Geoscience Frontiers 12 (2021) 441–452

HOSTED BY

Contents lists available at ScienceDirect

Geoscience Frontiers

journal homepage: www.elsevier.com/locate/gsf

Research Paper

Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms

Pin Zhang [a], Zhen-Yu Yin [a,*], Yin-Fu Jin [a], Tommy H.T. Chan [b], Fu-Ping Gao [c,d]

[a] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[b] School of Civil Engineering & Built Environment, Science and Engineering Faculty, Queensland University of Technology (QUT), Brisbane, Qld, 4001, Australia
[c] Key Laboratory for Mechanics in Fluid Solid Coupling Systems, Institute of Mechanics, Chinese Academy of Sciences, Beijing, 100190, China
[d] China School of Engineering Science, University of Chinese Academy of Sciences, Beijing, 100049, China

ARTICLE INFO

Keywords:
Compressibility
Clays
Machine learning
Optimization
Random forest
Genetic algorithm

ABSTRACT

Compression index C_c is an essential parameter in geotechnical design for which the effectiveness of correlation is still a challenge. This paper suggests a novel modelling approach using machine learning (ML) technique. The performance of five commonly used machine learning (ML) algorithms, i.e. back-propagation neural network (BPNN), extreme learning machine (ELM), support vector machine (SVM), random forest (RF) and evolutionary polynomial regression (EPR) in predicting C_c is comprehensively investigated. A database with a total number of 311 datasets including three input variables, i.e. initial void ratio e_0, liquid limit water content w_L, plasticity index I_p, and one output variable C_c is first established. Genetic algorithm (GA) is used to optimize the hyper-parameters in five ML algorithms, and the average prediction error for the 10-fold cross-validation (CV) sets is set as the fitness function in the GA for enhancing the robustness of ML models. The results indicate that ML models outperform empirical prediction formulations with lower prediction error. RF yields the lowest error followed by BPNN, ELM, EPR and SVM. If the ranges of input variables in the database are large enough, BPNN and RF models are recommended to predict C_c. Furthermore, if the distribution of input variables is continuous, RF model is the best one. Otherwise, EPR model is recommended if the ranges of input variables are small. The predicted correlations between input and output variables using five ML models show great agreement with the physical explanation.

1. Introduction

Compressibility of soils is described using the compression index C_c, which is generally determined by the oedometer test in geotechnical design. The accurate prediction of C_c would facilitate the understanding of soil volume change and it is of significance to calculate consolidation settlement in engineering practice such as foundation (Yang et al., 2019), tunneling (Shen et al., 2014; Wu et al., 2020; Zhang et al., 2020a), embankment (Yin et al., 2015; Zhu et al., 2020). Compared with natural clays, properties of reconstituted clays represent the inherent properties, because they are inherent to the soil and independent of the natural deposits with chemical environment (Hong et al., 2010; Yin et al., 2011; Zhu et al., 2016; Yin et al., 2017a). The properties of reconstituted clays thus provide a basis for understanding the in situ state of natural clays and the influence of structure on its in situ properties (Burland, 1990). Therefore, this study merely focuses on the compressibility of reconstituted clays.

Numerous empirical correlations have been proposed to predict C_c based on influential or state parameters of the soils. A linear correlation between C_c and liquid limit water content w_L proposed by Skempton and Jones (1944) must be the most widely accepted one. Thereafter, C_c predictive correlations based on plasticity index I_p (Wroth and Wood, 1978; Sridharan and Nagaraj, 2000; Nath and DeDalal, 2004; Tiwari and Ajmera, 2012), shrinkage limit I_s (Sridharan and Nagaraj, 2000), void ratio at the liquid limit e_L (Nagaraj and Murthy, 1983, 1986; Burland, 1990) were proposed, as summarized in Table 1.

Recently, machine learning (ML) algorithms have been extensively used to develop soil properties prediction models and improve the prediction accuracy in comparison with conventional empirical formulations because of their capability of capturing the non-linear

* Corresponding author.
E-mail address: zhenyu.yin@polyu.edu.hk (Z.-Y. Yin).
Peer-review under responsibility of China University of Geosciences (Beijing).

https://doi.org/10.1016/j.gsf.2020.02.014
Received 13 September 2019; Received in revised form 9 February 2020; Accepted 23 February 2020
Available online 21 March 2020

Table 1
Existing C_c empirical formulations for reconstituted clays.

Variable	Formulation	Reference
w_L	$C_c = 0.007(w_L - 10)$	Skempton and Jones (1944)
I_p	$C_c = 0.5I_pG_s$	Wroth and Wood (1978)
	$C_c = 0.014(I_p + 3.6)$	Sridharan and Nagaraj (2000)
	$C_c = 0.015I_p - 0.0198$	Nath and DeDalal (2004)
	$C_c = 0.0014I_p$	Tiwari and Ajmera (2012)
I_s	$C_c = 0.007(I_s + 18)$	Sridharan and Nagaraj (2000)
e_L	$C_c = 0.2237e_L$	Nagaraj and Murthy (1983)
	$C_c = 0.2343e_L$	Nagaraj and Murthy (1986)
	$C_c = 0.0256e_L - 0.04$	Burland (1990)

Note: w_L = liquid limit; G_s = specific gravity; I_p = plasticity index; I_s = shrinkage limit; e_L = void ratio at the liquid limit.

relationships among high-dimensional variables (Chen et al., 2019a, b; Zhang et al., 2019). Applications include the prediction of water content (Arsoy et al., 2013; Zhou et al., 2016), temperature (Kundu et al., 2017; Feng et al., 2019), creep index (Zhang et al., 2020b), shear strength (Pham et al., 2018), unconfined compression strength (Gunaydin et al., 2010; Ghorbani and Hasanzadehshooiili, 2018), cyclic behavior (Zhang et al., 2020c) and assessment of soil liquefaction (Alavi and Gandomi, 2012). Overall, ML algorithms used in these research works cover back-propagation neural network (BPNN), extreme learning machine (ELM), support vector machine (SVM), random forest (RF) and evolutionary polynomial regression (EPR). In regard to the application of ML algorithms in predicting C_c, Park and Lee (2011) first adopted BPNN for predicting C_c, Yin et al. (2016b) proposed an integrated EPR and real-coded genetic algorithm, and Kirts et al. (2018) developed a model based on SVM. Hitherto, only these three ML algorithms have been tried for predicting C_c. Nevertheless, the performance of ML algorithms in a prediction issue is different and there is no unique theory to identify which one is the optimum algorithm. The application of other ML algorithms except the BPNN, EPR and SVM may improve the C_c predictive accuracy and further facilitates understanding of the correlations between C_c and influential factors. Hence, a comprehensive study of different ML algorithms in predicting C_c is worth investigating.

This study focuses on the comparison of the performance of five commonly used ML algorithms, i.e. BPNN, ELM, SVM, RF and EPR, in predicting C_c of reconstituted clays. A database including various reconstituted clays is first established. Meta-heuristic genetic search algorithm is employed to optimize hyper-parameters of five ML algorithms. The average prediction errors for the 10-fold cross-validation (CV) sets are used as the fitness function in the GA. The performance of ML algorithms is also compared with existing empirical prediction formulations of C_c. The correlations between influential factors and C_c using ML models are particularly investigated.

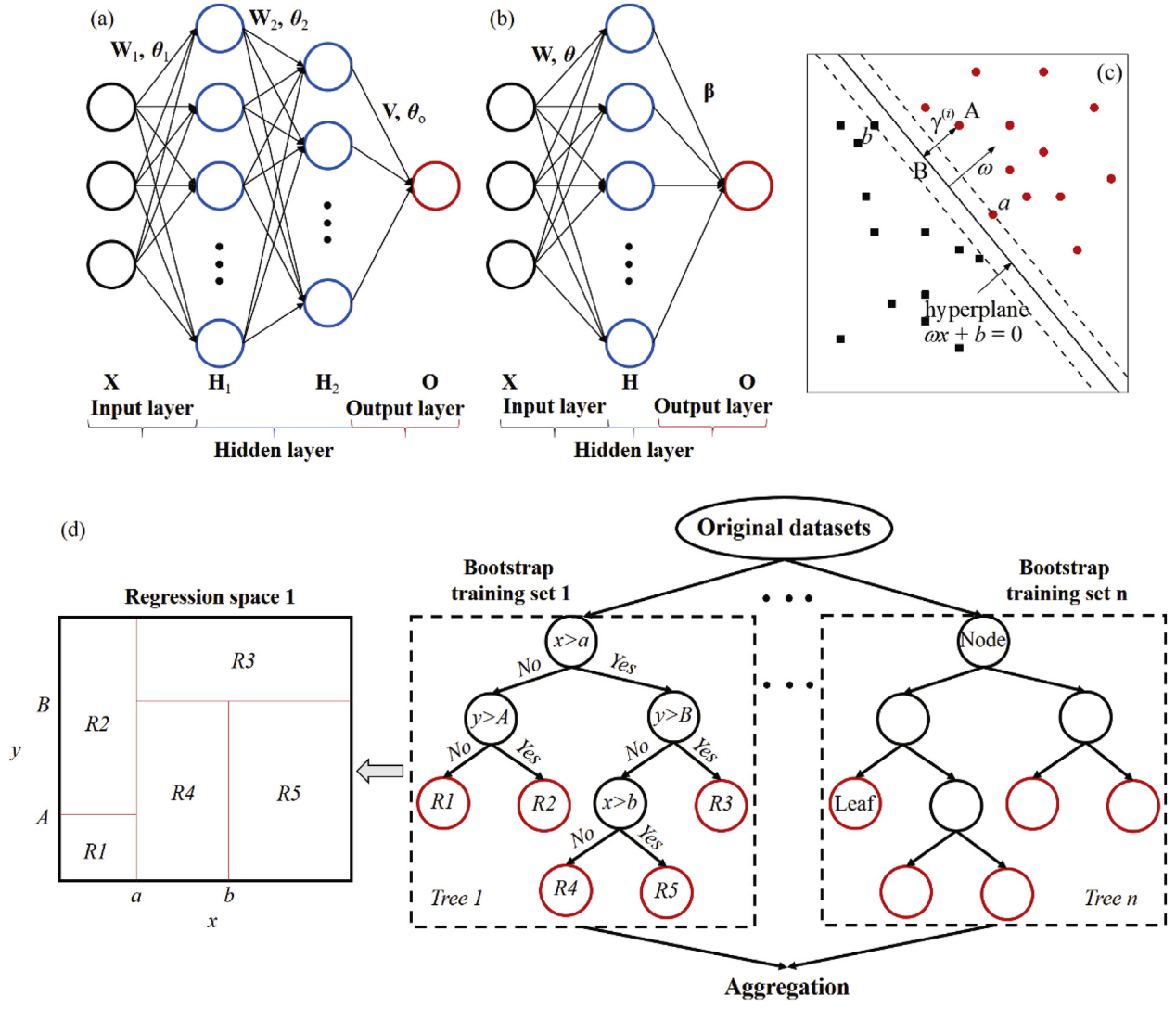

Fig. 1. Schematic view of ML algorithms for: (a) BPNN; (b) ELM; (c) SVM; (d) RF.

