0.01 g (maximum weight of 600 g). The direct shear test utilized ZJ strain control direct shear instrument with a 61.7 × 20 mm cutting ring to cut transparent soil samples. According to the operation procedures specified in the “Geotechnical Test Regulations”, four groups of consolidated quick direct shear strength tests with different stress states were carried out on transparent soils with pre-consolidation pressure ranging from 0 to 12.5kpa, of which four samples were taken from each group of stress states. Vertical pressure was applied step by step under the pressure of 1, 2, 3, and 4 kPa respectively. Then, the fixing pin was pulled out and the shearing was conducted, with the shearing rate kept at 0.8 mm/min. Generally, the shear loss of the sample was completed within 3~5 minutes. The transparency test adopted UV-2100 ultraviolet-visible spectrophotometer manufactured. The process is shown in Fig. 2.

(a) Samples made



(b) Vacuum equipment

(c) Cuvette sample

(d) Direct shear test

Fig. 2 Preparation and test drawing of transparent soil

3. Theoretical overview

3.1 Elastic net regression theory

In 2005, Zou and Hastie (2005) combined the ridge regression with the Lasso method to propose an elastic constraint estimation, Elastic Net. The parameters of the elastic constraint are estimated by

$$\tilde{\beta} = arg_{\beta} \min\left(\sum_{i}^{n}\left(y_i - \sum_{j=1}^{P}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{P}|\beta_j| + \lambda_2\sum_{j=1}^{P}{\beta_j}^2\right) \quad (3)$$

Equivalents to finding

$$\sum_{i}^{n}\left(y_i - \sum_{j=1}^{P}\beta_j x_{ij}\right)^2 \quad (4)$$

Satisfies the current demand

$$(1-\lambda)\sum_{j=1}^{P}|\beta_j| + \lambda\sum_{j=1}^{P}{\beta_j}^2 \le t \quad (5)$$

reaching the smallest of $\beta_{j,j=1,2,...,p}$. It is easy to see that when λ = 1, the elastic constraint estimation is the Ridge Regression; when λ = 0, the elastic constraint estimation is the Absolute Constraint Estimate. Therefore, the estimation takes into account the characteristics of both absolute constraint estimation and Ridge estimation.

3.2 Spearman correlation analysis

As a common statistical method for correlation analysis in civil engineering, the Spearman method is used to measure the strength of a linear relationship between two or more variables. The method's basic idea is to calculate the Spearman correlation coefficient based on the difference of each pairwise rank of two pairs of columns (Melucci 2009).

As a non-parametric correlation coefficient, spearman's correlation coefficient does not require the data source to meet strict assumptions such as normal distribution, linear constraint, and homoskedasticity, thus it is more widely applied (Delicado and Smrekar 2009). For a sample with the size of n, n raw data are converted to rank data. The correlation coefficient ρ can be expressed as

$$\rho = \frac{\Sigma_i(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{(\Sigma_i(r_i - \bar{r})^2\Sigma_i(s_i - \bar{s})^2)}} \quad (6)$$

where r_i and s_i are the ranks of x_i and y_i, respectively. i = 1, 2, ... n. When an equal value appears in the variables, the rank corresponding to this value is the average of the ranks corresponding to these values. The values of ρ are within the range of [-1, 1]. The greater the $|\rho|$, the greater the correlation. $|\rho| = 0$ indicates the minimum correlation; $\rho > 0$ indicates the positive correlation and vice versa.

In this study, R software is used to perform Spearman correlation analysis on the transparency, shear strength, DAP, SR, and SSR of transparent soil.

3.3 Overview of machine learning theory

3.3.1 Support vector regression

As a supervised learning algorithm, the Support Vector Regression is a new general algorithm proposed by Vapnik (2013) based on the statistical theory of VC dimension and the principle of structural risk minimization. It can better solve problems, such as small samples, nonlinearity, high dimensionality, and local minimal points. Thus, it is regarded as a better algorithm to replace ANN. The process of SVM is to map the original training data from the input space to a feature space with higher dimensions. The classification is then performed by choosing an optimal hyperplane toward maximizing the class margins (Kanevski *et al*. 2008).

First, k sample data is given, and its value is expressed as $\{x_k, y_k\}$, where $x_k \in R^n$ is n-dimensional vector and $y_k \in R$ is the corresponding output variable. The dataset is mapped to a high-dimensional feature space H via a nonlinear mapping φ, and linear regression is performed in this space. The specific function form can be expressed as

$$\mathrm{f(x)} = \left(\omega, \varphi(x_k)\right) + b, \quad \varphi: R^n \to H, \omega \in R^n, \quad (7)$$

where b is the offset. The separating hyperplane for linear separable data can be defined as

$$y_i(X_i^T w + c) \ge 1 - \xi_i \quad (8)$$

where X_i^T is transpose of the ith sample; w is orientation of the hyperplane; c is offset of the hyperplane from the origin; and ξ_i is positive slack variables. The following equation can be used to select the optimal hyperplane

$$\text{Minimize } \frac{1}{2}w^2 + C\sum_{i=1}^{l}\xi_i \quad (9)$$

under the constraints $y_i(X_i^T w + c) \ge 1 - \xi_i$, where C is penetrating parameter defining the trade-off between the number of mis-classification and the maximization of margin.

3.3.2 XGBoost

The XGBoost (eXtreme Gradient Boosting) algorithm is an optimization of the GBDT algorithm, which was first proposed by Chen *et al*. (2015). The biggest difference between XGBoost and GBDT is that the latter only uses the first-order derivative of the loss function to calculate the residuals, while the former uses both the first- and second-order derivatives of the loss function. Therefore, the

XGBboost algorithm allows for a second-order derivation of the loss function.

XGBoost uses the second-order Taylor expansion for the loss function. For the *t*-th loss function.

$$L^{(t)} = \sum_{i=1}^{k} l\left(y_i, y_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{10}$$

The following equation can be obtained by performing a quadratic Taylor expansion

$$L^{(t)} \simeq \sum_{i=1}^{k}\left[l\left(y_i, y_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x)\right)\right] + \Omega(f_t) \tag{11}$$

where $g_i = \partial_{y^{(t-1)}} l\left(y_i, y^{(t-1)}\right)$, $h_i = \partial^2{}_{y^{(t-1)}} l(y_i, y^{(t-1)})$.

In addition, XGBoost has other improvements over GBDT, including: (a) The complexity of the tree model is added to the optimization target as a regularization term; (b) The approximate algorithm is used to find the best segmentation point; (c) The sparse value is considered, which improves the efficiency; (d) The sorting and block of data are beneficial to parallelization; and (e) Cache and memory are optimized.

Based on these optimizations, the XGBoost model can learn faster than GBDT for the same data size without any decrease in training accuracy. Compared with traditional data mining algorithms, XGBoost has the advantage of high accuracy, scalability. XGBoost has been extensively used in online public opinion and e-commerce prediction, while its application in civil engineering is not yet wide enough, leaving a huge room for research.

A benefit of using XGBoost is that after the trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute. Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. This importance is computed for each attribute in the dataset which can lead to ranking and comparing the attributes. Importance, for a single decision tree, is obtained regarding the amount that each attribute split point improves the performance measure, and is weighted by the number of observations. The performance measure may be the purity (Gini index) used for choosing the split points or another more specific error function.

The Gini index is usually used as an evaluation indicator. The formula for calculating the Gini index is shown in Eq. (12).

$$GI_m = 1 - \sum_{k=1}^{K} \hat{p}_{mk}^2 \tag{12}$$

In Eq. (12), K is the number of categories in the sample set. $\hat{p}_{mk}^2$ is the probability estimate that the sample at node m belongs to class k. The Gini index importance score for the characteristic X_j, measured as the average of the node splitting impurity of this variable across all trees, as shown in Eq. (13). where GI_l and GI_r denote the Gini index of two new nodes after node m branching, respectively.

$$VIM_j^{Gini} = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} GI_m - GI_l - GI_r \tag{13}$$

The feature importance is then averaged across all of the decision trees within the model.

