Front. Struct. Civ. Eng. 2022, 16(2): 224–238
https://doi.org/10.1007/s11709-022-0812-6

RESEARCH ARTICLE

Dimensionality reduction and prediction of soil consolidation coefficient using random forest coupling with Relief algorithm

Hai-Bang LY[a], Huong-Lan Thi VU[a], Lanh Si HO[a,b*], Binh Thai PHAM[c]

[a] *Department of Civil Engineering, University of Transport Technology, Hanoi 100000, Vietnam*
[b] *Civil and Environmental Engineering Program, Hiroshima University, Hiroshima 739-8527, Japan*
[c] *Department of Science, Technology and International Cooperation, University of Transport Technology, Hanoi 100000, Vietnam*
[*] *Corresponding author. E-mail: lanhhs@utt.edu.vn*

ABSTRACT The consolidation coefficient of soil (C_v) is a crucial parameter used for the design of structures leaned on soft soi. In general, the C_v is determined experimentally in the laboratory. However, the experimental tests are time-consuming as well as expensive. Therefore, researchers tried several ways to determine C_v via other simple soil parameters. In this study, we developed a hybrid model of Random Forest coupling with a Relief algorithm (RF-RL) to predict the C_v of soil. To conduct this study, a database of soil parameters collected from a case study region in Vietnam was used for modeling. The performance of the proposed models was assessed via statistical indicators, namely Coefficient of determination (R^2), Root Mean Squared Error (*RMSE*), and Mean Absolute Error (*MAE*). The proposal models were constructed with four sets of soil variables, including 6, 7, 8, and 13 inputs. The results revealed that all models performed well with a high performance ($R^2 > 0.980$). Although the RF-RL model with 13 variables has the highest prediction accuracy ($R^2 = 0.9869$), the difference compared with other models was negligible (i.e., $R^2 = 0.9824$, 0.9850, 0.9825 for the cases with 6, 7, 8 inputs, respectively). Thus, it can be concluded that the hybrid model of RF-RL can be employed to predict C_v based on the basic soil parameters.

KEYWORDS soil consolidation coefficient, machine learning, random forest, Relief

1 Introduction

Due to rapid economic growth and urbanization, most construction structures have to be built on grounds that often face soft soil conditions. One of the most important characteristics of soft soil related to the failure of structures is the settlement. Therefore, in the design stage of the construction project, the estimation or calculation of settlement is a crucial task. The settlement of soft soil (soft clay) is generally related to the consolidation problem caused by changing of volume due to dissipating of pore water under changing of the effective stress. The magnitude and rate of consolidation settlement are attributed to the compression index (C_c), which is calculated through the consolidation coefficient (C_v). The C_v is defined as a parameter used to denote the degree where saturated clay experiences consolidation when it is subjected to an increment of pressure [1]. Thus, the C_v is considered one of the most vital parameters used to calculate and predict the settlement due to the consolidation of soft soil [2]. In the laboratory, the Oedometer Test (one-dimensional consolidation test) is generally employed to determine the C_v [3], while in the field, the Cone and piezocone penetration test (CPTu) is used [4,5]. However, the values of C_v determined from these tests have significant variations; besides, these tests are expensive and time-consuming. From the above difficulties, many previous researchers have tried to estimate the C_v by establishing the relationship of C_v with other basic soil parameters. For example, some authors used Atterberg's limits to estimate the C_v [1]. In another study, the C_v was forecasted using void ratio and

Article history: Received Oct 16, 2021; Accepted Dec 1, 2021

overburden pressure [6]. Besides, simple regression such as the back analysis method and some empirical models were employed to estimate the C_v [2,7,8]. Nevertheless, these simple regressions or models have some limitations and weaknesses, such as could only be applied for a simple case with only several parameters, or these methods were based on a partial number of linear or non-linear equations [9,10].

It is known that machine learning (ML) and artificial intelligence (AI) have been applied popularly in engineering problems, particularly in geotechnical engineering to estimate basic parameters of soil and rock such as shear strength, soil permeability coefficient, compressibility coefficient, and consolidation coefficient of soil [11–15]. Regarding the prediction of soil shear strength, many authors have successfully used different ML and AI techniques for these problems. For example, decision tree (DT) and artificial neural network (ANN) have been employed and compared to estimate shear strength, and results indicated that ANN performed better than DT [16]. Another author indicated that functional networks had a higher prediction than ANN but lower than support vector machine (SVM) [17]. Furthermore, gradient boosting is a state-of-the-art ML approach applied in estimating soil parameters. For example, extreme gradient boosting (XGboost) has been applied successfully in forecasting soil's undrained shear strength and compression index [18,19]. In the problems of estimating the compression coefficient of soil, many authors have extensively used various ML and AI techniques for predicting soil compression coefficient (C_c). Pham et al. [20] compared ANN, ANFIS, and SVM models with Monte Carlo sensitivity analysis. The authors revealed that the SVM model was the best in estimating C_c compared to ANN and ANFIS models. By using a hybrid model of ML such as Harris hawks optimization (HHO-ANN), grasshopper optimization algorithm (GOA-ANN), and PSO-MLP, other authors have indicated that these hybrid models could be potential alternative methods for estimating C_c [21,22].

In the study on estimating the consolidation coefficient of soil, limited previous studies have used ML and AI. Pham et al. [23] used multi-layer perceptron neural network-biogeography-based optimization (MLP-BBO) in comparison with backpropagation multi-layer perceptron neural networks (Bp-MLP Neural Nets), radial basis functions neural networks (RBF-Neural Nets), Gaussian process (GP), M5 Tree, and support vector regression (SVR) using Atterberg limits, water content, and clay content. They suggested that MLP-BBO was the best model for this case. In another study, they evaluated and compared the performance of ANN-BBO, ANN, ANFIS, and SVM, they found that all the models have done well, but, the ANN-BBO model was the best [24]. Besides, the random forest (RF) technique was also firstly applied to forecast the C_v [25]. They reported that RF is a good algorithm for estimating this soil parameter. It is accepted that RF is one of the effective ensemble ML methods that is popularly used for regression and classification in geotechnical engineering [26–28], recently in predicting the shear strength of soil [29]. Besides, to have outstanding RF modeling, it is needed to use or combine with an optimization algorithm to fine-tune hyperparameters. Relief is an attribute estimation algorithm, which has been known as being both non-parametric [30] and non-myopic [31]. In other words, Relief evaluates the characteristic of a specified feature from the perspective of other features, and it does not need to make an assumption in terms of sample size or population distribution. Furthermore, it was indicated that the efficiency of the Relief algorithm had been explained by the fact that this algorithm does not obviously discover feature subsets [32].

Based on the above literature, C_v's prediction is vital; however, there is a limited study using a hybrid model of ML and AI. Thus, in this study, we aimed at developing a new hybrid ML model of RF coupling with Relief to accurately estimate the C_v of soil using data collected from a project in Vietnam. In addition, this study can fill the gap of literature in estimating C_v using ML and AI methods. The hybrid model in this study was a combination of RF and a Relief algorithm. The soil data obtained from field and laboratory tests were adopted to build the datasets for training and testing. The common criteria, namely *MAE*, *RMSE*, and R^2 were used to evaluate the performance of models.

2 Database construction

In this research, soil samples obtained from Hanoi-Hai Phong expressway project in Vietnam were employed for modeling. In this study, we used thirteen important inputs (i.e., variables), including depth of the sample, clay fraction, moisture content, bulk density, dry density, specific gravity, void ratio, porosity, degree of saturation, Atterberg's limits, and the output is of the model is the C_v. These parameters directly affect the consolidation coefficient. The consolidation coefficient is strongly influenced by soil type, saturation degree (i.e., clay fraction), void ratio, and porosity. In general, the higher void ratio leads to a higher consolidation, which causes a greater consolidation settlement. The detail of these inputs and output can be found in Table 1. The mean values, standard deviation, and skewness of all inputs were also described in Table 1. Besides, the multi-correlation between input and output variables was also analyzed and presented in Fig. 1. There are strong correlations among input variables, such as moisture content versus void ratio, porosity, and liquid limit (R > 0.8). Also, high correlations between void ratio and porosity versus liquid limit, plastic limit, and plasticity index are observed. Furthermore, clay fraction also has a good relationship with moisture content, void ratio, and porosity. These high mutual correlations among input variables can be attributed to the physical relationships between them, for example, a higher void ratio and porosity of soil will lead to a higher moisture content of the soil. From Fig. 1, it can be observed that the highest multi-correlation between input and output was found for the bulk and dry density. This may be related to effective stress, which is one of the most critical parameters governing soil consolidation. Depth of sample and liquidity index also has a strong correlation with output.