2. Methodology

2.1. Machine learning algorithms

2.1.1. Back-propagation neural network

Back-propagation neural network (BPNN) is a type of feedforward neural network, characterized by propagating errors from the output layer for finding a set of weights that ensure that the output value produced by the network is the same as the actual output value (Rumelhart et al., 1986). The generalization ability of BPNN is strong, but the determination of the values of numerous parameters including weights, biases and hyper-parameters is a hard task. A BPNN consists of an input layer, any number of hidden layers and an output layer, as shown in Fig. 1a. The performance of BPNN depends on the number of hidden layers and hidden neurons. If they are fixed, the values of weights and bias can be determined by gradient descend. The outputs of the hidden and output layers are expressed as:

$$\mathbf{H}_1 = f(\mathbf{W}_1\mathbf{X} + \boldsymbol{\theta}_1) \tag{1}$$

$$\mathbf{O} = g(\mathbf{V}\mathbf{H} + \boldsymbol{\theta}_o) \tag{2}$$

where, $\mathbf{H}$ = the hidden layer output matrix; $\mathbf{X}$, $\mathbf{O}$ = actual input and output matrix, respectively; $\mathbf{W}_1$, $\mathbf{V}$ = weights matrix on the connections between input and hidden neurons, between hidden and output neurons, respectively; $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_o$ = bias vectors on the connections between input and hidden neurons, between hidden and output neurons, respectively. f, g = activation functions in hidden and output layers, respectively, that are, *tansig* and *purlin* in this study, which can be formulated as:

$$tansig: f(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{3}$$

$$purlin: g(x) = x \tag{4}$$

2.1.2. Extreme learning machine

Extreme learning machine (ELM) is a modification of the single-hidden layer feedforward neural network, as shown in Fig. 1b. The weights of input layer and the biases of hidden layer are assigned randomly, and the weights of the output layer are determined analytically through simple generalized inverse operation of the hidden layer output matrices (Huang et al., 2006), thereby the learning speed can be thousands of times faster than traditional feedforward network algorithms, but the generalization ability is sacrificed (Chen et al., 2019b). ELM can be represented as:

$$\mathbf{H} = f(\mathbf{W}\mathbf{X} + \theta) \tag{5}$$

$$\min_{\boldsymbol{\beta}} \|\mathbf{H}\boldsymbol{\beta} - \mathbf{O}\| \tag{6}$$

where, $\mathbf{H}$ = the hidden layer output matrix; $\mathbf{X}$, $\mathbf{O}$ = actual input and output matrix, respectively; $\boldsymbol{\beta}$ = weight matrix connecting the hidden and the output layers. The learning process of ELM algorithm is achieved by calculating $\boldsymbol{\beta}$.

2.1.3. Support vector machine

Support vector machine (SVM) develops upon structural risk minimization, thereby it can be used to train model with small datasets and the computational complexity depends on the number of support vector rather the number of input parameters, but computational cost of SVM is expensive for the training of numerous datasets. Datasets are mapped to a high-dimension space by a kernel trick, where a linear decision surface or hyperplane is constructed (Cortes and Vapnik, 1995), as shown in Fig. 1c. In this figure, $\gamma^{(i)}$, which is term as geometric margin, donates the distance of training sample ($x^{(i)}$, $y^{(i)}$) to the decision boundary, and it is orthogonal to the hyperplane. For all samples in the training set, the smallest geometric margin is represented by:

$$\gamma = \min_{i=1,2,\ldots,m} \gamma^{(i)} \tag{7}$$

The optimal SVM classifier is the one which can separate the positive and negative points of the training set with a largest "gap", that is, γ reaches the maximum value, which can be obtained by:

$$\begin{gathered} \min_{\gamma,\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} y(i)((\omega)^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, m, \ \xi_i \geq 0 \end{gathered} \tag{8}$$

where, m = a total number of training samples; ω, b = weights and biases; ξ = slack parameter; C = penalty parameter. Radial basis function (RBF) kernel is utilized in this research for mapping the inputs to a high-dimension space, which is formulated as:

$$\text{RBF}: K(x, z) = \exp\left(-\gamma\|x - z\|^2\right) \tag{9}$$

2.1.4. Random forest

Random forest (RF) is an ensemble learning algorithm with the integration of bootstrap aggregating (Breiman, 1996) and random subspace (Ho, 1998) methods. The prediction performance of RF is strong due to the integration of numerous decision trees, but the output of RF is limited to the combination of values of output parameters (Zhang, 2019). In bagging, n bootstrap sets are made by sampling with replacement N training examples from the training set. The number of samples in the bootstrap training set is arbitrary, less than the original training set. Thereafter each bootstrap set is used to build a decision tree, as shown in Fig. 1d. Each node in a decision tree represents a classification criterion and the leaves of the tree represents the output labels, thereby a decision tree classifies a bootstrap training sample by testing random features at each node. Consequently, a regression space can be determined. The ultimate result can be obtained by aggregating the outputs of all trees (Liaw and Wiener, 2002), which can be expressed as:

$$y = \frac{1}{n}\sum_{i=1}^{n} y_i(x) \tag{10}$$

where, $y_i(x)$ = predicted output of a tree for an input vector x.

2.1.5. Evolutionary polynomial regression

Evolutionary polynomial regression (EPR) is a type of genetic programming including two-stage technique for constructing symbolic models: (i) structure identification, and (ii) parameter estimation (Giustolisi and Savic, 2006). In the first stage, genetic algorithm (GA) is used to search for symbolic structures of EPR and parameters values are estimated by solving a Least Squares (LS) linear problem in the second stage. Unlike the ML algorithms mentioned above, a simple formulation is particularly presented in EPR algorithm for describing the correlation between input and output variables, which is convenient to be used, but EPR is not suitable to solve high-dimensional and complex problems. A general EPR expression can be formulated as:

$$y = \sum_{j=1}^{m} a_j \cdot z_j + a_0 \tag{11}$$

where, y = predicted output; z_j = jth transformed variable; a_j = an adjustable parameter for the jth term; a_0 = an optional bias.

The key objective of the EPR is to search the best form of the function, i.e. the number of transformed variables and a combination of vectors of independent input variables. Herein, the transformed variable is obtained by:

$$z_j = x_1^{\mathbf{ES}(j,1)} \cdot \ldots \cdot x_i^{\mathbf{ES}(j,i)} \cdot \ldots \cdot x_k^{\mathbf{ES}(j,k)} \tag{12}$$

where, x_i = ith input variable; k = a total number of input variables; $\mathbf{ES}_{m\times k}$ = exponents matrix, it is determined by GA. Thereafter, the adjustable parameters and an optional bias can be determined by the least square regression.

2.2. Genetic algorithm

Genetic algorithm (GA) is a meta-heuristic search and optimization technique inspired by nature evolution (Holland, 1975). GA is one of the most accepted global optimization algorithm and has been extensively utilized in geotechnical engineering such as identification of parameters of constitutive models (Jin et al., 2016b, 2017; Yin et al., 2016a, 2017b), model selection (Jin et al., 2016a), slope (Tran and Srokosz, 2010; Liu et al., 2019), embankment (Guo et al., 2018; Müthing et al., 2018), tunneling (Koopialipoor et al., 2017; Liu and Liu, 2019), pile foundation (Jin et al., 2018a, 2018b), excavation (Jin et al., 2019). Therefore, the GA is employed to optimize hyper-parameters in this study. Fig. 2 presents the flowchart of the GA algorithm. GA starts from generating a population of individuals. Each individual is represented by a chromosome based on a coding scheme (real-coded GA). The performance of each individual can be evaluated by the fitness value. The best individuals in the population – those with lowest fitness value – are selected and then modified through crossover and mutation operations at each generation. A new population is thus created. The process continues until satisfies the termination condition, that is, whether or not reaches the maximum generation and fitness value converges at the constant value (100 generations for BPNN, SVM, RF and EPR, 1000 generations for ELM due to the slow converge rate in this study). Common genetic operators used in GA, roulette wheel selection and real valued recombination methods are adopted in this research. The values of parameters used in GA is presented in Table 2.

2.3. Evaluation indicators

To evaluate the performance of ML models, three commonly used indicators "Mean Absolute Error (MAE)", "Ranking Distance (RD)" and "Nash–Sutcliffe model Efficiency coefficient (NSE)" are adopted. MAE is an unbiased measure to evaluate the average prediction error of model. RD and NSE can be used to assess the accuracy and precise of model (Nash and Sutcliffe, 1970; Orr and Cherubini, 2003). The combination of such three indicators enables to comprehensively evaluate model performance. The expression of these three measures can be obtained by

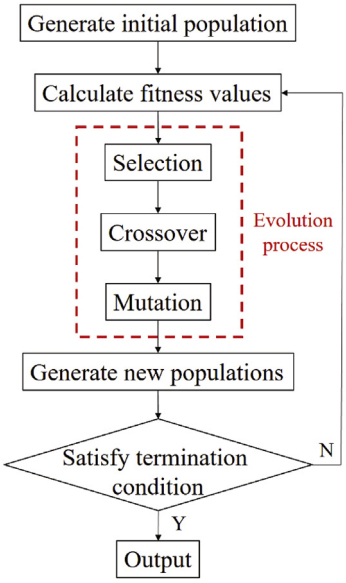

Fig. 2. Flowchart of GA algorithm.