3.3.3 Random Forest

The Random Forest algorithm was proposed by Breiman (2001) in 2001 as an extension of the Bagging method. Basically, it forms multiple trees by integrating learning ideas. Its basic unit is a decision tree containing multiple decision trees trained by the Bagging integration learning technique. When the samples to be classified are entered, the final classification result is determined by voting on the output of a single decision tree.

The implementation process of the Random Forest algorithm is as follows: First, the number *n* of decision trees in the Random Forest is determined. Second, *N* training sets with the same scale are obtained by random sampling and replacement. Third, *n* attribute subsets are obtained by randomly selecting k attributes in the attribute set each time, and $k = \log_2 V$. Forth, each training set corresponds to a subset of attributes and is trained by a decision tree algorithm to obtain a decision tree that grows to its maximum without any pruning operations. Fifth, the random forest model is obtained by synthesizing these decision trees. Sixth, for a prediction sample, a classification result is obtained from each of these decision trees. The result is obtained by a voting mechanism, i.e., the class that appears most frequently in the classification results of these decision trees is the final output of the Random Forest model.

3.3.4 Gradient Boosting Decision Tree (GBDT)

GBDT (Gradient Boosted Decision Trees) is an improved nonlinear model proposed by Friedman (2001) based on the Boosting algorithm. GBDT consists of several decision trees, each small in scale. The essence of GBDT is that, during each iteration, each tree learns the conclusions and residuals of all trees in the previous cycle. Then a decision tree is re-established in the gradient direction that reduces residuals and improves prediction accuracy. When a sample is entered, the model will first assign an initial value, which will then traverse each decision tree. Subsequently, the latter will adjust and correct the predicted value and accumulate the results of each decision tree to obtain the final prediction.

Regarding the regression problem of mean square error, the specific flow chart is as follows:

(a) Initialization model as a constant

$$F_0(X) = arg_\gamma min \sum_{i=1}^{k} L(y_i, \gamma), \tag{14}$$

Iteratively generate m classifiers

a) Calculate the residuals r_{mi}

$$r_{mi} = -\left[\frac{\partial L(y_i, F(x_i)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \text{ for } i = 1,2 \ldots, k \tag{15}$$

The negative gradient of the loss function is taken as the

value of the current model and as an estimation of the residual.

b) Based on $\{(x_i, r_{mi})\}_{i=1}^{k}$, a base classifier $h_m(x)$ is generated;

c) Calculate the optimal γ_m

$$\gamma_m = arg_{\gamma} min \sum_{i=1}^{k} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right) \quad (16)$$

d) Update model

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (17)$$

(c) Output the final GBDT classifier $F_M(x)$.

The negative gradient of the loss function is used as the value of the current model and as an estimation of the residuals.

The advantage of the GBDT algorithm is that Boosting method can effectively improve the classification accuracy without the need for sophisticated feature engineering and feature transformation. The model is highly expressive due to the use of nonlinear changes.

3.3.5 Grid search CV

In machine learning models, it is often necessary to select the optimal hyper-parameters to tune the model. Improper selection of hyper-parameters can lead to underfitting or overfitting, resulting in poor model performance. As the main method of hyper-parameter selection, Grid-Search CV adjusts the parameters within a specified parameter range in the order of step size for each combination in the list of hyper-parameter combinations. Then it trains the learner using the adjusted parameters, and performs CV sub-cross-validation to select the hyper-parameter combination with the highest average score as the optimal choice, thus finding the parameter with the highest accuracy on the validation set.

3.3.6 Evaluation of model performance

The accuracy of a model is assessed by Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2 (R Squared). These four indicators are popular and important in model validation. The formulas are shown in Table 5, where y_i and y refer to the measured and mean values of dependent variables (the transparency and the shear strength of soil), respectively. $pr_i$and p_r are output values from the model.

Table 5 Evaluation index formula

Standard	Formula expression
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(pr_i - y_i)^2}$
MSE	$\frac{1}{n}\sum_{i=1}^{n}(pr_i - y_i)^2$
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert(pr_i - \overline{y_i})\rvert$
R^2	$\frac{\sum_i(\overline{y_i} - pr_i)^2}{\sum_i(\overline{y_i} - y_i)^2}$

3.4 Data modelling and process

In order to generate a data set for modeling, the transparency of transparent soil is taken as the dependent variable (Y). Besides, DAP, SR, and SSR are represented by independent variables X_1, X_2, and X_3 respectively for transparency and Shear strength test. Variable data is divided into training data set (80%) and test data set (20%). The statistical data used are shown in Fig. 3.

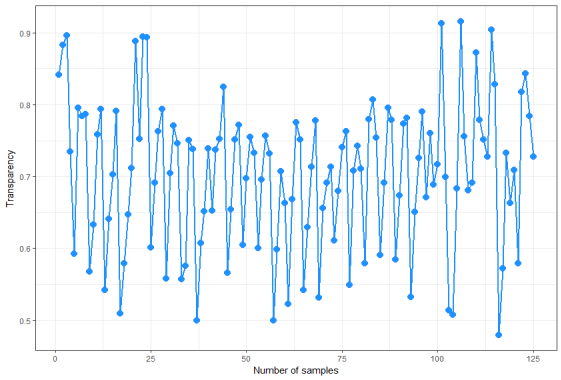
4. Results and discussion

4.1 Elasticnet regression analysis

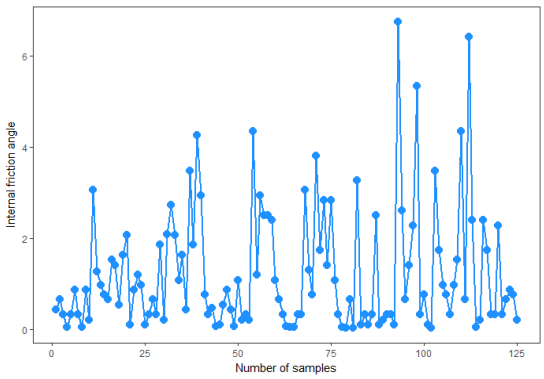
4.1.1 Model assumption

The following hypotheses are proposed for the regression model.

Assumption 1, Factor A (DAP) has a negative effect on transparency and a positive effect on shear strength (φ). As DAP increases, transparency decreases and shear strength (φ) increases. Since the cost of preparation varies with



(a) Transparency data map of 125 groups of transparent soil samples



(b) Internal friction angle of 125 groups of transparent soil samples

Fig. 3 Data diagram of 125 groups of transparent soil samples

particle size, the selection of appropriate particle size is critical for the formulation of transparent clays. As mentioned above, economic factors should also be considered when ensuring adequate transparency and shear strength. Therefore, in this paper, silica powders with particle sizes of 15 nm and 30 nm are used. The increase in DAP means smaller particle size in the solute and solution mix, which enables the solute to better integrate with the solution, resulting in decreased transparency and increased shear strength.

Assumption 2: Factor B (SR) has a positive effect on transparency and a negative effect on shear strength (φ). Transparency increases and shear strength (φ) decreases with increasing SR. In the process of preparing transparent soil, two solutions of white oil and tridecane need to be mixed in a certain ratio. An increase in the SR means an increase in the white oil content. According to the physical parameters of the two solutions, the density of tridecane is greater than that of white oil. Two solutions are colorless and transparent.

In a solution of the same mass, the transparency of the solution increases and the density decreases when the amount of less dense white oil increases. The increase in density lowers the density mass of the transparent soil at a given mass and raises the light transmittance, thus increasing the transparency and decreasing the shear strength (φ).