3 Background of the methods

3.1 Random forest

The RF algorithm was first introduced by Breiman. This is a non-parametric technique derived from classification and regression trees (CART) [33]. The RF model works on the basis of growing more trees, each with a bootstrap pattern [34]. A randomized process is used to generate a subset of predictors at each node, also known as each part of the tree. The mean value of the obtained results is the output of RF [35]. The RF consists of a classification tree and a regression tree with the following risk objective or function [36] (Fig. 2).

$$\min\left(J(k,t_{\mathrm{k}}) = \frac{n_{\mathrm{left}}}{n} MSE_{\mathrm{left}} + \frac{n_{\mathrm{right}}}{n} MSE_{\mathrm{right}}\right), \qquad (1)$$

where MSE_{left} is the Mean Squared Error of the left subset; MSE_{right} is the Mean Squared Error of the right subset; n_{left} is the sample of the left subset; n_{right} is the sample of the right subset.

3.2 Relief algorithm–Feature selection

The problem of incomplete data and limited to 2-class

Table 1 Statistical analysis of the inputs and output in this study

variable	task	notation	range	mean	St.D.[a)]	SK[b)]
depth of sample (m)	input	I_1	1.600–35.700	12.819	7.029	0.715
clay (%)	input	I_2	4.500–47.500	24.588	8.873	–0.315
moisture (%)	input	I_3	28.030–67.850	48.501	9.476	–0.487
bulk density (g/cm^3)	input	I_4	1.520–1.930	1.708	0.083	0.582
dry density (g/cm^3)	input	I_5	0.920–1.490	1.158	0.133	0.791
specific gravity	input	I_6	2.660–2.720	2.689	0.012	0.067
void ratio	input	I_7	0.805–1.891	1.351	0.256	–0.396
porosity (%)	input	I_8	44.600–65.410	56.919	5.022	–0.820
degree saturation (%)	input	I_9	84.110–99.920	96.461	3.091	–1.463
liquid limit (%)	input	I_{10}	30.110–76.190	52.497	10.984	–0.190
plastic limit (%)	input	I_{11}	15.060–37.060	27.750	4.611	–0.528
plasticity index (%)	input	I_{12}	9.400–47.150	24.791	7.976	0.283
liquidity index	input	I_{13}	0.520–1.660	0.853	0.168	1.850
coef. consolidation (cm^2/1000 s)	output	O	0.310–3.370	1.168	0.742	1.367

Notes: a) St.D. = Standard Deviation; b) SK = Skewness.

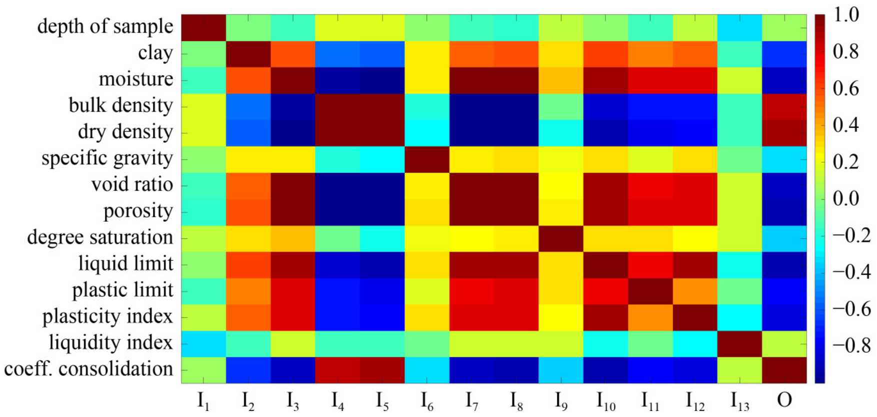


Fig. 1 Multi-correlation graph of input and output parameters employed in this study.

generates a massive difficulty for prediction algorithms (Fig. 3). To get around this problem, a solution is proposed, called the Relief [37]. Relief is an attribute estimation algorithm in the condition of having numerous unrelated random properties created by two scientists Kira and Rendell [32,38]. In his research, Ditterrich [39] argued that this is the most successful preprocessing algorithm ever, due to its simplicity and effectiveness [39]. In another research, Sun [40] supported that the key in the algorithm is the need to distinguish neighboring samples by repeatedly estimating the feature weights of the object according to their ability. In each iteration cycle, the algorithm will select a random sample of y, then the two closest neighbors of y: y_1 and y_2 will be found. Next, the weight of the ith feature will be updated.

$$w_i = w_i + \left|y^{(i)} - NM^{(i)}(y)\right| - \left|y^{(i)} - NH^{(i)}(y)\right|. \tag{2}$$

where w_i = P (different value of ith feature/NM) − P (different value of ith feature/NH); NH is termed the nearest hit and NM is termed the nearest miss [41].

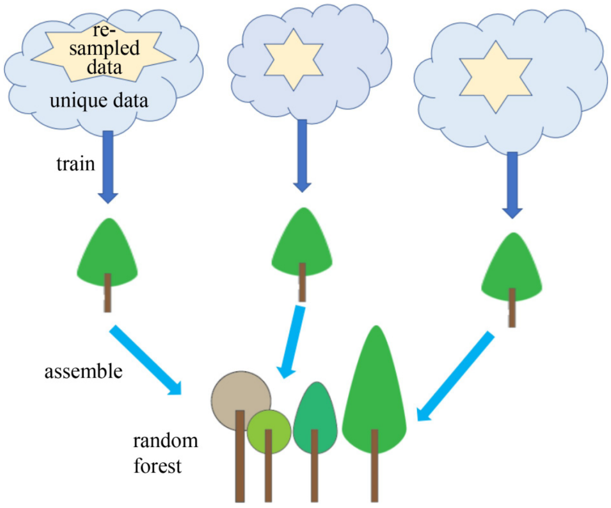


Fig. 2 RF algorithm.

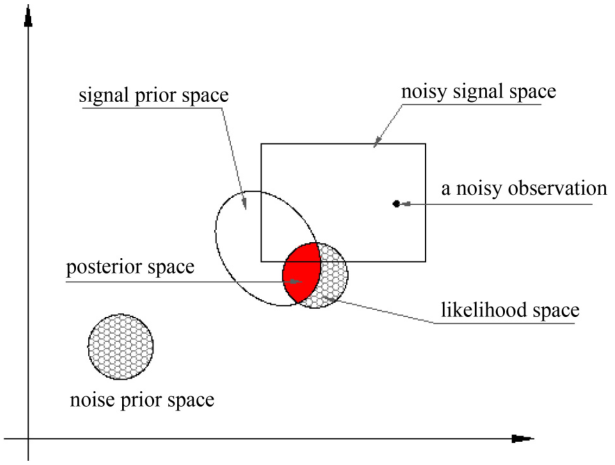


Fig. 3 Relief algorithm.

The algorithm can handle noise and missing data by further expanding the Relief.