Table 2
Values of parameters in the GA algorithm.

Algorithm	p_{cross}	$p_{mutation}$	Population	Generation
GA	0.7	0.1	20	100/1000

Note: 100 = maximum generation in BPNN, SVM, RF and EPR; 1000 = maximum generation in ELM.

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|r_i-p_i| \tag{13}$$

$$\mathrm{RD}=\sqrt{\left[1-\mu\left(\frac{p_i}{r_i}\right)\right]^2+\left(\delta\frac{p_i}{r_i}\right)^2} \tag{14}$$

$$\mathrm{NSE}=1-\frac{\sum_{i=1}^{n}(p_i-r_i)^2}{\sum_{i=1}^{n}(r_i-\bar{r})^2} \tag{15}$$

where, r = measured output value; p = predicted output value; $\bar{r}$ = mean of measured output values; n = a total number of datasets; μ = mean value of p_i/r_i; δ = standard deviation of p_i/r_i. Low values of MAE and RD, high value of NSE indicate a model with great performance.

2.4. K-fold cross validation

The whole process of establishing a ML model includes three phases: training, validation and test. The objective of validation is to improve the robustness of training models and avoid overfitting, the training model is thus more reliable for the test set. Herein, k-fold cross-validation (CV) method has been extensively used to validate model (Stone, 1974). In this method, the original training set is randomly divided into k sub-datasets. K–1 sub-datasets are used to train models and a remaining sub-dataset is used to validate models. Each sample thus has opportunity to train and validate models. K was set as 10 in this study according to the research conducted by Kohavi (1995).

At each round, ML models with a fixed set of hyper-parameters will be trained ten times with random nine sub-datasets as the training set, and the performance of ML models will be evaluated by the mean prediction error for the remaining one sub-dataset, which can be formulated by:

$$\mathrm{Fitness}=\frac{1}{10}\sum_{i=1}^{10}\mathrm{MAE}_i \tag{16}$$

where, MAE_i = prediction error for the ith validation set. Eq. (16) is defined as the fitness function in the GA algorithm.

3. Hybrid meta-heuristic and machine learning algorithms

3.1. Model framework

Fig. 3 presents the proposed process of establishing hybrid meta-heuristic and ML-based C_c prediction models. A database including influential factors and C_c is first formed. The selection of input variables is vitally important to the model performance, and the correlation of selected parameters to C_c will be examined by grey relational analysis (GRG). GRG can account for the geometric similarity of the time series of the two parameters, and a large GRG value indicates a strong correlation exists between two parameters, thereby it has been extensively applied to evaluate uncertain correlations among parameters (Jiang and He, 2012; Li and Chen, 2019). Given a reference sequence $x_r = x_r\,(x_r(1), x_r(2), \ldots, x_r(n))$ and a compared sequence $x_i = x_i\,(x_i(1), x_i(2), \ldots, x_i(n))$, the grey relational coefficient between two sequences at the jth (j = 1, 2, …, n) criterion can be obtained by:

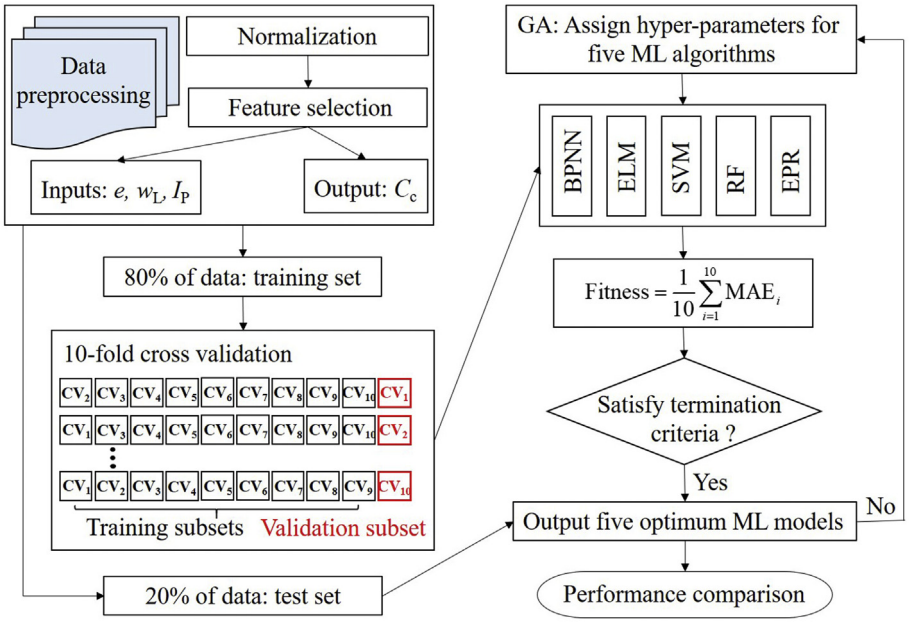

Fig. 3. Flowchart of proposed hybrid meta-heuristic and ML C_c prediction model.

$$\gamma(x_r(j),x_i(j))=\frac{\min\limits_i\min\limits_j|x_r(j)-x_i(j)|+\xi\max\limits_i\max\limits_j|x_r(j)-x_i(j)|}{|x_r(j)-x_i(j)|+\xi\max\limits_i\max\limits_j|x_r(j)-x_i(j)|} \tag{17}$$

where ξ = the resolving coefficient in the range [0, 1], usually considered to be 0.5. The GRG between sequences x_r and x_i is defined by:

$$\gamma(x_r,x_i)=\frac{1}{n}\sum_{j=1}^{n}\gamma(x_r(j),x_i(j)) \tag{18}$$

To train a well-performed ML based model, it is necessary to collect numerous datasets. Therefore, this study uses three parameters which can be extensively collected and exhibit strong relationships with C_c, i.e., e_0, w_L and I_p. It should be noted that the mineralogical composition of each sample and the void ratio at liquid limit are the parameter most directly related to intrinsic compression index (Giasi et al., 2003; Cerato and Lutenegger, 2004; Tiwari and Ajmera, 2012; Cao et al., 2018; Habibbeygi et al., 2018). However, the datasets including the void ratio at liquid limit are limited, which hinders to develop a ML based model involving the void ratio at liquid limit with excellent generalization ability. It has been reported that liquid limit w_L and plasticity index I_p exhibit clear relationships with C_c and have been used in empirical formulations (Skempton and Jones, 1944; Tiwari and Ajmera, 2012). In addition, the compressibility behavior is directly affected by initial void ratio e_0 (Nagaraj and Murthy, 1983; Burland, 1990; Tiwari and Ajmera, 2011). Herein, 80% of data are randomly selected for training model while the remaining are used to test model. All the datasets are first mapped to the interval (−1, 1) using Eq. (22). In this way, the computation cost can decrease dramatically and the different magnitude of input variables can be eliminated.

$$x_{norm}=\frac{x-x_{min}}{x_{max}-x_{min}}(\overline{x}_{max}-\overline{x}_{min})+\overline{x}_{min} \tag{19}$$

The objective at the process of training model is to identify the optimum hyper-parameters in five ML models. Table 3 summarizes the hyper-parameters in five ML algorithms and the range of these hyper-parameters. One hidden layer in BPNN is sufficient enough considering merely three input variables. The ranges of hidden neurons in BPNN and ELM are determined according to the published research works as listed in Table 4. The ranges of hyper-parameter in SVM and RF are large enough to determine the optimum values. The maximum number of terms in the EPR is set as 8 considering that the optimum number of terms is 5 in C_c prediction model proposed by Yin et al. (2016b) and too many terms can cause overfitting. At each round, GA randomly assigns hyper-parameters to ML algorithms. The performance of ML algorithms with this set of hyper-parameters will be evaluated by the fitness value. If the termination condition is satisfied, the optimum hyper-parameters in one ML algorithm can be determined. Otherwise, GA will assign a new set of hyper-parameters to ML algorithms. After the optimum hyper-parameters of five ML algorithms are determined, the performances of five ML C_c prediction models will be comprehensively compared by the test set.

Table 3
Hyper-parameters in five ML algorithms.

Algorithm	Hyper-parameters	Description	Range
BPNN	*l*_hidden layers	Number of hidden layers	1
	*n*_neurons	Number of neurons in hidden layers	1–8
ELM	*n*_neurons	Number of neurons in hidden layers	1–8
SVM	*C*_penalty	Penalty parameter of the error term	0–200
	*g*_width index	Width index of kernel functions	0–200
RF	*n*_tree	Number of trees in the forest	1–200
	*m*_try	Number of features for splitting at each node	1–3
EPR	*z*_transformed term	Number of transformed variables	1–8

Table 4
Methods for determining the number of hidden neurons.

Method	Reference	Number of neurons
$\leq 2N_i+1$	Nielsen (1987)	7
$\frac{2+N_i\times N_o+0.5N_o\times(N_o^2+N_i)-3}{N_o+N_i}$	Paola (1994)	1
$2N_i/3$	Wang (1994)	2
$\sqrt{N_i\times N_o}$	Masters (1994)	2
$2N_i$	Kaastra and Boyd (1996)	6

Note: N_i = the number of input variables; N_o = the number of output variables.

3.2. Data analysis

Compressibility tests for determining C_c have been conducted by numerous studies. To develop a general C_c prediction model for all reconstituted clays, the data used in this research were collected from various reconstituted clays (including silty clays) in the world. A total number of 331 datasets were particularly collected, and the statistical properties of these datasets are presented in Table 5. The diagonal line in Fig. 4 presents the histograms of all variables in the database and the values of mean and standard deviation (SD). Scatter plots of pairwise variables are also plotted in this figure. It can be observed that all variables cover a wide range of values and roughly presented a lognormal distribution, which sufficiently extend the applicability of the proposed model.

Grey relational grade (GRG) has been extensively employed to evaluate uncertain correlations among variables in a system (Jiang and He, 2012; Li and Chen, 2019). Table 6 presents the GRG values among input and output variables. It can be observed that w_L has the largest GRG value with 0.88, followed by I_p with 0.85 and e_0 with 0.81, respectively. Overall, GRG values of all input variables are much larger than 0.8, showing high correlations between input and output variables. It indicates that the selected influential factors are appropriate to predict C_c.