Table 6 Statistic description of 125 data sets

Variable	Obs	Obs	Std. Dev.	Min	Max
DAP	125	1.186	0.6779535	0.43	2.33
SR	125	1.208	0.3911006	0.72	1.82
SSR	125	7	1.419905	5	9
T	125	0.7028795	0.102608	0.4797625	0.9160666
S	125	0.1579079	0.0655533	0.0380729	0.3189737

Table 7 Spearman correlation coefficients

	DAP	SR	SSR
DAP	1.0000		
SR	0.0000	1.0000	
SSR	0.0000	0.0000	1.0000

4.1.2 Elastic regression model development

Based on the above assumptions, the regression models for transparency and shear strength were developed, as shown in Eqs. (18) and (19).

model 1:

$$T = \alpha_1 \times DAP \times \beta_1 \times SR + \gamma_1 \times SSR + \varepsilon_1 \quad (18)$$

model 2:

$$S = \alpha_2 \times DAP + \beta_2 \times SR + \gamma_2 \times SSR + \varepsilon_2 \quad (19)$$

Annotation: Transparency = T; Shear strength = S; Different aggregate proportions = DAP; Solvent ratio = SR; Solute solution ratio = SSR; modulus：α_n, β_n, γ_n, ε_n

4.1.3 Experimental analysis of models

(a) Sample description

The main characteristics of the 125 data sets include the three independent variables (DAP, SR, SSR) and the two dependent variables (T, S). See Table 6 for a summary of these characteristics.

(b) Correlation analysis

The results of the correlation analysis are shown in Table 7. It is shown that the correlation coefficients of three independent variables (DAP, SR, SSR) are 0, indicating that the three independent variables are independent of each other. Therefore, there is no multi-collinearity in the regression model.

(c) Regression results

The model was validated and compared using transparency as a criterion, and the results are shown in Tables 8-9. At a significance level α of 0.05, the parameter estimates of transparency and shear strength showed that the p-values of the three factors, DAP, SR, and SSR, all passed significance with less than 0.05 test, which validates

Table 8 Elastic net regression results for transparency and shear strength

Factors	Transparency			Shear strength		
	Estimate	Estimate Std.	Pr (> |t|)	Estimate	Estimate Std.	Pr (> |t|)
Intercept	0.7028	0.0066	< 2e-16 ***	0.1579	0.0042	< 2e-16 ***
Different aggregate proportions	-0.0144	0.0067	0.0322 *	0.0093	0.0042	0.0298 *
Solvent ratio	0.0529	0.0067	1.20e-12 ***	-0.0352	0.0042	1.55e-13 ***
Solute solution ratio	-0.0461	0.0067	2.37e-10 ***	0.0281	0.0042	1.05e-09 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 9 Goodness of fit of elastic net regression results

	R^2	RMSE	MSE	MAE
Transparency	0.6883015	0.1827612	0.03340166	0.3646985
Shear strength	0.4933289	0.1161859	0.01349917	0.2345066

our hypothesis.

Tables 8-9 show that two factors, namely, DAP and SSR, are negatively correlated with transparency. Transparency decreases with an increase in both factors. We also find that with every unit increase in SR, transparency increases by 0.0529, indicating that SR has a positive effect on transparency.

In addition, a positive influence of the two factors on the shear strength was observed. Besides, the shear strength is negatively affected by SR, which is quite the opposite of the case of transparency. With every unit increase in the SR, the shear strength decreases by 0.0352. These phenomena verify that the results of our elastic regression model are consistent with the phenomena observed in our usual experiments. Therefore, the effects of the three factors on transparency and shear strength can be verified by the elastic regression model, which quantitatively describes the coefficients of the effects of the three factors on transparency and shear strength.

Table 10 Results of correlation test between transparency and shear strength of transparent soil

Correlation coefficient	P
-0.9960955	0.000

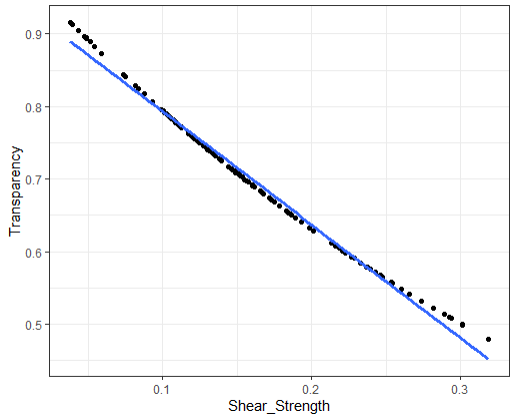


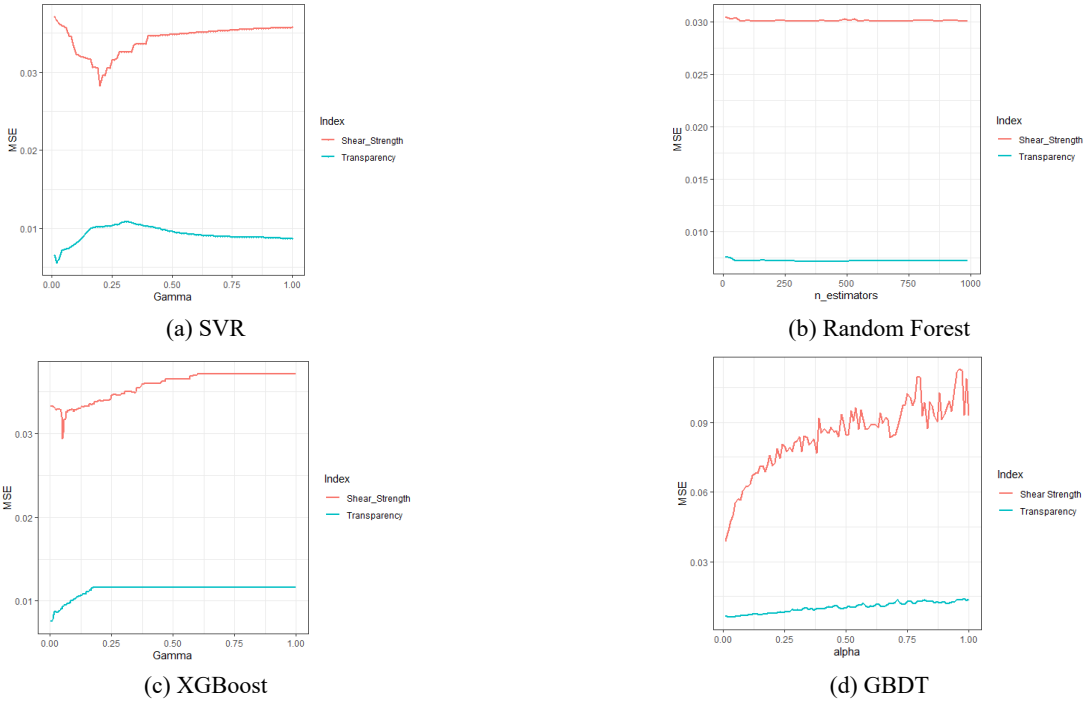
Fig. 4 Relationship between transparency and shear strength of transparent soil

4.2 Spearman's correlation test

The following assumptions regarding the relationship between transparency and shear strength are made.

Assumption 4: Transparency and shear strength are negatively correlated.

The transparency of transparent soil relates to the refractive index. During the preparation process, the solute consists of many particles, while the solution is a transparent liquid. During the fusion of solute and solution, the solution's density gradually increases, thus raising the shear strength (φ) and decreasing transparency. Table 10 shows Spearman's correlation coefficient and significance test for transparency and shear strength is -0.99, meaning that transparency and shear strength have a strong negative correlation. Moreover, the p-value (0.000) is lower than 0.05, indicating that the correlation is significant. These data verify the hypothesis that transparency and shear strength of transparent soil is adversely proportional. In other words, in a general trend, the larger the shear strength is the lower the transparency is. Fig. 4 visually depicts the process by which the transparency of transparent soils varies with shear strength.