The pseudocode of the Relief Algorithm is as follows [37].

```
1) Initialization: given $D = \{(y_k, z_k)\}_{k=1}^{K}$, set $w_i = 0$, $1 \leqslant i \leqslant I$, number of iterations $P$;
2) for $p = 1: P$
3) Randomly select a pattern $y$ from $D$;
4) Find the nearest hit $NH(y)$ and miss $NM(y)$ of $y$;
5) for $i = 1: I$
6) Compute: $w_i = w_i + \left|y^{(i)} - NM^{(i)}(y)\right| - \left|y^{(i)} - NH^{(i)}(y)\right|$
7) end
8) end
```

3.3 Performance indicators

In this paper, to evaluate the effectiveness of the model, we use the following types of indicators: Coefficient of determination (R^2) measures the square correlation between the design value and the predicted value [41], the value of R^2 changes from 0 to 1, the model is said to be more accurate as R^2 is closer to 1 [42]. In addition, we also used root mean square error ($RMSE$) [43] and mean absolute error (MAE) [44]. The proposed algorithms get better results when the value of $RMSE$ and MAE are small [45,46]. Where $RMSE$ shows the difference in value between reality and prediction shown in the equation below. Besides, MAE presents the average error between fact and prediction. These coefficients are calculated through the following equations [47,48]:

$$MAE = \frac{1}{k}\sum_{i=1}^{k}(c_i - \overline{c_i}), \tag{3}$$

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(c_i - \overline{c_i})^2}, \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{k}(c_i - \overline{c_i})^2}{\sum_{i=1}^{k}(c_i - \overline{c})^2}, \tag{5}$$

where k infers the number of the samples, c_i and $\overline{c_i}$ are the actual and predicted outputs, respectively, and $\overline{c}$ is the average value of the c_i.

4 Methodology flow chart

The current study is conducted according to the proposed methodology, which consists of three main steps as follows: 1) data preparation; 2) constructing of the models; and 3) validating the proposed models (Fig. 4).

1) Data preparation: in this step, the sample data taken from the laboratory was adopted to build a training and testing (validating) dataset. The dataset was divided with a ratio 70/30, in which 70% used for training and 30% for testing.

2) Constructing the models: in this step, the training dataset was used for training the models on the basis of RF coupling with Relief algorithm.

3) Validating the proposed models: in this step, the testing dataset was employed to validate the proposed model. Statistical criteria consisting of *RMSE*, *MAE*, R^2 were employed to validate the models.

5 Results

5.1 Feature selection by Relief algorithm

In general, to prevent multicollinearity and overfitting problems, dimension reduction such as principal component analysis (PCA) and feature selection approaches using particle swarm optimization, genetic algorithm, RF, Relief algorithm [49–51]. This study used the Relief algorithm to perform the feature selection. The nearest neighbor is a crucial factor of the Relief algorithm. When looking at the test data set, it can be noticed that nearest neighbors ranging from 5 to 500 gave different results. Realize that it is possible to find 3 data sets from the weights of the 13 input variables in the data set. Figure 5(a) uses 5 to 35 closest neighbors (the lowest number of neighbors), the weights of the variables are also the worst in the 3 data sets. The weight of inputs reaches a relatively high value of 0.042 when there is only 1 input variable, but when using from 1 to 8 input variables, the weights of inputs decrease drastically and relatively equally, only reaching from 0.005–0.1, even when the number of input variables is 6 and the weight is less than 0. When the number of input variables was 9, the weights of the variables reached 0.02 and started to increase, then peaked at about 0.43 when the number of input is 10, then decreased gradually and spiked again to reach the maximum value of 0.81 when the number of input is 13.

Figure 5(b) looks for 40 to 85 the nearest neighbors. The resulting graph shows that the weights of the inputs are higher than in Fig. 5(a). Most notably, when the number of input variables is 6, the weight of the input reaches the min value of less than 0, and the maximum value is 0.08 when the number of input variables is 10.

Likewise, Fig. 5(c) looks for 90 to more than 500 nearest neighbors. The resulting graph shows that the weights of the inputs are higher in Figs. 6(a) and 6(b). Especially when the number of input variables is 6, the weight of the input reaches the min value, while it reaches a maximum value of 0.09 when the number of input variables is 10.

In Table 2, the RF is denoted RF, the Relief algorithm denoted by RL, the numbers 13, 6,7, 8, are the number of different input variables used in the data set. Overall, using Relief, 4 data sets are finally selected.

5.2 Convergence and statistical analysis

In this work, it was decided to conduct 50 simulations, after which one could determine whether or not the number was adequate, and the proposed ML model was reliable. This approach has been proposed in several recent studies [52,53], and proved to be a reliable method to investigate the "response" of the model under the random sampling effect. The convergence analysis is conducted with R^2 and *RMSE* for the training and testing datasets and is depicted in Fig. 6. Regarding model RF-RL13, R^2 converged within 0.5% of error only after 8 simulations. And *RMSE* required 11 simulations to converge in 5% of error (Figs. 6(a) and 6(b)). For

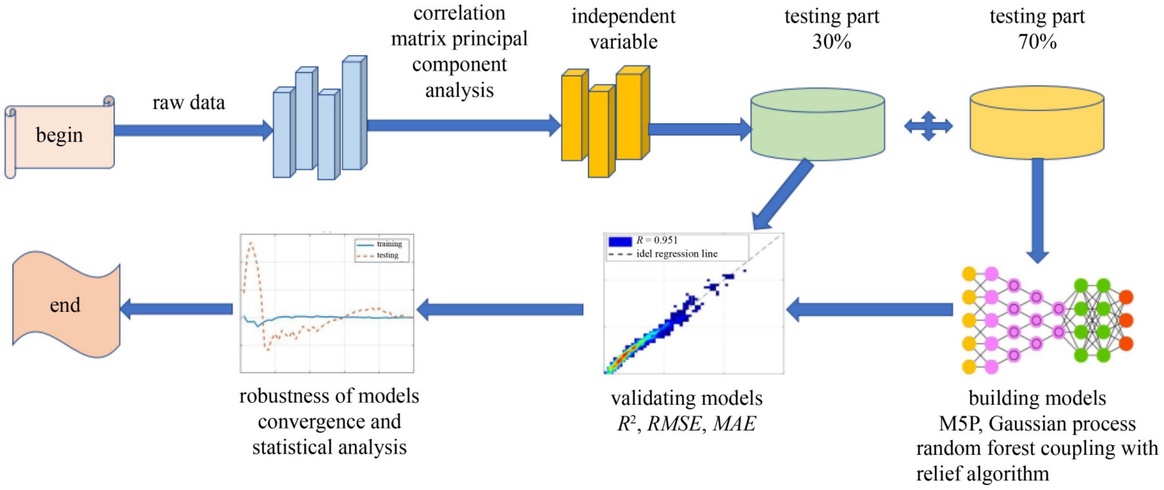

correlation matrix principal component analysis
independent variable
testing part 30%
testing part 70%
raw data
begin
R = 0.951
idel regression line
training
testing
end
robustness of models convergence and statistical analysis
validating models
R^2, *RMSE*, *MAE*
building models
M5P, Gaussian process random forest coupling with relief algorithm

Fig. 4 Methodology flow chart.