4. Results

4.1. Determination of hyper-parameters

Fig. 5 shows the evolution of the fitness values in five ML algorithms. It can be observed that the evolution of fitness values is obviously different and ultimately converges at various values. The maximum generation in BPNN is 72. The convergence value decreases with the increase in the hidden neurons, but it holds steadily when the number of neurons exceeds 7. Hence, the optimum number of hidden neurons in BPNN is here identified as 7. In ELM, the convergence rate is relatively slow, which is attributed to the principle of ELM as mentioned before. The performance of ELM depends on the weights of input layer and biases of hidden layer, thereby the parameters which need to be optimized are numerous, causing slow convergence rate. Overall, the variation in the fitness value can be negligible when the generation exceeds 500. The optimum number of hidden neurons in ELM is here identified as 3. In contrast, the fitness values of SVM and RF keep constant from the initial generation, which is due to the fact that only two parameters need to be optimized in these two algorithms. In EPR, the maximum generation is 67. The convergence value decreases with the increase in the number of terms and reaches minimum value with six terms, then starts to increase with the continuous increase in the number of terms.

Once the fitness value converges at a steady value, the corresponding hyper-parameters are defined as the optimum values. Table 7 summarizes the optimum hyper-parameters, convergence values and computational cost in five ML algorithms. It can be observed that the minimum convergence value (0.003265) appears in the RF algorithm, followed by BPNN (0.04163), ELM (0.04252), SVM (0.06657) and EPR (0.08291). Compared with other ML algorithms, EPR presents a clear C_c prediction formulation. The optimum EPR formulation with six terms in this study could then be expressed as:

$$C_c = 0.185419 + 158.8564\frac{1}{w_L^2 I_p} - 2.23074\frac{1}{e_0^2 w_L I_p} + 0.30408\left(\frac{I_p}{w_L}\right)^2 + 0.030177e_0^2 + 0.0000792\frac{w_L^2}{e_0} - 0.00461\frac{w_L}{e_0} \quad (20)$$

Table 5
Statistical properties of parameters.

Parameter	Max.	Min.	Mean
e_0	4.643	0.663	2.18
w_L	166.2	25	67.95
I_p	113.2	0.23	31.30
C_c	0.12	1.34	0.46

4.2. Prediction of C_c for the validation and test sets

To reveal the reason behind the difference in the convergence value in five ML prediction models, Fig. 6 presents the distribution of MAE values for the ten CV sets in five ML algorithms. It can be observed that the evolution of MAE in five ML algorithms is roughly identical. The ranges of MAE in BPNN, ELM and RF are less than that in SVM and EPR. Meanwhile, in ten CV sets, the corresponding MAE values produced by BPNN, ELM and RF are less than that in SVM and EPR. Furthermore, the maximum MAE value of outlier is 0.29131 appeared in EPR, Followed by SVM, the maximum MAE value of the outlier is 0.21774. These factors cause the convergence values in SVM and EPR are obviously larger than that in BPNN, ELM and RF.

Once the optimum hyper-parameters are determined in five ML algorithms. The test set will be used to evaluate the feasibility and applicability of these models. Fig. 7 presents the scatter plots of the predicted C_c for training and testing sets using five optimum models, and the values of three performance indicators are also presented in the figures. Such values are summarized in Table 8. It is clear that RF model outperforms the remaining four ML models with the highest value of NSE and the lowest values of MAE and RD. The predicted C_c for the training set show perfect agreement with the measured C_c, in which both MAE and RD are approximately identical to zero and the NSE is equal to 1. The predicted C_c for the test set is also close to the P = M line. The MAE and RF values for the testing set, 0.0143 and 0.08, respectively, are much less than that generated by remaining four ML algorithms. The predicted C_c for the training and test sets using BPNN also exhibits excellent agreement with the measured C_c with small values of MAE and RD. The predicted C_c scatters around the P = M line, and the NSE value for the testing set generated by BPNN is the lowest among five ML algorithms. Overall, regarding the testing set, NSE values produced by five optimum ML models are roughly identical; RF yields the lowest MAE and RD values, followed by BPNN, ELM, EPR and SVM. Due to the scarce datasets for the C_c larger than 0.8, the predicted C_c using ELM, SVM and EPR based models obviously deviates from the measured C_c as the measured C_c exceeds 0.8, whereas BP and RF based models still present high accuracy for predicting large C_c. Such factors indicate the generalization ability of BP and RF based models outperform ELM, SVM and EPR based models.

5. Discussions

5.1. Comparison with empirical formulations

To compare the predictive ability of ML models with the empirical formulations, four commonly used empirical formulations are used for predicting C_c. Fig. 8 presents scatterplot of the predicted C_c for the test set using four empirical formulations, and the relevant correlations. MAE, RD and NSE values are also included in the figure. It can be observed that the predicted C_c is widely distributed and deviates from the P = M line. The prediction error using empirical formulation proposed by Skempton and Jones (1944) is lowest with MAE = 0.1065, RD = 0.35 and NSE = 0.61, but they are much larger than the prediction error of five ML-based models.

5.2. Parametric investigation

A robust ML model exhibits smooth functions to describe the correlations between input and output variables, and exhibits physical explanation for these correlations (Shahin et al., 2005). Therefore, the correlations between three input variables and C_c in five optimum ML

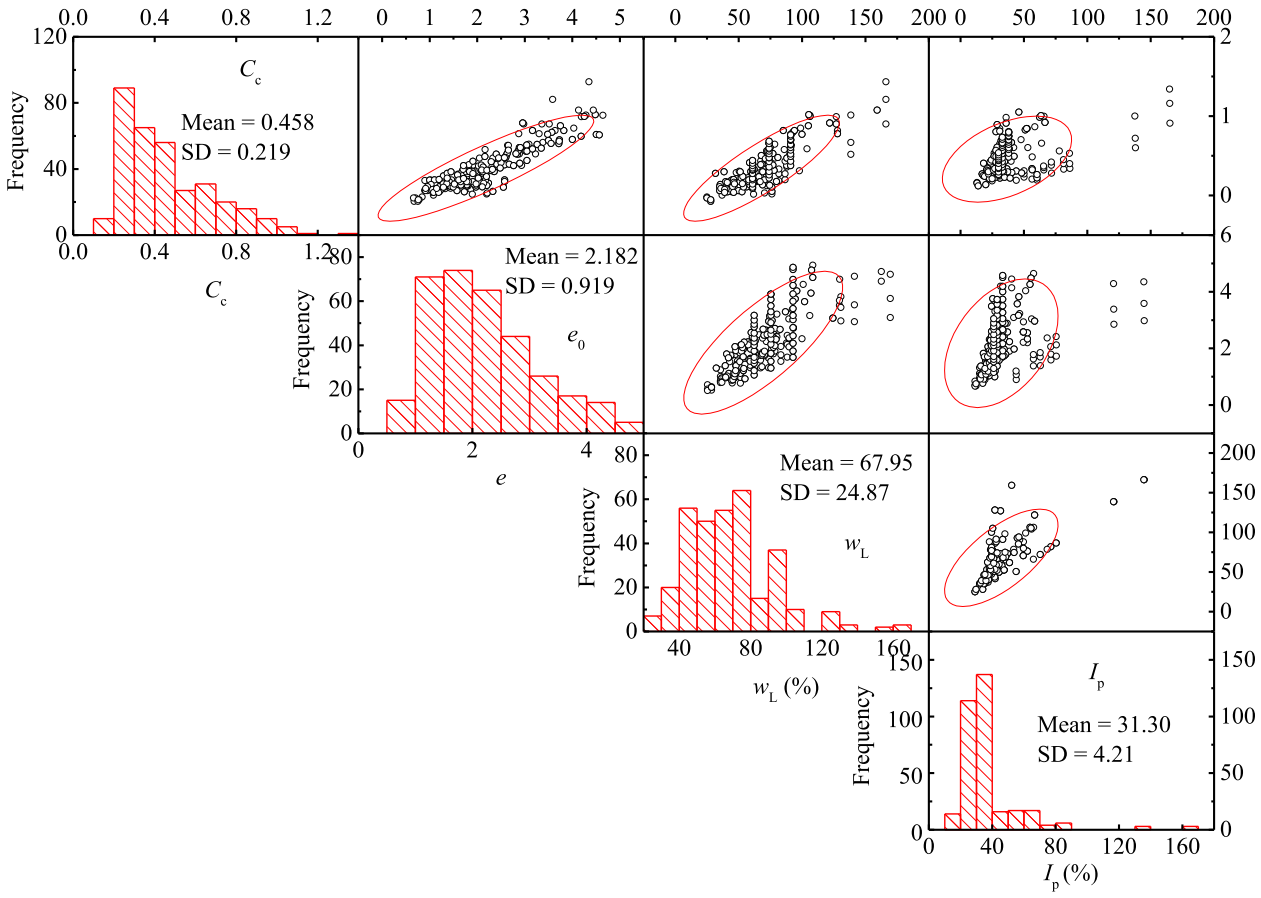

Fig. 4. Distribution of input and output variables.

Table 6
GRG values between input and output variables.

Variables	e_0	w_L	I_p
GRG	0.81	0.88	0.85

algorithms are investigated, as shown in Fig. 9. Herein, the values of studied parameter vary from the minimum to maximum, and the other parameters maintain at mean values. The smooth correlations between three input variables and C_c can be observed in ML models except RF, where the variation of predicted C_c is dramatic, although the general trend is similar to other ML models. This is attributed to the principle of RF algorithm. The performance of the RF algorithm depends on the classification conditions (values of variables) at each node as mentioned in the *Random forest* section. Furthermore, the classification conditions are affected by the distribution of input variables. Continuous distribution of input variables will generate continuous values of classification conditions, whereas the discrete distribution will generate discrete values of classification conditions. It means that the predicted results using RF may wrongly enter output labels if the values of new datasets do not appear in the original database, because the classification conditions do not include the values of new datasets. It can be observed from Fig. 4 that the distribution of e_0 is much more continuous than w_L and I_p. Therefore, in Fig. 9, the correlation between e and predicted C_c is much smoother than the remaining variables w_L and I_p.