(a) SVR (b) Random Forest

(c) XGBoost (d) GBDT

Fig. 4 Transition of average MSE values along with generations on the training set

Table 11 Hyper-parameters of ML Algorithms

Algorithms	Parameters	Explanation	Settings	Best parameter	
				Transparency	Shear strength
XGBoost	learning_rate	The weight of the model produced in each iteration	0.1-1	0.178	0.118
	max_depth	The maximum depth of the tree	1-100	1	1
	min_child_weight	The minimum weight sum in each leaf	1-100	1	21
	gamma	The minimum loss function required for node splitting	0.01-1	0.01	0.05
	reg_alpha	Weight of L1 regular term	0.01-1	0.024	0.02
	n_estimators	The number of boosting stages	10-500	10	80
SVR	kernel	Kernel type used in the algorithm	rbf, linear	rbf	rbf
	C	Penalty parameter of the error term	1-100	19.2	1
	gamma	The distribution after the data is mapped to the new feature space	0.01-1	0.02	0.2
Random Forest	n_estimators	The number of trees	10-1000	410	270
	criterion	Node division criteria	gini, entropy	gini	gini
	max_depth	The maximum depth of the tree	5-50	10	20
	min_samples_split	The minimum number of samples that can be divided by nodes	2-10	5	8
GBDT	learning_rate	The weight of the model produced in each iteration	0.1-1	0.198	0.01
	max_depth	The maximum depth of the tree	1-100	3	5
	min_samples_split	The minimum number of samples that can be divided by nodes	100-2000	1800	1400
	min_samples_leaf	Minimum number of leaf nodes	10-100	90	70
	n_estimators	The number of trees	10-1000	100	190
	alpha	The quantile value of the loss function	0.01-1	0.03	0.01

Table 12 Comparison of training sets of different models

Dependent variable	Model	R^2	RMSE	MSE	MAE
Transparency	SVR	0.5783	0.0658	0.0043	0.0537
	RF	0.4235	0.0843	0.0071	0.0652
	XGBoost	0.5933	0.0646	0.0041	0.0507
	GBDT	0.4371	0.0769	0.0060	0.0641
Shear strength (φ)	SVR	0.4215	0.1714	0.0293	0.1250
	RF	0.4117	0.1734	0.0301	0.1215
	XGBoost	0.4748	0.1681	0.0282	0.1103
	GBDT	0.4052	0.1843	0.0389	0.1316

Table 13 Comparison of testing sets of different models

Dependent variable	Model	R^2	RMSE	MSE	MAE
Transparency	SVR	0.4700	0.0868	0.0055	0.0590
	RF	0.4137	0.0863	0.0080	0.0712
	XGBoost	0.4821	0.0746	0.0048	0.0582
	GBDT	0.4231	0.0829	0.0074	0.0679
Shear strength (φ)	SVR	0.4013	0.1894	0.0469	0.1487
	RF	0.4075	0.1873	0.0364	0.1468
	XGBoost	0.4318	0.1684	0.0306	0.1105
	GBDT	0.4273	0.1796	0.0345	0.14357

4.3 Prediction of tansparency and shear strength using machine learning algorithms

4.3.1 Hyper-parameter tuning of machine learning algorithms

The Grid-Search CV method was adopted for Hyper-parameter tuning of the four machine learning algorithms, with the process shown in Fig. 5. It can be seen that by hyper-parameter tuning, all four algorithms obtained the minimum value of MSE and the model effect was improved. The hyper-parameters tuning results are shown in Table 11.

4.3.2 Performance of machine learning algorithms

In this section, RMSE, MAE, MSE, and R^2 goodness of fit superiority indexes are used to examine the transparency and shear strength prediction effects of the four algorithms. The results are shown in Tables 12-13. The following is a

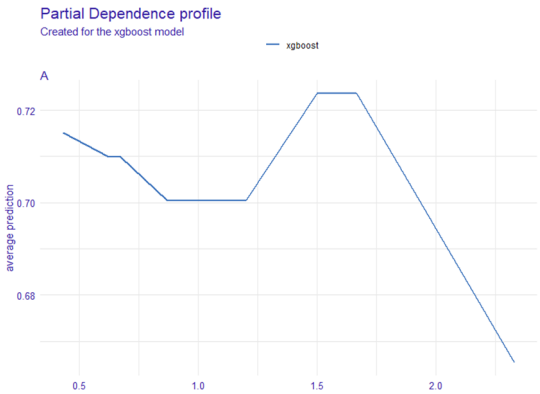
comparison of the indicators in the two tables:

The models are validated and compared using RMSE as a criterion. In terms of transparency, it can be observed that all the models in the training set (Table 12) have RMSE value less than 0.09, and XGBoost has the best prediction performance with the lowest prediction error (RMSE = 0.0646). In the test set (Table 13), XGBoost has the lowest prediction error (RMSE = 0.0746), followed by GBDT (RMSE = 0.0829), RF (RMSE = 0.0863), and SVR (RMSE = 0.0868).

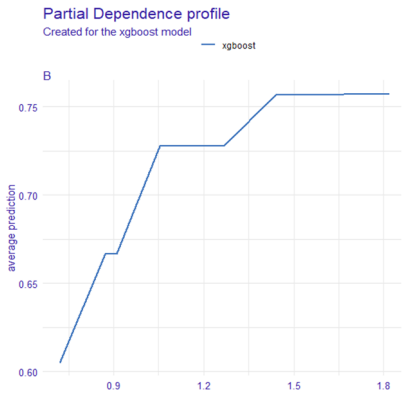
With regard to shear strength, in the training set (Table 12), XGBoost has the lowest prediction error (RMSE = 0.1681), followed by SVR (RMSE = 0.1714), RF (RMSE = 0.1734), and GBDT (RMSE = 0.1843). As for the test set (Table 13), XGBoost (RMSE=0.1684) had the lowest prediction error, followed by GBDT (RMSE = 0.1796), RF (RMSE = 0.1873), and SVR (RMSE = 0.1894).

The goodness of fit can be determined by R^2 and the results of the training set are shown in Table 12. In terms of transparency, XGBoost has the highest R^2 value (0.5933), followed by SVR (R^2 = 0.5783), GBDT (R^2 = 0.4371), and RF (R^2 = 0.4235). The data mean that XGBoost has the best fit goodness. For the test set (Table 13), XGBoost (R^2 = 0.4821), SVR (R^2 = 0.4700), GBDT (R^2 = 0.4231), and RF (R^2 = 0.4137).

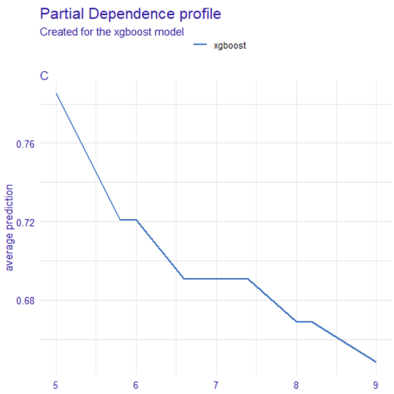
In terms of the Shear strength, Table 12 shows that XGBoost has the highest goodness of fit with the R^2 value being 0.4748, followed by SVR (R^2 = 0.4215), RF (R^2 = 0.4117), and GBDT (R^2 = 0.4052). For the test set (Table 13), XGBoost (R^2 = 0.4318), GBDT (R^2 = 0.4273), RF



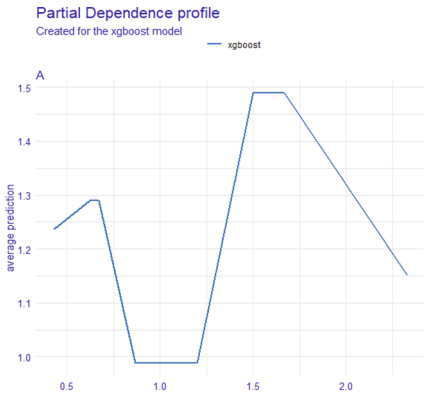
(a) Variation in the impact of factor A on transparency



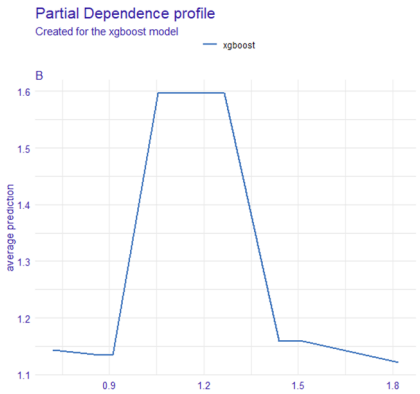
(b) Variation in the impact of factor B on transparency



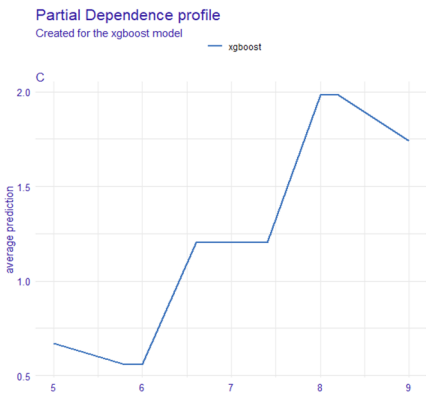
(c) Variation in the impact of factor C on transparency



(d) Variation in the impact of factor A on shear strength



(e) Variation in the impact of factor B on shear strength



(f) Variation in the impact of factor C on shear strength

Fig. 6 Variation curves for transparency and shear strength based on XGBoost model analysis of each factor

(R^2 = 0.4075) and SVR (R^2 = 0.4013).