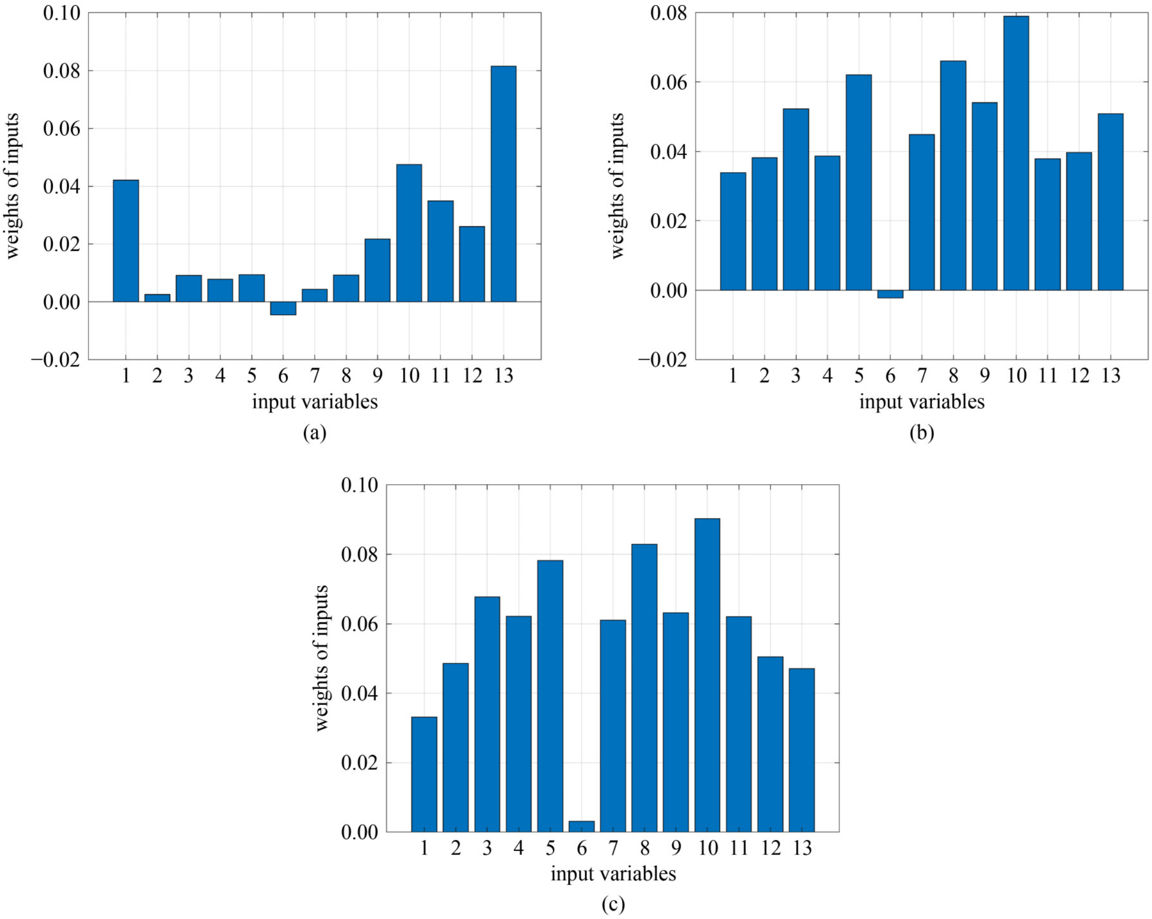


Fig. 5 Weights of input variables in the function of different nearest neighbors for: (a) from 5 to 35 nearest neighbors; (b) from 40 to 85 nearest neighbors; (c) more than 90 nearest neighbors.

RF-RL6, R^2 converged within 0.35% of error only after 13 simulations. Moreover, *RMSE* required 32 simulations to converge in 0.35% of error (Figs. 6(c) and 6(d)). For RF-RL7: R^2 converged within 0.7% of error only after 13 simulations. Furthermore, *RMSE* required 13 simulations to converge in 0.7% of error (Figs. 6(e) and 6(f)). For RF-RL8, R^2 converged within 0.3% of error only after 27 simulations. And *RMSE* required 27 simulations to converge in 0.3% of error (Figs. 6(g) and 6(h)). As a result, 50 simulations for each case are conducted, and the convergence analysis ensures the reliability of the prediction results of all RF-related models proposed herein.

Specifically, when the number of input variables is 13, as can be seen in Fig. 7, the obtained R^2 is the highest, for the training set R^2 = 0.9905; with the testing set R^2 = 0.9856. In the same case, the obtained *RMSE* is the lowest, for the training set *RMSE* = 0.098; for the test set *RMSE* = 0.2168. At the same time, obtained *MAE* is also the smallest, for the training set *MAE* = 0.0602, for the testing set *RMSE* = 0.1148. However, the best case could be RF-RL7 with 7 inputs, but the prediction accuracy remained the same as the raw dataset with 13 inputs. As a result, reducing 6 inputs may not affect the prediction results. Using four different datasets, the results of training and testing statistics are shown in Table 3. It can be shown that the training was effective, and that high accuracy was achieved in all cases.

5.3 Prediction accuracy

The simulation of C_v by the RF-RL model using the training and testing dataset is shown in Fig. 8. Here, the graphs above (Fig. 8) show the best predictions presented here with 4 data sets. It can be seen that there is a high correlation between the predicted and actual values, the predicted results are almost the same as the experimental results for both the training and testing set.

Figure 9 shows the comparison between the predicted and measured values of C_v of all data for the cases with 6, 7, 8, and 13 input variables using RF coupling with Relief (hereafter call as RF-RL). The correlation coefficient values (R) of all cases are very high (R > 0.99), which means that RF-RL model has a high prediction ability. The prediction ability in this study is higher than those reported in previous studies [23–25]. This may be related to the advantages of Relief, as previously mentioned, Relief is a non-parametric and non-myopic algorithm

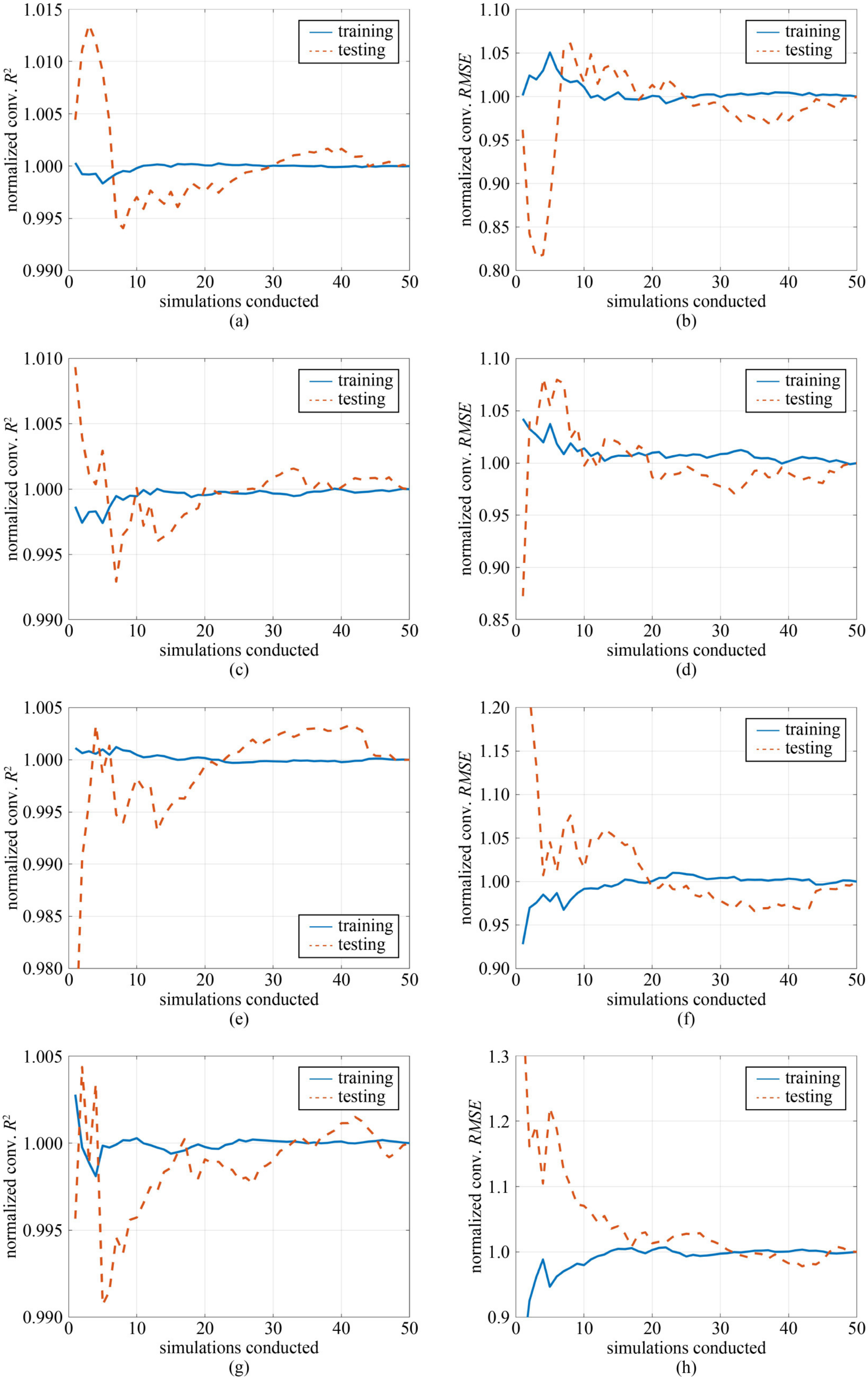


Fig. 6 Analysing of the convergence of prediction results with respect to statistical criteria of different datasets over 50 simulations: (a) R^2 of RF-RL13; (b) *RMSE* of RF-RL13; (c) R^2 of RF-RL6; (d) *RMSE* of RF-RL6; (e) R^2 of RF-RL7; (f) *RMSE* of RF-RL7; (g) R^2 of RF-RL8; (h) *RMSE* of RF-RL8.