From the perspective of other four ML algorithms, it can be observed from Fig. 9a that C_c increases monotonically with the increase in e_0. The difference among four ML models can be negligible when the e_0 is less than 4. When the e_0 is larger than 4, the increase in the C_c in the EPR model is obviously larger than the remaining three ML models. In Fig. 9b, C_c initially decreases with the increase in the w_L, after reaching the minimum value, C_c starts to increase with the continuous increase in the w_L. In Fig. 9c, the predicted C_c roughly increases monotonically with the increase in I_p. Similar to the correlation between C_c and e_0, the increase in the C_c in the RF model is much larger than that in the remaining three ML models. Overall, the correlations presented in Fig. 9 are consistent with physical explanation, indicating robustness and reasonability of the proposed RF models.

To investigate the generalization ability of five optimum models, a database including a total number of 10,000 random samples is established, in which it is assumed that each variable complies with lognormal distributions (Zhang et al., 2009; Cao and Wang, 2014; Zhang et al., 2018). Herein, the values of mean and standard deviation for each variable are consistent with the measured values presented in Fig. 4. Fig. 10 shows the distribution of predicted C_c using five optimum ML models. It can be observed that the predicted C_c presents a clear lognormal distribution except the RF model, where numerous predicted C_c falls into the range of 0.45–0.525. The reason behind this is similar to the correlations between the input and out variables in the RF model. In EPR model, the maximum predicted C_c reaches 3.9, severely losing reliability, because the increase in the input variables causes the continuous increase in the C_c in EPR model as mentioned before. In contrast, the predicted C_c using the remaining four ML algorithms does not exceed the range of C_c in the original database. The distribution of predicted C_c using BPNN, ELM and SVM is roughly identical. Overall, the performance of ML models is reliable for the unseen datasets, and the values of mean and standard error for 10,000 random samples generated by five ML models are roughly equal to the values of mean and standard error of the measured C_c (0.458 and 0.219, respectively).

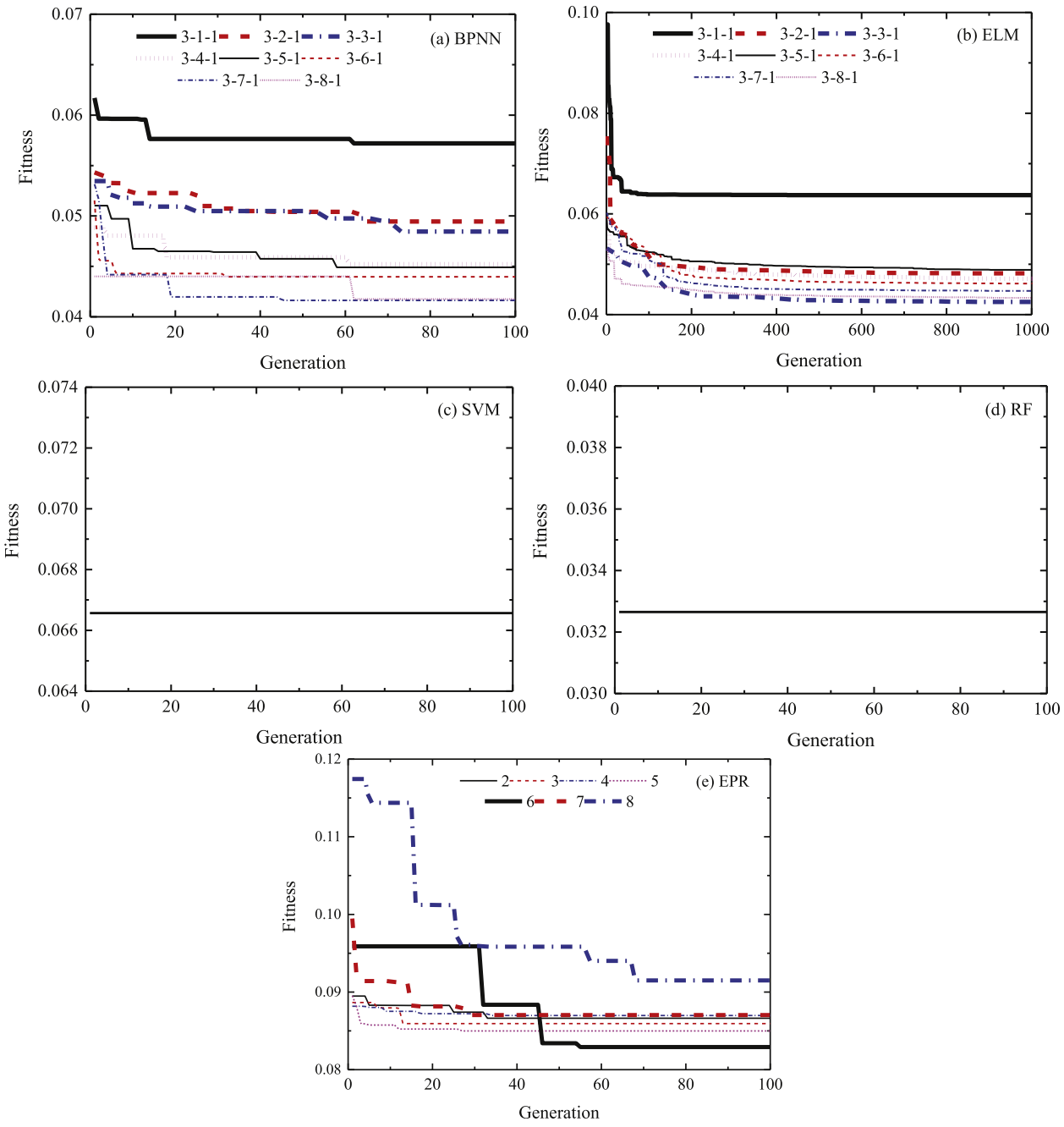

Fig. 5. Evolution of fitness values for: (a) BPNN; (b) ELM; (c) SVM; (d) RF; (e) EPR.

Table 7
Optimum values of hyper-parameters in five ML algorithms.

Algorithm	Hyper-parameters	Value	Convergence value
BPNN	*l*_hidden layers	1	0.04163
	*n*_neurons	7	
ELM	*n*_neurons	3	0.04252
SVM	*C*_penalty	132	0.06657
	*g*_width index	84	
RF	*n*_tree	131	0.03265
	*m*_try	1	
EPR	*z*_transformed term	6	0.08291

5.3. Sensitivity analysis

Sensitivity analysis (SA) aims at investigating how model output uncertainty can be apportioned to the uncertainty in each input variable (Saltelli and Sobol, 1995). There are two primary SA methods, namely local sensitivity analysis (LSA) and global sensitivity analysis (GSA). The calculation of LSA is to obtain the partial derivatives of the model response with respect to input variables at a given point. LSA method is useless if the relation between input and output variables is non-linear or the correlations between input variables are strong. In contrast, GSA can take into consideration the whole variable space, thereby the coupled effect among input variables can be considered. In this study, a simple linear correlation between input and output does not exist. Meanwhile, three input variables e_0, w_L and I_p obviously exist strong correlation, especially for w_L and I_p. Therefore, GSA method is employed to investigate the significance of input variables to clay compressibility in five ML models.

Variance-based GSA method is used in this study, which has been extensively used in geotechnical engineering (Zhang et al., 2017; Hamdia

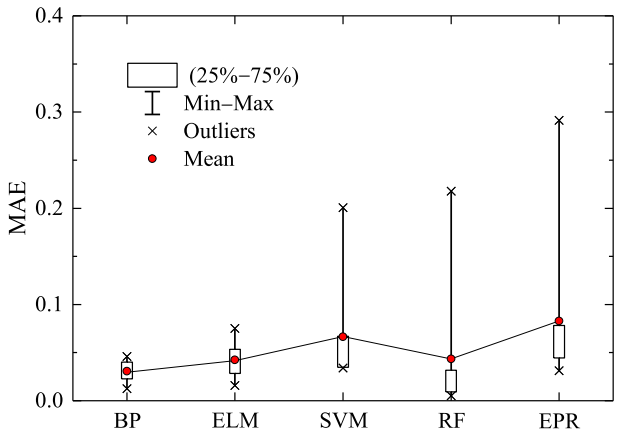

Fig. 6. Distributions of MAE values in 10 CV sets.

et al., 2018; Zhao et al., 2018). The total-order index S_{Ti} in variance-based GSA method measures the effect of the input parameter $\boldsymbol{X}_i$ on the output variable as well as the coupled effect of the $\boldsymbol{X}_i$ and other variables on the output variable. The calculation of S_{Ti} proposed by Jansen (1999) is adopted, as shown in Eqs. (24–26). The superiority of this estimator has been demonstrated by Saltelli et al. (2010).

$$S_{Ti} = 1 - \frac{V(E(Y|\boldsymbol{X}_{\sim i}))}{V(Y)} = \frac{E(V(Y|\boldsymbol{X}_{\sim i}))}{V(Y)} \tag{21}$$

$$E(V(Y|X_{\sim i})) = \frac{1}{2N}\sum_{j=1}^{N}\left(f(\mathbf{A})_j - f\left(\mathbf{A}_{\mathbf{B}}^{(i)}\right)_j\right)^2 \tag{22}$$

$$V(Y) = \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A})_j^2 - \left(\frac{1}{N}\sum_{j=1}^{N} f(\mathbf{A})_j\right)^2 \tag{23}$$

where, $\mathbf{A}$, $\mathbf{B}$ = two independent sampling matrices with $N/2$ (N is the total number of data in the training set) sets of data (see Eqs. (24) and (25)). $\mathbf{A}_{\mathbf{B}}^{(i)}$ = all columns are from $\mathbf{A}$ except the ith column which is from $\mathbf{B}$ (see Eq. (26)); f = prediction models, that are, optimum five optimum ML models in this study.

Table 8
Summary of indicators for five ML algorithms.