The results show that all the models have a good performance in predicting the transparency and shear strength of transparent soils. The values of RMSE and R^2 show that XGBoost is the best predictor, outperforming other models (GBDT, SVR, RF). Similarly, the MAE, MSE pattern is consistent with the RMSE.

4.3.3 The change process of transparency and shear strength with influencing factors

Machine Learning (ML) models are widely used in classification or regression. But such unverified black box models usually lack direct interpretability. Here, we use DALEX package in *R* language, which contains various methods that help to understand the link between input variables and model output to interpret any model and help to explain its behavior.

Fig. 6 shows the influencing process of factor A (DAP), B (SR), and C (SSR) on transparency (T) and shear strength (φ) observed by the optimal model XGBoost.

Figs. 6(a)-(c) shows how transparency varies with the three factors respectively. According to Fig. 6(a), when factor A is in the range of [0.43, 0.87], the transparency showed a stepwise downward trend and then continues to increase, reaching a maximum value of 0.725 in the range of [1.5, 1.7]. Fig. 6(b) shows that when factor B is in the range [0.72, 1.4], transparency increases in a stepwise manner and reaches its maximum value at [0.76], ultimately stabilizing at the range [1.4, 1.8]. Fig. 6(c) illustrates that transparency peaks when factor C is 5% Then it decreases in a stepwise pattern as factor C increases until reaching the lowest point at 9%.

Figs. 6(d)-(f) shows how shear strength varies with the three factors respectively. Fig. 6(d) depicts that the shear strength decreases linearly when factor A is within [0.62, 0.70]. Then it reaches the minimum within [0.7, 1.2]. When A = 1.2, it begins to increase linearly and starts to decrease after peaking within [1.5, 1.7]. Fig. 6(e) shows that when factor B is in the range of [0.72, 0.91], φ shows a slowly decreasing trend. After 0.91, it starts to surge straightly until reaching the maximum value of 1.59 when factor B is in the range of [1.05, 1.3]. When factor B reaches 1.40, φ begins to decline first linearly then flatly. Fig. 6(f) shows that when factor C is [5%, 6%], φ shows a slowly decreasing trend, followed by a stepwise increase after factor C = 6%. Then φ reaches a maximum value of 1.9 when factor C is in the range of [8%, 8.1%] and then begins to decline slowly.

4.3.4 Feature Importance Analysis (Score) of Influencing Factors

This section discusses the importance of each influencing factor on the transparency and shear strength of transparent soils. As mentioned above, SSR and SR had the most significant influence on the transparency and shear strength. It has been proved that XGBoost is the model with the best prediction effect, this paper utilizes the model to obtain the importance scores of DAP, SR, SSR on transparency and shear strength. The importance values for solvent ratio, solute solution ratio, and different aggregate proportions were 0.44, 0.45, and 0.11, respectively, for transparency, and 0.24, 0.57, and 0.19 for shear strength.

The results show that the SSR is the most significant variable affecting transparency with the importance score being 0.45. The importance score of the SR is 0.44, ranking second. DAP has the lowest importance score of 0.11, 75.6% lower than the one ranking first. This shows that the DAP has the least influence on transparency. Therefore, we must first determine the best proportion of SSR before preparing transparent soils to obtain the optimal transparency. The second-ranking factor should not be overlooked; after all, it is only 0.01 points lower than the first-place importance score.

The three variable factors affected shear strength in the same order of significance as they affected transparency. However, the difference in scores among the three is larger. The highest score of 0.57 was 137.5% higher than the second-place score of 0.24. Therefore, in the preparation of transparent soil, we must first determine the optimal SSR to obtain the optimal shear strength.

4.4 Discussion

4.4.1 Analysis of factors influencing transparency and shear strength based on elastic regression models and Spearman's correlation analysis

According to the elastic regression model analysis, two factors namely DAP and SSR, have a negative effect on transparency and shear strength. The main reason for this phenomenon is that DAP is 15 nm and 30 nm silica powder mass ratio, which means that the more the 15 nm particles, the smaller the solute particle size, and the more solute dissolved in the solution, resulting in reduced transparency of the prepared transparent soil. Therefore, smaller solute particle size doesn't necessarily decrease the transparency in experiment. Similarly, an increase in SSR leads to more solute, which affects the refractive index of the prepared soil, thus leading to a decrease in transparency.

In contrast, the factor SR has a positive effect on transparency. Because the factor SR refers to the ratio of white oil and tridecane, an increase in the SR means an increase in the white oil content. It is known that the density of tridecane is greater than the density of white oil. With the same solution quality, the higher the white oil content, the lower the density of the corresponding prepared solution, resulting in a less dense and more transparent prepared transparent soil. This phenomenon is consistent with our observation, and the elastic regression model clearly illustrates the coefficients that quantitatively describe the effect of the three factors on transparency.

Similarly, the regression analysis based on the elastic regression model shows that the two factors DAP and SSR have a positive effect on the results of shear strength. The main reason for this phenomenon is that DAP is the mass ratio of silica powder with the particle size of 15 nm and 30nm respectively. that the more 15 nm particles means that the smaller the solute particle size, the more it dissolves in the solution, resulting in increased friction between particles within the transparent soil. Therefore, when preparing transparent soils, one feasible method to increase the shear strength is to increase the particle size of smaller solutes. Similarly, an increase in the SSR means more solute, which

results in greater interparticle friction in transparent soils and thus a higher shear strength. This phenomenon is consistent with the data obtained by our experiments, and the elastic regression model clearly illustrates the coefficients that quantitatively describe the effect of the three factors on transparency.

4.4.2 Machine learning-based prediction of transparency and shear strength

Based on Lambert-Beer law, the transparency of transparent soils is measured by ultraviolet-visible spectrophotometer, which provides a quantitative basis for studying the factors influencing transparency. Four machine learning methods are used to verify the transparency of transparent soils. The analysis of the model verification results shows that XGBoost has the best performance, followed by SVR, GBDT, and RF. However, when it comes to the predictive ability for transparency and shear strength of transparent soils, XGBoost and SVR performed acceptably, while RF and GBDT performed poorly.

According to SVR results, although the optimization method that solves constrained quadratic programming functions in SVR is better and more comprehensive, and has the ability to deal with overtraining problems, the SVR is constrained by many parameters (e.g., hidden layer number, learning rate, number of training hours, hidden node number, momentum term, weight initialization technology, transfer function). These are difficult to optimize (Samui 2008) simultaneously in the learning models. Furthermore, although Das *et al*. (2011) think the SVM model is better in predicting residual strength of soft soil and Kanungo *et al*. (2014) believes that ANN is a promising method for predicting soil shear strength parameters, XGBoost instead of SVR was the best performer in predicting transparency and shear strength of transparent soils in our study.

Although machine learning techniques such as RF, XGBoost, SVR, and GBDT are advanced methods in forecasting problems, their performance is highly dependent on the quality (Mair *et al*. 2000) of the data used. In terms of geotechnical engineering, the use of variables determined by different tests can lead to results deviation of different samples from the same soil mass, thus affecting the use of the model. In this study, these four models show acceptable predictive ability. However, the sample data can be expanded to improve the goodness of fit in the models. Also, the over-sampling method is adopted to process the unbalanced dataset (He *et al*. 2008) to improve the predictive ability of the model.