Table 2 Summary of different datasets selected using Relief algorithm

case	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}
RF-RL13	√	√	√	√	√	√	√	√	√	√	√	√	√
RF-RL6	√	–	–	–	–	–	–	–	√	√	√	√	√
RF-RL7	–	–	√	–	√	–	√	√	√	√	–	–	√
RF-RL8	–	–	√	√	√	–	√	√	√	√	√	–	–

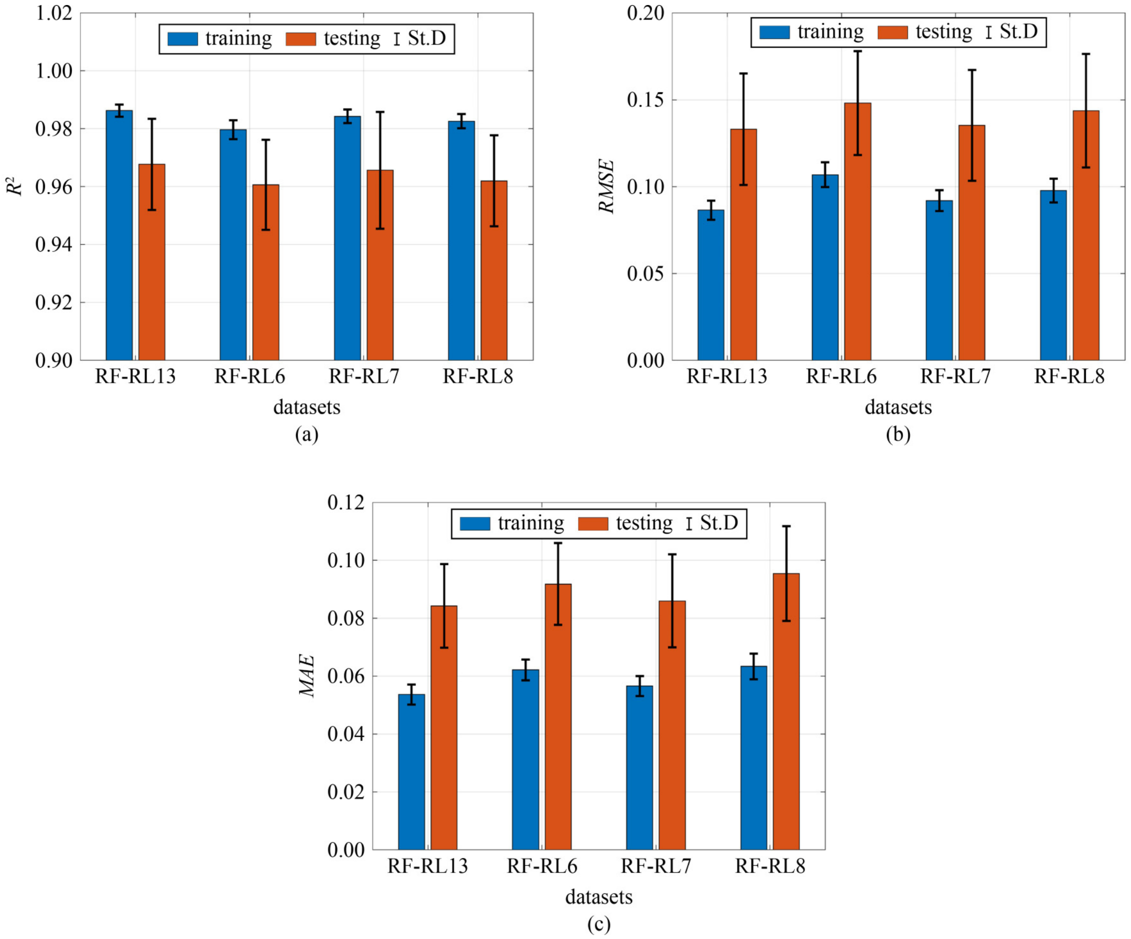


Fig. 7 Comparisons of the prediction accuracy over 50 simulations in different cases with respect to (a) R^2, (b) *RMSE*, and (c) *MAE*.

[30,31], which can reduce the limitation of RF. From Fig. 9, it can be concluded that all proposed models performed well and the model with 13 input variables has the best performance with the highest correlation coefficient R of 0.9934.

Table 4 shows the results of training and testing statistics with 4 data sets of input variables. It can be seen that the training is good, and perfect accuracy (high R^2) is obtained in all the cases.

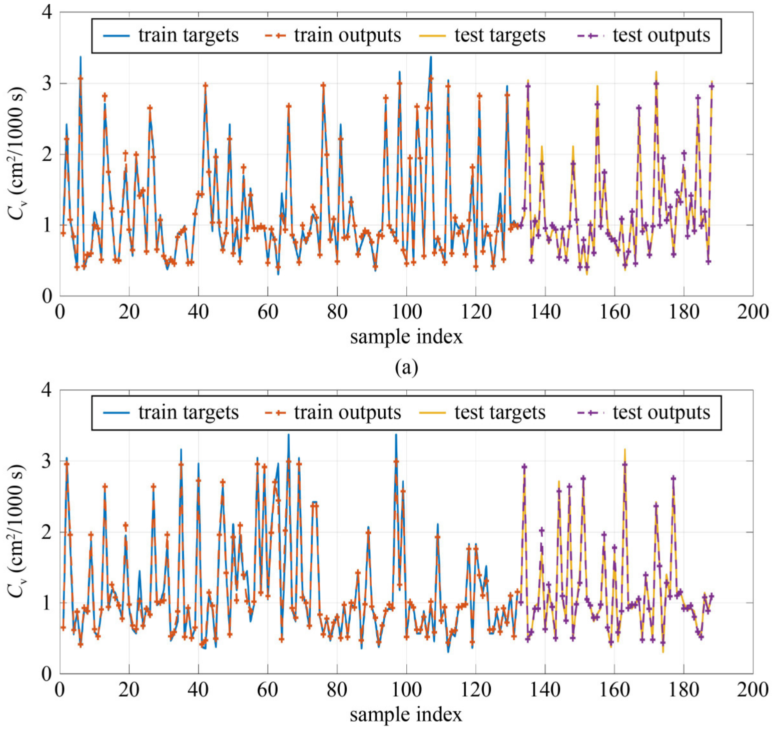
The prediction results of RF are then compared with those of well-known ML techniques, including light gradient boosting machine (LightGBM), CatBoost, and deep neural network (Deep NN). Overall, LightGBM and CatBoost are gradient boosting DTalgorithms, however, RF's nature entails the use of its bagging method to form the model, which is different from LightGBM and CatBoost. Deep NN are ANNs that, in their natural form, include many hidden layers. While the Python programming language is used to develop the LightGBM and CatBoost algorithms, the Matlab programming language is used to create Deep NN. After several trial and error tests, it was decided to use the default setting of hyperparameters of LightGBM and CatBoost from the original library, whereas the strategy to construct the Deep NN was based on the relevant literature [54]. Overall, the parameters selected to conduct the simulation are presented in Table 5.