ML algorithm	Training set			Testing set		
	MAE	RD	NSE	MAE	RD	NSE
BPNN	0.0306	0.13	0.98	0.0325	0.15	0.99
ELM	0.0412	0.17	0.96	0.0457	0.17	0.98
SVM	0.0470	0.22	0.96	0.0465	0.22	0.98
RF	0.0008	0.00	1.00	0.0143	0.08	0.98
EPR	0.0542	0.20	0.95	0.0523	0.17	0.98

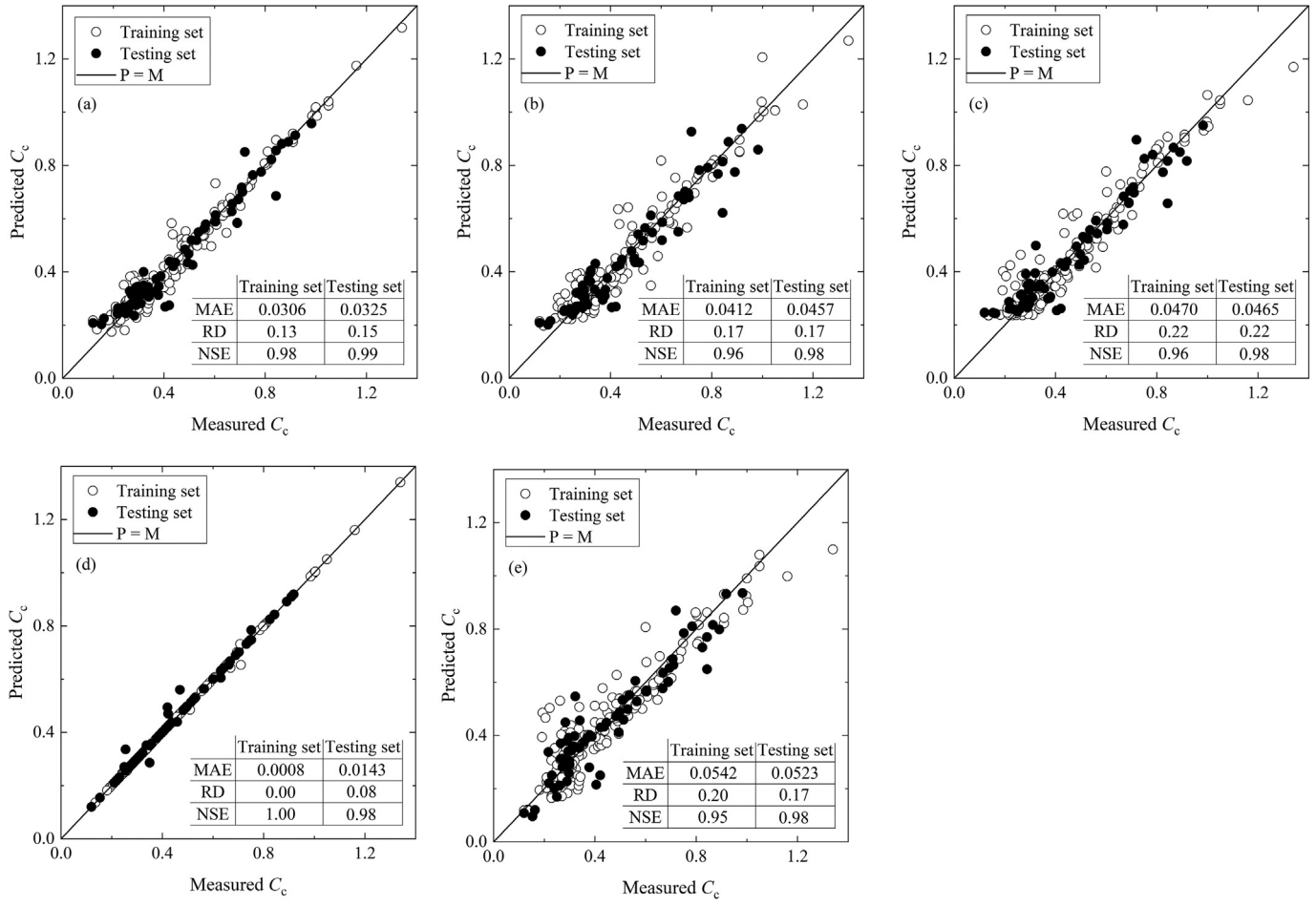

Fig. 7. Predicted C_c for training and testing sets by different methods: (a) BPNN; (b) ELM; (c) SVM; (d) RF; (e) EPR.

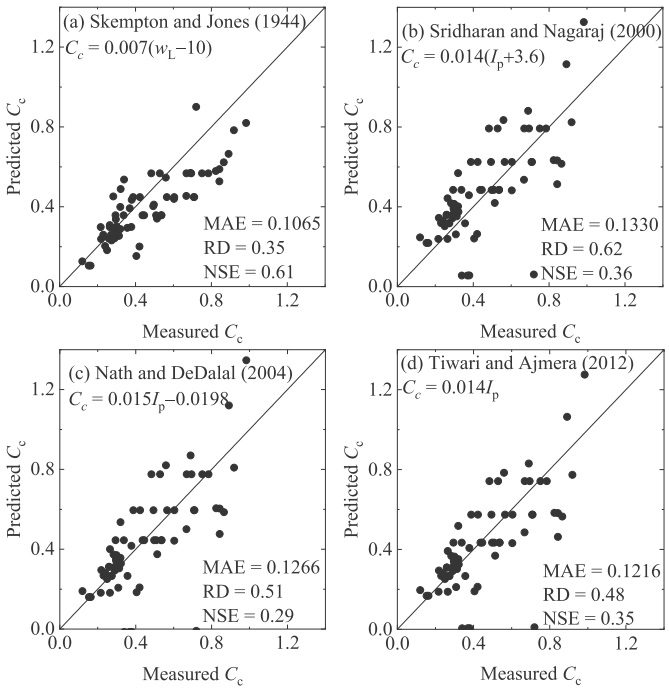

Fig. 8. Predicted C_c using empirical formulations.

$$\mathbf{A} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots x_i^{(1)} & \cdots x_k^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots x_i^{(2)} & \cdots x_k^{(2)} \\ \cdots & \cdots & \cdots & \\ x_1^{(N/2)} & x_2^{(N/2)} & \cdots x_i^{(N/2)} & \cdots x_k^{(N/2)} \end{bmatrix} \tag{24}$$

$$\mathbf{B} = \begin{bmatrix} x_1^{(N/2+1)} & x_2^{(N/2+1)} & \cdots x_i^{(N/2+1)} & \cdots x_k^{(N/2+1)} \\ x_1^{(N/2+2)} & x_2^{(N/2+2)} & \cdots x_i^{(N/2+2)} & \cdots x_k^{(N/2+2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(N)} & x_2^{(N)} & \cdots x_i^{(N)} & \cdots x_k^{(N)} \end{bmatrix} \tag{25}$$

$$\mathbf{A}_B^{(i)} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots x_i^{(N/2+1)} & \cdots x_k^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots x_i^{(N/2+2)} & \cdots x_k^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(N/2)} & x_2^{(N/2)} & \cdots x_i^{(N)} & \cdots x_k^{(N/2)} \end{bmatrix} \tag{26}$$

A and B matrices (5000 × 3) derive from 10,000 random samples generated in the *Parametric investigation* section. In order to comprehensively compare the significance of input variables in five ML models, the sum of all input variables total-order index is assumed to be 1. The total-order index proportion of each input variable in five ML models is shown in Fig. 11. It is clear that the significance of three input variables in five ML models is similar. C_c depends heavily on the e_0, and the significance of w_L and I_p is roughly identical. C_c describes soil volume change behavior and the change of soil volume relates to the compression of soil void, e_0 thus affects C_c dramatically. The significance of w_L and I_p to C_c is roughly identical in five ML models, which is consistent with experimental results (Kootahi and Moradi, 2016). The results of GAS in the five ML models harmonize with physical explanation, further indicating the reliability of ML models in predicting C_c.

6. Conclusions

This study comprehensively investigated the performance of five commonly used machine learning (ML) algorithms, i.e. back-propagation neural network (BPNN), extreme learning machine (ELM), support vector machine (SVM), random forest (RF) and evolutionary polynomial regression (EPR) in predicting C_c. Genetic algorithm (GA) was adopted to optimize the hyper-parameters in five ML algorithms. The average prediction errors for the 10-fold cross-validation (CV) sets were used as the fitness function in the GA, which could effectively enhance the robustness of ML models and avoid overfitting problem.

Five ML models with only three input parameters including initial void ratio e_0, water content w_L and plasticity index I_p obviously outperform the C_c empirical prediction formulations. For the test set, RF yields

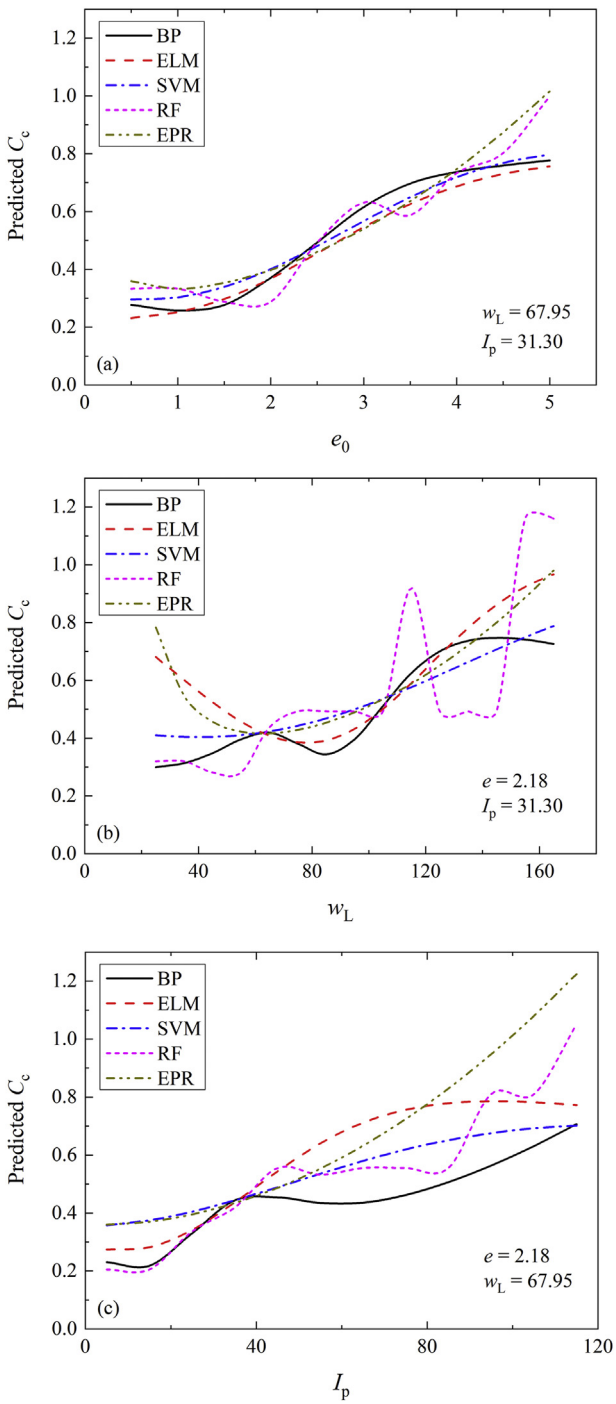

Fig. 9. Correlation between predicted C_c and three input variables respectively for: (a) initial void ratio e_0; (b) liquid limit w_L and (c) plasticity index I_p.