4.4.3 On the applicability of machine learning

Compared with previous methods of transparency observation, the transparency test method in this paper is more concise. Traditionally, the preparation of transparent soils and model test not only take a long time but also requires a large number of experimental materials, which is time-consuming and laborious. Therefore, using advanced machine learning method is effective in determining the soil mechanics parameters such as transparency and shear strength and reduces the experimental cost of transparent soils.

Machine learning analysis will further improve the success rate of transparent soil preparation. Machine learning will play an important role in the field of transparent soil model tests, and its level of prediction can also extend the application of artificial intelligence systems. However, the establishment of machine learning model requires more test information such as data points, than traditional methods, which has application limitations for indoor testing of rock and soil.

Transparent soils are man-made soils with unpredictable properties. Due to the influence of many factors (e.g., experimental conditions, equipment, the experience of test personnel, etc.), sometimes its properties cannot be measured accurately in the laboratory. In this study, the average error rate of the advanced machine learning model for predicting shear strength is below the standard deviation, which is acceptable for geotechnical issues. Therefore, these machine learning techniques can also be applied to the simulation of transparency and shear strength of transparent soils.

Surmounting some limitations and executing innovative ideas may result in even more applicable methods. Data optimization, for example, normally leads to less complex networks which consequently are trained in a shorter time and probably higher accuracy. This idea can be achieved by taking the most important factors into the equation, as well as purifying the dataset.

It is true that machine learning that was employed here presented a nice performance for the testing stage but the generalization potential can also be tested by new and different laboratory results.

Last but not least, although several well-known techniques are employed and compared in this study conducting comparative works is recommended for outlining the most efficient technique. There are still potential methods like ANN and neuro-fuzzy tools that should be considered as well. Moreover, the training algorithms in this study were conventional which needs to be compared with the hybrid version of the used models. For this purpose, the training task can be relegated to optimization techniques like particle swarm optimization (Moayedi *et al*. 2019, Bai *et al*. 2021) and differential evolution (Moayedi *et al*. 2020). However, an important factor in dealing with such algorithms is time-efficiency. Hybrid methods usually are slower than conventional versions, due to the time elapsed for finding the optimal solution.

5. Conclusions

The transparency of transparent soils was measured by UV-visible spectrophotometer according to the Lambert-Bill law, which provides a quantitative basis for studying the factors influencing transparency. Based on the elastic net analysis, the influence coefficients of three factors, namely, DAP, SR, SSR on transparency are -0.014459, 0.052942, and -0.046142, on shear strength are 0.009326, -0.035286, 0.028069, respectively. According to the Spearman correlation test, the correlation coefficient between transparency and shear strength is -0.99.

The predictive performances of four machine learning algorithms XGBoost, GBDT, SVR, and RF on the transparency and shear strength of transparent soils were compared. The results show that XGBoost performs best and is the best model. Based on the model, the SSR has the largest importance score on the effects of transparency and shear strength of transparent soil, with importance scores of 0.45 and 0.57 for transparency and shear strength, respectively. For transparency, the score of SSR is 0.11 and 0.34 higher than that of DAP and SR, respectively; for shear strength, the score of SSR was 0.33 and 0.38 higher than that of DAP and SR, respectively. Based on the interpretable analysis of the XGBoost model, the variation process of transparency and shear strength with factors DAP, SR, SSR is investigated. The transparency reaches its maximum when factor DAP is within the range of [1.5, 1.7], SR is within the range of [1.4, 1.8], and SSR is 5%, respectively. The shear strength (φ) reaches its maximum when DAP is within the range of [1.5, 1.7], SR is within the range of [1.0, 1.4], SSR is within the range of [7.9%, 8.2%], respectively. SSR is the variable that has the most significant influence on the transparency and shear strength of transparent soils.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central University (No. 2018CDYJSY0055), the National Natural Science Foundation of China (No. 51478066) and Chongqing Natural Science Foundation of China (No. cstc2018jscx-msybX0271).

Data availability

Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

References

Abedini, M. and Zhang, C. (2021), "Dynamic performance of concrete columns retrofitted with FRP using segment pressure technique", *Compos. Struct.*, **260**, 113473. https://doi.org/10.1016/j.compstruct.2020.113473

Bai, B., Guo, Z., Zhou, C., Zhang, W. and Zhang, J. (2021), "Application of adaptive reliability importance sampling-based extended domain PSO on single mode failure in reliability engineering", *Inform. Sci.*, **546**, 42-59. https://doi.org/10.1016/j.ins.2020.07.069

Bhattacharya, S., Maddikunta, P.K.R., Kaluri, R., Singh, S., Gadekallu, T.R., Alazab, M. and Tariq, U. (2020), "A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU", *Electronics*, **9**(2), 219. https://doi.org/10.3390/electronics9020219

Breiman, L. (2001), "Random forests", *Machine Learning*, **45**(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bui, D.T., Ghareh, S., Moayedi, H. and Nguyen, H. (2019), "Fine-tuning of neural computing using whale optimization algorithm for predicting compressive strength of concrete", *Eng. Comput.*, **37**, 701-712. https://doi.org/10.1007/s00366-019-00850-w

Chahnasir, E.S., Zandi, Y., Shariati, M., Dehghani, E., Toghroli, A., Mohamad, E.T., Shariati, A., Safa, M., Wakil, K. and Khorami, M. (2018), "Application of support vector machine with firefly algorithm for investigation of the factors affecting the shear strength of angle shear connectors", *Smart Struct. Syst., Int. J.*, **22**(4), 413-424. https://doi.org/10.12989/sss.2018.22.4.413

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H. (2015), "Xgboost: extreme gradient boosting", R package version 0.4-2 1-4.

Chen, H., Qiao, H., Xu, L., Feng, Q. and Cai, K. (2019), "A fuzzy optimization strategy for the implementation of RBF LSSVR model in vis–NIR analysis of pomelo maturity", *IEEE Transact. Indust. Inform.*, **15**(11), 5971-5979. https://doi.org/10.1109/TII.2019.2933582

Chou, J.S., Yang, K.H. and Lin, J.Y. (2016), "Peak shear strength of discrete fiber-reinforced soils computed by machine learning and metaensemble methods", *J. Comput. Civil Eng.*, **30**(6), 04016036. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000595

Das, S., Samui, P., Khan, S. and Sivakugan, N. (2011), "Machine learning techniques applied to prediction of residual strength of clay", *Open Geosciences*, **3**(4), 449-461. https://doi.org/10.2478/s13533-011-0043-1

Delicado, P. and Smrekar, M. (2009), "Measuring non-linear dependence for two random variables distributed along a curve", *Statist. Comput.*, **19**(3), 255. https://doi.org/10.1007/s11222-008-9090-y

Ding, L., Huang, L., Li, S., Gao, H., Deng, H., Li, Y. and Liu, G. (2020), "Definition and application of variable resistance coefficient for wheeled mobile robots on deformable terrain", *IEEE Transact. Robotics*, **36**(3), 894-909. https://doi.org/10.1109/TRO.2020.2981822

Ezzein, F.M. and Bathurst, R.J. (2011), "A transparent sand for geotechnical laboratory modeling", *Geotech. Test. J.*, **34**(6), 590-601. https://doi.org/10.1520/GTJ103808

Ezzein, F.M. and Bathurst, R.J. (2014), "A new approach to evaluate soil-geosynthetic interaction using a novel pullout test apparatus and transparent granular soil", *Geotext. Geomembr.*, **42**(3), 246-255. https://doi.org/10.1016/j.geotexmem.2014.04.003

Friedman, J.H. (2001), "Greedy function approximation: a gradient boosting machine", *Annals of statistics*, 1189-1232.