Comparison is conducted using the 10-fold cross-validation (CV) technique. It is used to guarantee the suggested models' generality in the face of the variability in the training dataset while building it. It is also used to ensure that the model's high prediction accuracy (if any) is not “accidentally” achieved by using a specific mix of samples in the training dataset. The results for the 10-fold CV are presented in Fig. 10. LightGBM has the lowest R^2 accuracy and the highest *RMSE* values in the third, sixth, and ninth CVs, indicating that it is the most unstable algorithm for this problem. For the prediction of soil consolidation coefficient, Deep NN and CatBoost models perform quite well in terms of R^2 and *RMSE*, with numerous CV iterations obtaining greater accuracy than the suggested RF model. The RF model, on the other hand, seems to be the most stable model for this problem, as seen by its low values of standard deviation in terms of R^2 and *RMSE*.

Table 6 summarizes the prediction findings for the

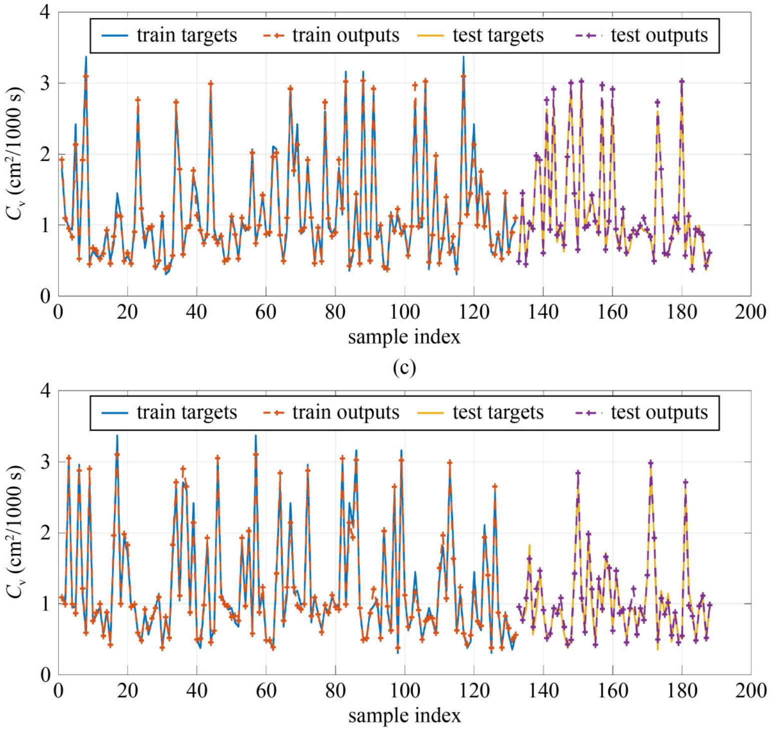
Table 3 Summary of different quality assessment criteria over 50 simulations in different cases

criteria	RF-RL13		RF-RL6		RF-RL7		RF-RL8	
	train	test	train	test	train	test	train	test
R^2								
min	0.9809	0.9085	0.9736	0.9208	0.9792	0.8827	0.9784	0.9044
average	0.9862	0.9677	0.9796	0.9606	0.9842	0.9656	0.9825	0.9620
max	0.9905	0.9856	0.9868	0.9885	0.9918	0.99	0.9893	0.9829
std	0.0021	0.0157	0.0033	0.0155	0.0024	0.0202	0.0025	0.0157
RMSE								
min	0.0714	0.0853	0.0887	0.0736	0.0687	0.0827	0.0763	0.0985
average	0.0865	0.1331	0.1069	0.1482	0.092	0.1354	0.0978	0.1437
max	0.098	0.2168	0.1184	0.2257	0.1048	0.247	0.1074	0.2418
std	0.0054	0.0321	0.0072	0.0299	0.006	0.0318	0.0068	0.0328
MAE								
min	0.0447	0.0603	0.0554	0.0478	0.0451	0.0597	0.0536	0.0683
average	0.0537	0.0842	0.0621	0.0918	0.0566	0.086	0.0634	0.0954
max	0.0602	0.1148	0.0701	0.1288	0.0652	0.1415	0.0698	0.1304
std	0.0034	0.0144	0.0035	0.0141	0.0034	0.016	0.0044	0.0163



(a)

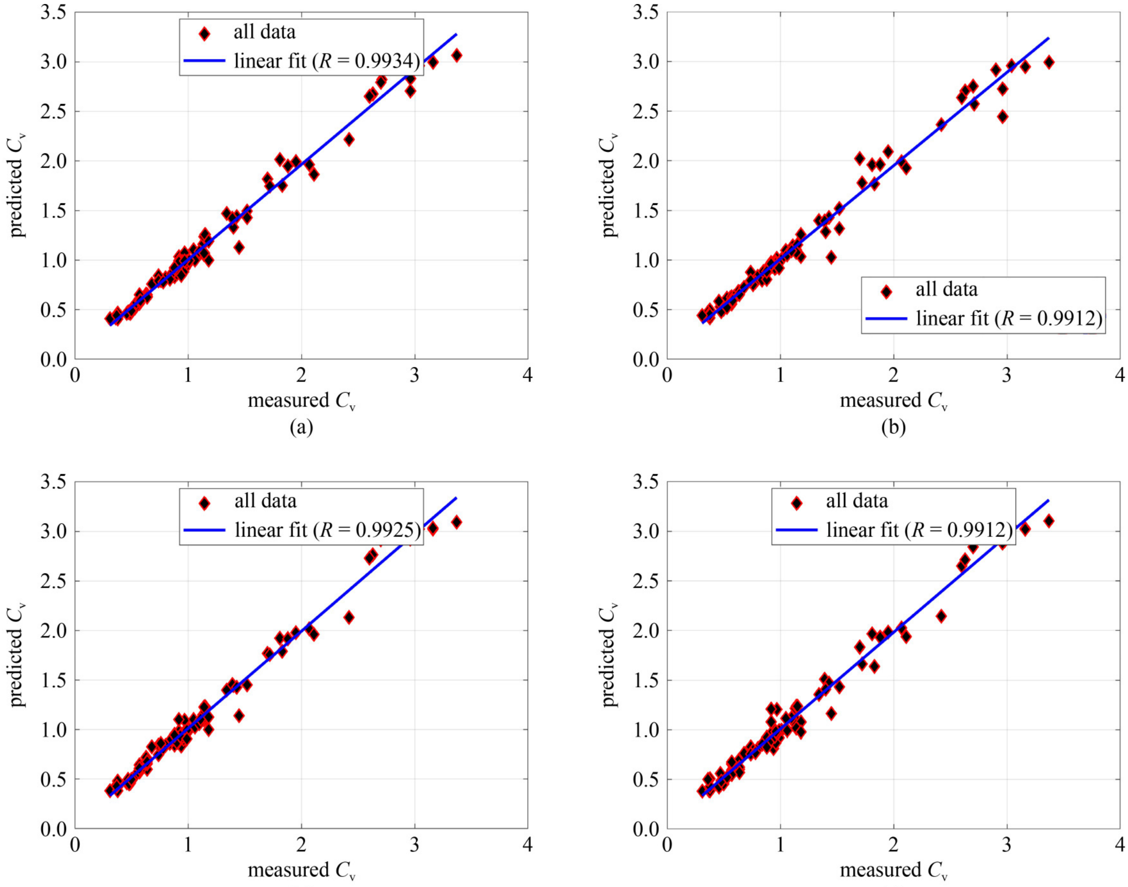
(b)



(c)

(d)

Fig. 8 Target and output values plots for training and testing datasets for the best predictor in different cases: (a) RF-RL13; (b) RF-RL6; (c) RF-RL7; and (d) RF-RL8.