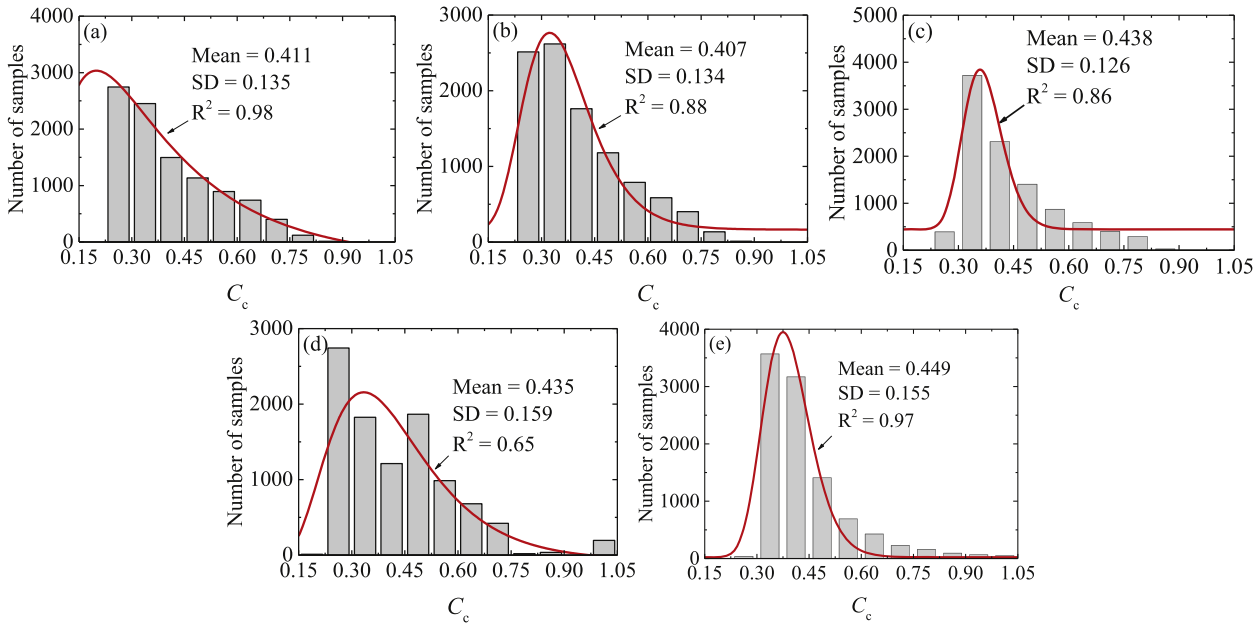

Fig. 10. Distribution of predicted C_c for randomly generated data using: (a) BPNN; (b) ELM; (c) SVM; (d) RF; (e) EPR.

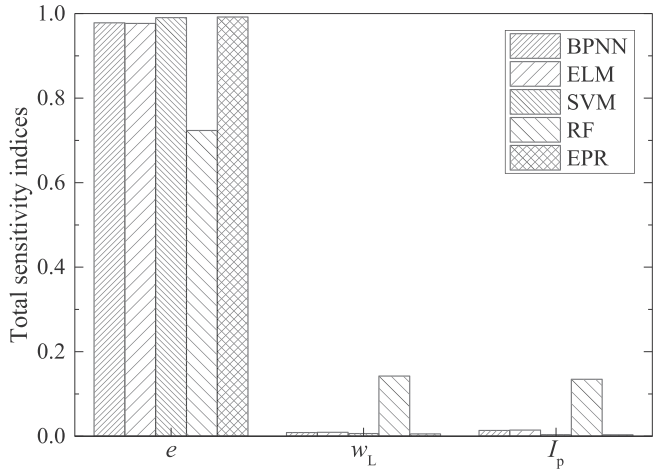

Fig. 11. Total sensitivity indices of three input variables.

the lowest MAE, RD and highest NSE values, 0.0143, 0.17 and 0.98, respectively, followed by BPNN, ELM, EPR and SVM. Parametric investigation indicates that predicted correlations between three input variables e_0, w_L, I_p for C_c using five ML models are roughly identical, showing a good agreement with the physical explanation, which indicates the generalization ability of ML-based models are acceptable.

From the perspective of the predicted C_c for the 10,000 random datasets using five ML models, BPNN, ELM, SVM and RF models are suitable for interpolation. and extrapolation. Results of variance-based global sensitivity analysis indicate that the values of predicted C_c in ML models primarily depend on the e_0, and the significance of w_L and I_p is roughly identical. Overall, if the ranges of input variables in database are large, RF-based model is recommended to predict C_c. Otherwise, if the ranges of input variables in database are small, EPR model is recommended. The explicit formulation of EPR-based model presented in this study can be conveniently used in engineering practice, and user-friendly application programming interface of RF-based model is developed for ensuring its easy usage, which can be download at following link: https://www.researchgate.net/publication/337918766_API_for_compression_index_prediction.

Users can easily run the code and achieve the results presented in this study. Furthermore, for the new datasets, the whole process can be easily achieved by updating the datasets in the original database in the Excel document.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The financial support provided by the RIF project (Grant No. PolyU R5037-18F) from the Research Grants Council (RGC) of Hong Kong is gratefully acknowledged.

References

Alavi, A.H., Gandomi, A.H., 2012. Energy-based numerical models for assessment of soil liquefaction. Geosci. Front. 3, 541–555.
Arsoy, S., Ozgur, M., Keskin, E., Yilmaz, C., 2013. Enhancing TDR based water content measurements by ANN in sandy soils. Geoderma 195–196, 133–144.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Burland, J.B., 1990. On the compressibility and shear strength of natural clays. Geotechnique 40, 329–337.
Cao, S., Song, W., Yilmaz, E., 2018. Influence of structural factors on uniaxial compressive strength of cemented tailings backfill. Construct. Build. Mater. 174, 190–201.
Cao, Z.J., Wang, Y., 2014. Bayesian model comparison and selection of spatial correlation functions for soil parameters. Struct. Saf. 49, 10–17.
Cerato, A.B., Lutenegger, A.J., 2004. Determining intrinsic compressibility of fine-grained soils. J. Geotech. Geoenviron. 130, 872–877.
Chen, R.P., Zhang, P., Kang, X., Zhong, Z.Q., Liu, Y., Wu, H.N., 2019a. Prediction of maximum surface settlement caused by EPB shield tunneling with ANN methods. Soils Found. 59, 284–295.
Chen, R.P., Zhang, P., Wu, H.N., Wang, Z.T., Zhong, Z.Q., 2019b. Prediction of shield tunneling-induced ground settlement using machine learning techniques. Front. Struct. Civ. Eng. 13, 1363–1378.
Cortes, C., Vapnik, V., 1995. Support-Vector networks. Mach. Learn. 20, 273–297.
Feng, Y., Cui, N., Hao, W., Gao, L., Gong, D., 2019. Estimation of soil temperature from meteorological data using different machine learning models. Geoderma 338, 67–77.
Ghorbani, A., Hasanzadehshooiili, H., 2018. Prediction of UCS and CBR of microsilica-lime stabilized sulfate silty sand using ANN and EPR models; application to the deep soil mixing. Soils Found. 58, 34–49.
Giasi, C.I., Cherubini, C., Paccapelo, F., 2003. Evaluation of compression index of remoulded clays by means of Atterberg limits. Bull. Eng. Geol. Environ. 62, 333–340.