Fu, X. and Yang, Y. (2020), "Modeling and analysis of cascading node-link failures in multi-sink wireless sensor networks", *Reliabil. Eng. Syst. Safety*, **197**, 106815. https://doi.org/10.1016/j.ress.2020.106815

Gadekallu, T.R., Rajput, D.S., Reddy, M.P.K., Lakshmanna, K., Bhattacharya, S., Singh, S., Jolfaei, A. and Alazab, M. (2020), "A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU", *J. Real-Time Image Process.*, 1-14. https://doi.org/10.1007/s11554-020-00987-8

He, H., Bai, Y., Garcia, E.A. and Li, S. (2008), "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", *Proceedings of 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, China, June, pp. 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

He, S., Guo, F. and Zou, Q. (2020), "MRMD2. 0: a python tool for machine learning with feature ranking and reduction", *Current Bioinformat.*, **15**(10), 1213-1221. https://doi.org/10.2174/1574893615999200503030350

Iskander, M. (2010), *Modelling with transparent soils: Visualizing soil structure interaction and multi phase flow, non-intrusively*, Springer Science & Business Media.

Iskander, M.G., Liu, J. and Sadek, S. (2002), "Transparent

amorphous silica to model clay", *J. Geotech. Geoenviron. Eng.*, **128**(3), 262-273. https://doi.org/10.1061/(ASCE)1090-0241(2002)128:3(262)

Ju, Y., Shen, T. and Wang, D. (2020), "Bonding behavior between reactive powder concrete and normal strength concrete", *Constr. Build. Mater.*, **242**, 118024. https://doi.org/10.1016/j.conbuildmat.2020.118024

Kanevski, M., Pozdnukhov, A. and Timonin, V. (2008), "Machine learning algorithms for geospatial data. Applications and software tools", *Proceedings of the 4th International Congress on Environmental Modelling and Software*, Barcelona, Catalonia, Spain, July.

Kanungo, D.P., Sharma, S. and Pain, A. (2014), "Artificial Neural Network (ANN) and Regression Tree (CART) applications for the indirect estimation of unsaturated soil shear strength parameters", *Frontiers Earth Sci.*, **8**(3), 439-456. https://doi.org/10.1007/s11707-014-0416-0

Kiran, S., Lal, B. and Tripathy, S. (2016), "Shear strength prediction of soil based on probabilistic neural network", *Indian J. Sci. Technol.*, **9**(41). https://doi.org/10.17485/ijst/2016/v9i41/99188

Kuo, Y.L., Jaksa, M.B., Lyamin, A.V. and Kaggwa, W.S. (2009), "ANN-based model for predicting the bearing capacity of strip footing on multi-layered cohesive soil", *Comput. Geotech.*, **36**(3), 503-516. https://doi.org/10.1016/j.compgeo.2008.07.002

Li, T., Xu, M., Zhu, C., Yang, R., Wang, Z. and Guan, Z. (2019), "A deep learning approach for multi-frame in-loop filter of HEVC", *IEEE Transact. Image Process.*, **28**(11), 5663-5678. https://doi.org/10.1109/TIP.2019.2921877

Li, B.H., Liu, Y., Zhang, A.M., Wang, W.H. and Wan, S. (2020a), "A survey on blocking technology of entity resolution", *J. Comput. Sci. Technol.*, **35**(4), 769-793. https://doi.org/10.1007/s11390-020-0350-4

Li, C., Sun, L., Xu, Z., Wu, X., Liang, T. and Shi, W. (2020b), "Experimental investigation and error analysis of high precision FBG displacement sensor for structural health monitoring", *Int. J. Struct. Stabil. Dyn.*, **20**(06), 2040011. https://doi.org/10.1142/S0219455420400118

Liang, S., Foong, L.K. and Lyu, Z. (2020), "Determination of the friction capacity of driven piles using three sophisticated search schemes", *Eng. Comput.*, 1-13. https://doi.org/10.1007/s00366-020-01118-4

Liu, L., Li, J., Yue, F., Yan, X., Wang, F., Bloszies, S. and Wang, Y. (2018), "Effects of arbuscular mycorrhizal inoculation and biochar amendment on maize growth, cadmium uptake and soil cadmium speciation in Cd-contaminated soil", *Chemosphere*, **194**, 495-503. https://doi.org/10.1016/j.chemosphere.2017.12.025

Lo, H.C.J., Tabe, K., Iskander, M. and Yoon, S.H. (2010), "A transparent water-based polymer for simulating multiphase flow", *Geotech. Test. J.*, **33**(1), 1-13. https://doi.org/10.1520/GTJ102375

Lv, Z. and Qiao, L. (2020), "Deep belief network and linear perceptron based cognitive computing for collaborative robots", *Appl. Soft Comput.*, **92**, 106300. https://doi.org/10.1016/j.asoc.2020.106300

Lv, X., Li, N., Xu, X. and Yang, Y. (2020), "Understanding the emergence and development of online travel agencies: a dynamic evaluation and simulation approach", *Internet Res.*, **30**(6), 1783-1810. https://doi.org/10.1108/INTR-11-2019-0464

Ma, H.J. and Xu, L.X. (2020), "Decentralized adaptive fault-tolerant control for a class of strong interconnected nonlinear systems via graph theory", *IEEE Transact. Automatic Control*, **66**(7), 3227-3234. https://doi.org/10.1109/TAC.2020.3014292

Ma, H.J. and Yang, G.H. (2015), "Adaptive fault tolerant control of cooperative heterogeneous systems with actuator faults and unreliable interconnections", *IEEE Transact. Automatic Control*, **61**(11), 3240-3255. https://doi.org/10.1109/TAC.2015.2507864

Ma, H.J., Xu, L.X. and Yang, G.H. (2019), "Multiple environment integral reinforcement learning-based fault-tolerant control for affine nonlinear systems", *IEEE Transact. Cybernetics*, **51**(4), 1913-1928. https://doi.org/10.1109/TCYB.2018.2889679

Ma, H.J., Yang, G.H. and Chen, T. (2020), "Event-triggered optimal dynamic formation of heterogeneous affine nonlinear multi-agent systems", *IEEE Transact. Automatic Control*, **66**(2), 497-512. https://doi.org/10.1109/TAC.2020.2983108

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M. and Webster, S. (2000), "An investigation of machine learning based prediction systems", *J. Syst. Software*, **53**(1), 23-29. https://doi.org/10.1016/S0164-1212(00)00005-4

Melucci, M. (2009), "Weighted rank correlation in information retrieval evaluation", In: *Asia Information Retrieval Symposium*, pp. 75-86. https://doi.org/10.1007/978-3-642-04769-5_7

Moavenian, M.H., Nazem, M., Carter, J.P. and Randolph, M.F. (2016), "Numerical analysis of penetrometers free-falling into soil with shear strength increasing linearly with depth", *Comput. Geotech.*, **72**, 57-66. https://doi.org/10.1016/j.compgeo.2015.11.002

Moayedi, H., Mehrabi, M., Mosallanezhad, M., Rashid, A.S.A. and Pradhan, B. (2019), "Modification of landslide susceptibility mapping using optimized PSO-ANN technique", *Eng. Comput.*, **35**(3), 967-984. https://doi.org/10.1007/s00366-018-0644-0

Moayedi, H., Mehrabi, M., Bui, D.T., Pradhan, B. and Foong, L.K. (2020), "Fuzzy-metaheuristic ensembles for spatial assessment of forest fire susceptibility", *J. Environ. Manage.*, **260**, 109867. https://doi.org/10.1016/j.jenvman.2019.109867

Mohammadhassani, M., Nezamabadi-Pour, H., Suhatril, M. and Shariati, M. (2014), "An evolutionary fuzzy modelling approach and comparison of different methods for shear strength prediction of high-strength concrete beams without stirrups", *Smart Struct. Syst., Int. J.*, **14**(5), 785-809. https://doi.org/10.12989/sss.2014.14.5.785

Nehdi, M., El Chabib, H. and Said, A. (2006), "Evaluation of shear capacity of FRP reinforced concrete beams using artificial neural networks", *Smart Struct. Syst., Int. J.*, **2**(1), 81-100. https://doi.org/10.12989/sss.2006.2.1.081

Ni, Q., Hird, C.C. and Guymer, I. (2010), "Physical modelling of pile penetration in clay using transparent soil and particle image velocimetry", *Géotechnique*, **60**(2), 121-132. https://doi.org/10.1680/geot.8.P.052