(a)

(b)

(c)

(d)

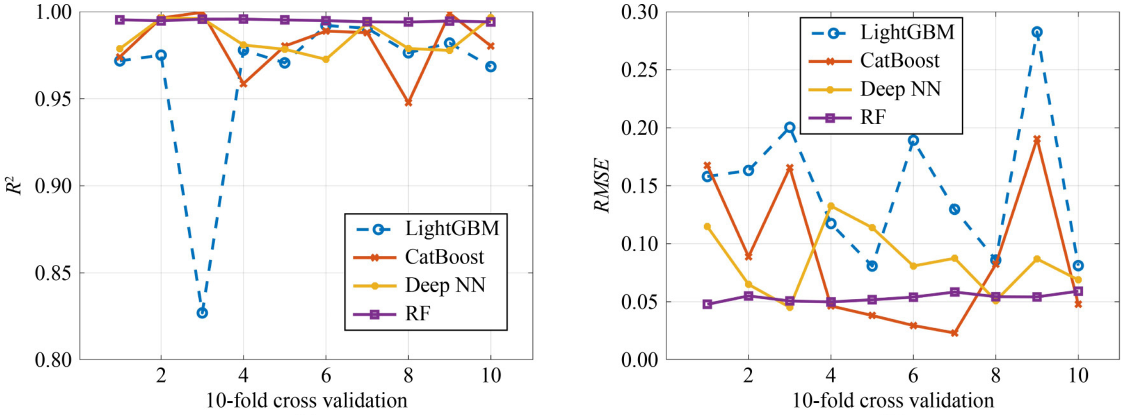
Fig. 9 Regression graphs for all data for the best predictor in different cases: (a) RF-RL13; (b) RF-RL6; (c) RF-RL7; (d) RF-RL8.

Table 4 Summary of different quality assessment criteria for the best predictor in different cases

case	set	*RMSE*	*MAE*	Err.Mean	Err.Std	R^2
RF-RL13	train	0.0865	0.0577	0.0009	0.0868	0.9868
	test	0.0880	0.0608	0.0051	0.0886	0.9856
	all data	0.0870	0.0586	0.0021	0.0872	0.9869
RF-RL6	train	0.1132	0.0701	0.0010	0.1137	0.9787
	test	0.0736	0.0478	–0.0063	0.0740	0.9885
	all data	0.1030	0.0634	–0.0012	0.1033	0.9824
RF-RL7	train	0.0932	0.0615	0.0005	0.0935	0.9841
	test	0.0856	0.0603	–0.0304	0.0808	0.9900
	all data	0.0910	0.0611	–0.0087	0.0908	0.9850
RF-RL8	train	0.0982	0.0683	0.0010	0.0985	0.9850
	test	0.0985	0.0696	–0.0183	0.0976	0.9709
	all data	0.0983	0.0687	–0.0048	0.0984	0.9825

Table 5 Summary of different parameters for the algorithms used in this study

algorithm	description of parameters
RF	Minimum number of samples to be at a leaf node = 2; Number of trees in the forest = 500; Measure of quality of split = *MSE*; Number of samples to split = 2; Number of features to consider in modeling = 13.
LightGBM	Type of boosting: Gradient Boosting DT; Maximum tree leaves = 30; No maximum tree depth; Learning rate = 0.1; Number of trees = 100.
Deep NN	Number of inputs = 13; Number of output = 1; Number of hidden layers = 3; Neurons in the three hidden layers, respectively, 20, 12, and 6 for hidden layer 1, 2, and 3; Training algorithm = Broyden–Fletcher–Goldfarb–Shanno algorithm; Leaning rate = Constant; Number of training epoch = 500; Activation function = ReLu.
CatBoost	Minimum number of samples to be at a leaf node = 1; Learning rate = 0.03; Maximum tree leaves = 64; Iterations = 1000; Evaluation metric = *RMSE*; Estimation method = Newton method.



(a) (b)

Fig. 10 Results of 10-fold cross-validation for the training part using LightGBM, CatBoost, Deep NN, and RF algorithms in this study: (a) R^2; (b) *RMSE*.

three approaches, whereas Fig. 11 depicts the corresponding regression graphs for the whole dataset. As can be seen, all three models perform well during the training phase, particularly the CatBoost algorithm (i.e., R^2 = 0.9981). On the other hand, the prediction accuracy metrics for the testing phase are lower than those for the training phase, ensuring that no overfitting occurs. The best model is Deep NN for the testing results (R^2 = 0.9791), followed by CatBoost (R^2 = 0.9788) and LightGBM (R^2 = 0.9454). With the proposed RF model's prediction accuracy and the little difference in error metrics, it is possible to infer that the RF model outperforms Deep NN, CatBoost, and LightGBM algorithms in predicting the soil consolidation coefficient.

To sum up, 188 samples may not be large, but the results of modeling indicated that a combined model of RF and Relief could predict well the consolidation coefficient of soil with very high accuracy. Because of the little difference in prediction performance between the initial dataset (13 inputs) and the reduced dataset (i.e., 6, 7, or 8 inputs), it can be concluded that the Relief approach might be used to minimize the input space.

Table 6 Summary of different quality assessment criteria for the best predictor in different cases

algorithm	set	*RMSE*	*MAE*	R^2
Deep NN	train	0.0869	0.0545	0.9837
	test	0.1211	0.0755	0.9791
	all data	0.0985	0.0609	0.9823
CatBoost	train	0.0688	0.0549	0.9981
	test	0.1251	0.0799	0.9788
	all data	0.0940	0.0624	0.9850
LightGBM	train	0.0818	0.0507	0.9871
	test	0.1833	0.1030	0.9454
	all data	0.1219	0.0666	0.9729

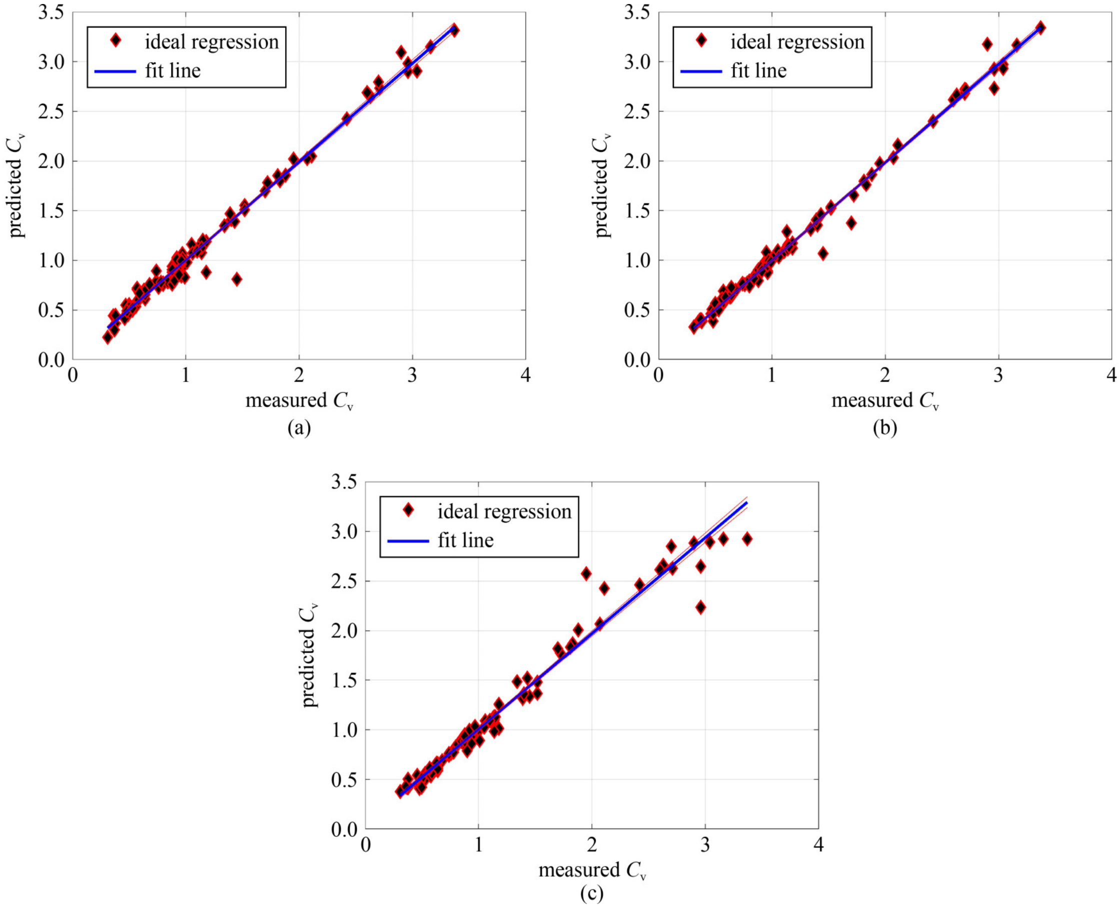


Fig. 11 Regression graphs for all data for different cases: (a) Deep NN; (b) CatBoost; (c) LightGBM.