Giustolisi, O., Savic, D.A., 2006. A symbolic data-driven technique based on evolutionary polynomial regression. J. Hydroinf. 8, 235–237.
Gunaydin, O., Gokoglu, A., Fener, M., 2010. Prediction of artificial soil's unconfined compression strength test using statistical analyses and artificial neural networks. Adv. Eng. Software 41, 1115–1123.
Guo, X., Dias, D., Carvajal, C., Peyras, L., Breul, P., 2018. Reliability analysis of embankment dam sliding stability using the sparse polynomial chaos expansion. Eng. Struct. 174, 295–307.
Habibbeygi, F., Nikraz, H., Koul, B.K., Iovine, G., 2018. Regression models for intrinsic constants of reconstituted clays. Cogent Geosci. 4, 1546978.
Hamdia, K.M., Ghasemi, H., Zhuang, X.Y., Alajlan, N., Rabczuk, T., 2018. Sensitivity and uncertainty analysis for flexoelectric nanostructures. Comput. Methods Appl. Math. 337, 95–109.
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844.
Holland, J.H., 1975. Adaptation in Natural and Artificial System. University of Michigan Press, Ann Arbor, Michigan.
Hong, Z.S., Yin, J., Cui, Y.J., 2010. Compression behaviour of reconstituted soils at high initial water contents. Geotechnique 60, 691–700.
Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70, 489–501.
Jansen, M.J.W., 1999. Analysis of variance designs for model output. Comput. Phys. Commun. 117, 35–43.
Jiang, H., He, W., 2012. Grey relational grade in local support vector regression for financial time series prediction. Expert Syst. Appl. 39, 2256–2262.
Jin, Y.-F., Wu, Z.-X., Yin, Z.-Y., Shen, J.S., 2017. Estimation of critical state-related formula in advanced constitutive modeling of granular material. Acta. Geotech. 12, 1329–1351.
Jin, Y.-F., Yin, Z.-Y., Shen, S.-L., Hicher, P.-Y., 2016a. Investigation into MOGA for identifying parameters of a critical-state-based sand model and parameters correlation by factor analysis. Acta Geotech 11, 1131–1145.
Jin, Y.-F., Yin, Z.-Y., Shen, S.-L., Hicher, P.-Y., 2016b. Selection of sand models and identification of parameters using an enhanced genetic algorithm. Int. J. Numer. Anal. Model. 40, 1219–1240.
Jin, Y.-F., Yin, Z.-Y., Wu, Z.-X., Daouadji, A., 2018a. Numerical modeling of pile penetration in silica sands considering the effect of grain breakage. Finite Elem. Anal. Des. 144, 15–29.
Jin, Y.-F., Yin, Z.-Y., Zhou, W.-H., Huang, H.-W., 2019. Multi-objective optimization-based updating of predictions during excavation. Eng. Appl. Artif. Intell. 78, 102–123.
Jin, Y.F., Yin, Z.Y., Wu, Z.X., Zhou, W.H., 2018b. Identifying parameters of easily crushable sand and application to offshore pile driving. Ocean Eng. 154, 416–429.
Kaastra, I., Boyd, M., 1996. Designing a neural network for forecasting financial and economic time series. Neurocomputing 10, 215–236.
Kirts, S., Panagopoulos, O.P., Xanthopoulos, P., Nam, B.H., 2018. Soil-compressibility prediction models using machine learning. J. Comput. Civ. Eng. 32 (1), 04017067.
Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp. 1137–1143
Koopialipoor, M., Jahed Armaghani, D., Haghighi, M., Ghaleini, E.N., 2017. A neuro-genetic predictive model to approximate overbreak induced by drilling and blasting operation in tunnels. Bull. Eng. Geol. Environ. 78, 981–990.
Kootahi, K., Moradi, G., 2016. Evaluation of compression index of marine fine-grained soils by the use of index tests. Mar. Georesour. Geotechnol. 35, 548–570.
Kundu, S., Khare, D., Mondal, A., 2017. Future changes in rainfall, temperature and reference evapotranspiration in the central India by least square support vector machine. Geosci. Front. 8, 583–596.
Li, Z., Chen, L., 2019. A novel evidential FMEA method by integrating fuzzy belief structure and grey relational projection method. Eng. Appl. Artif. Intell. 77, 136–147.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R. News 23, 18–21.
Liu, C., Jiang, Z., Han, X., Zhou, W., 2019. Slope displacement prediction using sequential intelligent computing algorithms. Measurement 134, 634–648.
Liu, K., Liu, B., 2019. Intelligent information-based construction in tunnel engineering based on the GA and CCGPR coupled algorithm. Tunn. Undergr. Space Technol. 88, 113–128.
Masters, T., 1994. Practical Neural Network Recipes in C++. Academic Press, Boston, MA.
Müthing, N., Zhao, C., Hölter, R., Schanz, T., 2018. Settlement prediction for an embankment on soft clay. Comput. Geotech. 93, 87–103.
Nagaraj, T.S., Murthy, B.R.S., 1983. Rationalization of Skempton's compressibility equation. Geotechnique 33, 433–443.
Nagaraj, T.S., Murthy, B.R.S., 1986. A critical reappraisal of compression index equations. Geotechnique 36, 27–32.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - a discussion of principles. J. Hydrol. 10, 282–290.
Nath, A., DeDalal, S.S., 2004. The role of plasticity index in predicting compression behaviour of clays. Electron. J. Geotech. Eng. 9, 1–7.
Nielsen, R.H., 1987. Kolmogorov's mapping neural network existence theorem. In: Proceedings of the IEEE First International Conference on Neural Networks, San Diego, CA, USA, pp. 11–13.
Orr, T.L., Cherubini, C., 2003. Use of the ranking distance as an index for assessing the accuracy and precision of equations for the bearing capacity of piles and at-rest earth pressure coefficient. Can. Geotech. J. 40, 1200–1207.
Paola, J.D., 1994. Neural Network Classification of Multispectral Imagery. The University of Arizona, USA.
Park, H.I., Lee, S.R., 2011. Evaluation of the compression index of soils using an artificial neural network. Comput. Geotech. 38, 472–481.
Pham, B.T., Son, L.H., Hoang, T.-A., Nguyen, D.-M., Tien Bui, D., 2018. Prediction of shear strength of soft soil using machine learning methods. Catena 166, 181–191.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.
Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput. Phys. Commun. 181, 259–270.
Saltelli, A., Sobol, I.M., 1995. About the use of rank transformation in sensitivity analysis of model output. Reliab. Eng. Syst. Saf. 50, 225–239.
Shahin, M.A., Maier, H.R., Jaksa, M.B., 2005. Investigation into the robustness of artificial neural networks for a case study in civil engineering. In: Zerger, A., Argent, R.M. (Eds.), MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, pp. 79–83.
Shen, S.L., Wu, H.N., Cui, Y.J., Yin, Z.Y., 2014. Long-term settlement behaviour of metro tunnels in the soft deposits of Shanghai. Tunn. Undergr. Space Technol. 40, 309–323.
Skempton, A.W., Jones, O.T., 1944. Notes on the compressibility of clays. Quar. J. Geol. Soc. 100, 119–135.
Sridharan, A., Nagaraj, H.B., 2000. Compressibility behaviour of remoulded, finegrained soils and correlation with index properties. Can. Geotech. J. 37, 712–722.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. C-Appl. 36, 111–147.
Tiwari, B., Ajmera, B., 2011. Consolidation and swelling behavior of major clay minerals and their mixtures. Appl. Clay Sci. 54, 264–273.
Tiwari, B., Ajmera, B., 2012. New correlation equations for compression index of remolded clays. J. Geotech. Geoenviron. 138, 757–762.
Tran, C., Srokosz, P., 2010. The idea of PGA stream computations for soil slope stability evaluation. Cr. Mecanique 338, 499–509.
Wang, C., 1994. A Theory of Generalization in Learning Machines with Neural Application. The University of Pennsylvania, USA.
Wroth, C.P., Wood, D.M., 1978. The correlation of index properties with some basic engineering properties of soils. Can. Geotech. J. 15, 137–145.
Wu, H.N., Zhang, P., Chen, R.P., Lin, X.T., Liu, Y., 2020. Ground response to horizontal spoil discharge jet-grouting with impacts on the existing tunnels. J. Geotech. Geoenviron. Eng. 146 (7), 05020006.
Yang, J., Yin, Z.Y., Laouafa, F., Hicher, P.Y., 2019. Internal erosion in dike-on-foundation modeled by a coupled hydro-mechanical approach. Int. J. Numer. Anal. Model. 43, 663–683.
Yin, Z.-Y., Hattab, M., Hicher, P.-Y., 2011. Multiscale modeling of a sensitive marine clay. Int. J. Numer. Anal. Model. 35, 1682–1702.
Yin, Z.-Y., Jin, Y.-F., Shen, S.-L., Huang, H.-W., 2016a. An efficient optimization method for identifying parameters of soft structured clay by an enhanced genetic algorithm and elastic–viscoplastic model. Acta Geotech 12, 849–867.
Yin, Z.-Y., Zhu, Q.-Y., Zhang, D.-M., 2017a. Comparison of two creep degradation modeling approaches for soft structured soils. Acta. Geotech. 12, 1395–1413.
Yin, Z.Y., Jin, Y.F., Huang, H.W., Shen, S.L., 2016b. Evolutionary polynomial regression based modelling of clay compressibility using an enhanced hybrid real-coded genetic algorithm. Eng. Geol. 210, 158–167.
Yin, Z.Y., Jin, Y.F., S, S.J., Hicher, P.Y., 2017b. Optimization techniques for identifying soil parameters in geotechnical engineering: comparative study and enhancement. Int. J. Numer. Anal. Model. 42, 1–25.
Yin, Z.Y., Xu, Q., Yu, C., 2015. Elastic-Viscoplastic modeling for natural soft clays considering nonlinear creep. Int. J. GeoMech. 15 (5), A6014001.
Zhang, J., Zhang, L.M., Tang, W.H., 2009. Bayesian framework for characterizing geotechnical model uncertainty. J. Geotech. Geoenviron. 135, 932–940.
Zhang, L., Li, D.-Q., Tang, X.-S., Cao, Z.-J., Phoon, K.-K., 2018. Bayesian model comparison and characterization of bivariate distribution for shear strength parameters of soil. Comput. Geotech. 95, 110–118.
Zhang, L.M., Wu, X.G., Zhu, H.P., AbouRizk, S.M., 2017. Performing global uncertainty and sensitivity analysis from given data in tunnel construction. J. Comput. Civ. Eng. 31, 04017065.
Zhang, P., 2019. A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model. Appl. Soft Comput. 85, 105859.
Zhang, P., Chen, R.-P., Wu, H.-N., Liu, Y., 2020a. Ground settlement induced by tunneling crossing interface of water-bearing mixed ground: a lesson from Changsha, China. Tunn. Undergr. Space Technol. 96, 103224.
Zhang, P., Chen, R.P., Wu, H.N., 2019. Real-time analysis and regulation of EPB shield steering using Random Forest. Automat. Constr. 106, 102860.
Zhang, P., Yin, Z.-Y., Jin, Y.-F., Chan, T.H.T., 2020b. A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. Eng. Geol. 265, 105328.
Zhang, P., Yin, Z.Y., Jin, Y.F., Ye, G.L., 2020c. An AI-based model for describing cyclic characteristics of granular materials. Int. J. Numer. Anal. Model. 44 (9), 1315–1335.
Zhao, C.Y., Lavasan, A.A., Hölter, R., Schanz, T., 2018. Mechanized tunneling induced building settlements and design of optimal monitoring strategies based on sensitivity field. Comput. Geotech. 97, 246–260.
Zhou, W.-H., Garg, A., Garg, A., 2016. Study of the volumetric water content based on density, suction and initial water content. Measurement 94, 531–537.
Zhu, Q.Y., Jin, Y.F., Yin, Z.Y., 2020. Modeling of embankment beneath marine deposited soft sensitive clays considering straightforward creep degradation. Mar. Georesour. Geotechnol. 38 (5), 553–569.
Zhu, Q.Y., Yin, Z.Y., Hicher, P.Y., Shen, S.L., 2016. Nonlinearity of one-dimensional creep characteristics of soft clays. Acta. Geotech. 11, 887–900.