Padmini, D., Ilamparuthi, K. and Sudheer, K.P. (2008), "Ultimate bearing capacity prediction of shallow foundations on cohesionless soils using neurofuzzy models", *Comput. Geotech.*, **35**(1), 33-46. https://doi.org/10.1016/j.compgeo.2007.03.001

Pham, B.T., Qi, C., Ho, L.S., Nguyen-Thoi, T., Al-Ansari, N., Nguyen, M.D., Nguyen, H.D., Ly, H.B., Le, H.V. and Prakash, I. (2020), "A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil", *Sustainability*, **12**(6), 2218. https://doi.org/10.3390/su12062218

Pourghasemi, H.R. and Rahmati, O. (2018), "Prediction of the landslide susceptibility: Which algorithm, which precision?", *Catena*, **162**, 177-192. https://doi.org/10.1016/j.catena.2017.11.022

Reddy, G.T., Reddy, M.P.K., Lakshmanna, K., Rajput, D.S., Kaluri, R. and Srivastava, G. (2020a), "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis", *Evolution. Intell.*, **13**(2), 185-196. https://doi.org/10.1007/s12065-019-00327-1

Reddy, T., Bhattacharya, S., Maddikunta, P.K.R., Hakak, S., Khan, W.Z., Bashir, A.K., Jolfaei, A. and Tariq, U. (2020b), "Antlion

re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset", *Multimedia Tools Applicat*., 1-25. https://doi.org/10.1007/s11042-020-09988-y

Rong, G., Alu, S., Li, K., Su, Y., Zhang, J., Zhang, Y. and Li, T. (2020), "Rainfall Induced Landslide Susceptibility Mapping Based on Bayesian Optimized Random Forest and Gradient Boosting Decision Tree Models—A Case Study of Shuicheng County, China", *Water*, **12**(11), 3066. https://doi.org/10.3390/w12113066

Samui, P. (2008), "Prediction of friction capacity of driven piles in clay using the support vector machine", *Can. Geotech. J.*, **45**(2), 288-295. https://doi.org/10.1139/T07-072

Sharma, S., Ahmed, S., Naseem, M., Alnumay, W.S., Singh, S. and Cho, G.H. (2021), "A Survey on Applications of Artificial Intelligence for Pre-Parametric Project Cost and Soil Shear-Strength Estimation in Construction and Geotechnical Engineering", *Sensors*, **21**(2), 463. https://doi.org/10.3390/s21020463

Stanier, S.A. (2012), "Modelling the behaviour of helical screw piles", Ph.D. Thesis; University of Sheffield, Department of Civil and Structural Engineering.

Sun, L., Yang, Z., Jin, Q. and Yan, W. (2020), "Effect of axial compression ratio on seismic behavior of GFRP reinforced concrete columns", *Int. J. Struct. Stabil. Dyn.*, **20**(06), 2040004. https://doi.org/10.1142/S0219455420400040

Vapnik, V. (2013), *The Nature of Statistical Learning Theory*, Springer Science & Business Media.

Xu, M., Li, T., Wang, Z., Deng, X., Yang, R. and Guan, Z. (2018), "Reducing complexity of HEVC: A deep learning approach", *IEEE Transact. Image Process.*, **27**(10), 5044-5059. https://doi.org/10.1109/TIP.2018.2847035

Xu, F., Foong, L.K. and Lyu, Z. (2020), "A novel search scheme based on the social behavior of crow flock for feed-forward learning improvement in predicting the soil compression coefficient", *Eng. Comput.*, 1-14. https://doi.org/10.1007/s00366-020-01119-3

Yan, J., Pu, W., Zhou, S., Liu, H. and Greco, M.S. (2020), "Optimal resource allocation for asynchronous multiple targets tracking in heterogeneous radar networks", *IEEE Transact. Signal Process.*, **68**, 4055-4068. https://doi.org/10.1109/TSP.2020.3007313

Yang, Y., Yao, J., Wang, C., Gao, Y., Zhang, Q., An, S. and Song, W. (2015), "New pore space characterization method of shale matrix formation by considering organic and inorganic pores", *J. Natural Gas Sci. Eng.*, **27**, 496-503. https://doi.org/10.1016/j.jngse.2015.08.017

Yang, J., Li, S., Wang, Z., Dong, H., Wang, J. and Tang, S. (2020a), "Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges", *Materials*, **13**(24), 5755. https://doi.org/10.3390/ma13245755

Yang, W., Zhao, Y., Wang, D., Wu, H., Lin, A. and He, L. (2020b), "Using principal components analysis and IDW interpolation to determine spatial and temporal changes of surface water quality of Xin'anjiang river in Huangshan, China", *Int. J. Environ. Res. Public Health*, **17**(8), 2942. https://doi.org/10.3390/ijerph17082942

Ye, X.W., Jin, T. and Yun, C.B. (2019), "A review on deep learning-based structural health monitoring of civil infrastructures", *Smart Struct. Syst., Int. J.*, **24**(5), 567-585. https://doi.org/10.12989/sss.2019.24.5.567

Ye, X., Moayedi, H., Khari, M. and Foong, L.K. (2020), "Metaheuristic-hybridized multilayer perceptron in slope stability analysis", *Smart Struct. Syst., Int. J.*, **26**(3), 263-275. https://doi.org/10.12989/sss.2020.26.3.263

Yue, H., Wang, H., Chen, H., Cai, K. and Jin, Y. (2020), "Automatic detection of feather defects using lie group and fuzzy fisher criterion for shuttlecock production", *Mech. Syst. Signal Process.*, **141**, 106690. https://doi.org/10.1016/j.ymssp.2020.106690

Zhang, C. and Wang, H. (2020), "Swing vibration control of suspended structures using the Active Rotary Inertia Driver system: Theoretical modeling and experimental verification", *Struct. Control Health Monitor.*, **27**(6), e2543. https://doi.org/10.1002/stc.2543

Zhang, K., Wang, Q., Chao, L., Ye, J., Li, Z., Yu, Z., Yang, T. and Ju, Q. (2019a), "Ground observation-based analysis of soil moisture spatiotemporal variability across a humid to semi-humid transitional zone in China", *J. Hydrol.*, **574**, 903-914. https://doi.org/10.1016/j.jhydrol.2019.04.087

Zhang, X., Nguyen, H., Bui, X.N., Tran, Q.H., Nguyen, D.A., Bui, D.T. and Moayedi, H. (2019b), "Novel soft computing model for predicting blast-induced ground vibration in open-pit mines based on particle swarm optimization and XGBoost", *Natural Resour. Res.*, **29**(2), 711-721. https://doi.org/10.1007/s11053-019-09492-7

Zhang, C., Abedini, M. and Mehrmashhadi, J. (2020a), "Development of pressure-impulse models and residual capacity assessment of RC columns using high fidelity Arbitrary Lagrangian-Eulerian simulation", *Eng. Struct.*, **224**, 111219. https://doi.org/10.1016/j.engstruct.2020.111219

Zhang, C., Gholipour, G. and Mousavi, A.A. (2020b), "State-of-the-art review on responses of RC structures subjected to lateral impact loads", *Arch. Computat. Methods Eng.*, **28**(4), 2477-2507. https://doi.org/10.1007/s11831-020-09467-5

Zhang, C., Gholipour, G. and Mousavi, A.A. (2020c), "Blast loads induced responses of RC structural members: State-of-the-art review", *Compos. Part B: Eng.*, 108066. https://doi.org/10.1016/j.compositesb.2020.108066

Zhang, W., Tang, Z., Yang, Y. and Wei, J. (2021), "Assessment of FRP–concrete interfacial debonding with coupled mixed-mode cohesive zone model", *J. Compos. Constr.*, **25**(2), 04021002. https://doi.org/10.1061/(ASCE)CC.1943-5614.0001114

Zhao, H., Ge, L. and Luna, R. (2010), "Low viscosity pore fluid to manufacture transparent soil", *Geotech. Test. J.*, **33**(6), 463-468. https://doi.org/10.1520/GTJ102607

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net", Journal of the royal statistical society: series B (statistical methodology), **67**(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

CC