When dealing with predicting soil parameters problem that has a large number of data points, dimensionality reduction by Relief may be quite beneficial. Thus, this hybrid model can be applied for predicting other soil parameters in further studies. Furthermore, more data should be collected and studied in future studies to validate the efficiency of this hybrid model.

Sensitivity analysis is used to understand better the impact of input characteristics on the consolidation coefficient of soil. The significance of each feature is determined by the mean of the accumulation of each tree's impurity reduction. The analysis is shown in Fig. 12, utilizing the whole input space, which contains thirteen parameters.

Clearly, I_{10} (limit liquid) is the most significant parameter affecting the consolidation coefficient of soil, followed by I_3, I_{13}, I_{12}, and I_5, which are the next four most significant factors. In good agreement with the Relief model's outcomes, input I_{10} is always kept in all three scenarios (i.e., RF-RL6, RF-RL7, RF-RL8). Inputs such as I_3, I_{13}, and I_5 are also selected by the Relief model in two scenarios, along with I_7, I_8, and I_{11}, which are the

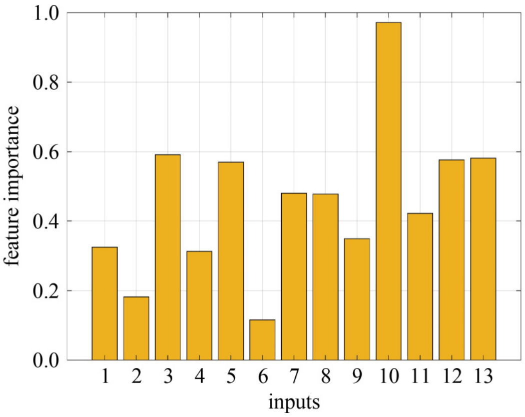


Fig. 12 Feature importance analysis conducted with 13 inputs using RF.

following three critical factors. Inputs I_6 and I_2 were discovered via feature importance analysis to be the least significant components, and these inputs are likewise removed in all three previously mentioned scenarios by the proposed Relief model. Again, it is shown that the Relief model is a good choice for dimensionality reduction in this study.

6 Conclusions

In this study, a model of RF Coupling With Relief Algorithm was employed to predict the consolidation coefficient of soil with 13 simple input variables of soils. Multi-correlation between input and output variables was carried out to understand the dependency of each input variable with output for better estimation of C_v.

The performance of the proposed models was assessed using different statistical indicators such as R^2, *RMSE*, and *MAE*. Four models of RF-RL with 6, 7, 8, and 13 input soil variables were evaluated. The proposed model results showed that all RF-RL models predicted C_v well with high accuracy, and the highest accuracy was found for the model with 13 input variables with ($R^2 = 0.9869$, $RMSE = 0.0870$, and $MAE = 0.0586$). However, thanks to the RL model, the remaining models with significant dimensionality reduction of the input space exhibit comparable prediction results, with only a slight difference in performance metrics. In addition, when compared to several benchmark ML models, such as deep neural networks, CatBoost, and LightGBM, the original RF model was more stable and effective in predicting the C_v of soil.

As a result, it can be said that the RF-RL is an excellent and inexpensive technique for predicting the consolidation coefficient of soil with high accuracy, which may be employed to estimate other vital soil parameters and geotechnical parameters such as shear strength or soil compressibility coefficient.

The refinement and growth of the ML model is a continual process that requires a great deal of study as well as extensive data collecting from various locations across the globe. This study is only conducted for a region in Vietnam; thus, it is needed to extend this study for many types of soil and other regions with a considerable number of samples to validate the finding of this paper. Furthermore, other based and hybrid models should be carried to compare with current models. In addition, other feature selection methods such as PCA, Deep autoencoder, t-distributed Stochastic Neighbor Embedding (t-SNE), Locally Linear Embedding (LLE) can be used for dimensionality reduction instead of the Relief algorithm used in this study for better performance of the ML models.

References

1. Casagrande A, Fadum R E. Notes on Soil Testing for Engineering Purposes. Cambridge, MA: Harvard University, 1940
2. Yang P, Zhang J, Hu H, Wu X, Cao X, Chang Y, Liu Y, Xu J. Coefficient analysis of soft soil consolidation based on measurement of stratified settlement. Geotechnical and Geological Engineering, 2016, 34(1): 383−390
3. Taylor D W. Research on Consolidation of Clays. Cambridge, MA: Massachusetts Institute of Technology, 1942
4. Cai G, Liu S, Puppala A J. Predictions of coefficient of consolidation from CPTU dissipation tests in quaternary clays. Bulletin of Engineering Geology and the Environment, 2012, 71(2): 337−350
5. Cai G, Liu S, Puppala A J. Consolidation parameters interpretation of CPTU dissipation data based on strain path theory for soft Jiangsu quaternary clays. Marine Georesources and Geotechnology, 2015, 33(4): 310−319
6. Raju P N, Pandian N S, Nagaraj T S. Analysis and estimation of the coefficient of consolidation. Geotechnical Testing Journal, 1995, 18(2): 252−258
7. Pistor C M, Yardimci M A, Güçeri S I. On-line consolidation of thermoplastic composites using laser scanning. Composites. Part A: Applied Science and Manufacturing, 1999, 30(10): 1149−1157
8. Sridharan A, Nagaraj H B. Coefficient of consolidation and its correlation with index properties of remolded soils. Geotechnical Testing Journal, 2004, 27: 469−474
9. Kanayama M, Rohe A, van Paassen L A. Using and improving neural network models for ground settlement prediction. Geotechnical and Geological Engineering, 2014, 32: 687−697
10. Psyllaki P, Stamatiou K, Iliadis I, Mourlas A, Asteris P, Vaxevanidis N. Surface treatment of tool steels against galling failure. In: Proceedings of the MATEC Web of Conferences. Les Ulis: EDP Sciences, 2018
11. Samaniego E, Anitescu C, Goswami S, Nguyen-Thanh V M, Guo H, Hamdia K, Zhuang X, Rabczuk T. An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications. Computer Methods in Applied Mechanics and

Engineering, 2020, 362: 112790

12. Anitescu C, Atroshchenko E, Alajlan N, Rabczuk T. Artificial neural network methods for the solution of second order boundary value problems. Computers, Materials & Continua, 2019, 59(1): 345−359
13. Nguyen-Thanh V M, Anitescu C, Alajlan N, Rabczuk T, Zhuang X. Parametric deep energy approach for elasticity accounting for strain gradient effects. Computer Methods in Applied Mechanics and Engineering, 2021, 386: 114096
14. Pham B T, Nguyen M D,Al-Ansari N, Tran Q A, Ho L S, Le H V, Prakash I. A comparative study of soft computing models for prediction of permeability coefficient of soil. Mathematical Problems in Engineering, 2021, 2021: 1−11
15. Pham B T, Ly H B, Al-Ansari N, Ho L S. A comparison of Gaussian process and M5P for prediction of soil permeability coefficient. Scientific Programming, 2021: 1−13
16. Kanungo D P, Sharma S, Pain A. Artificial Neural Network (ANN) and Regression Tree (CART) applications for the indirect estimation of unsaturated soil shear strength parameters. Frontiers of Earth Science, 2014, 8(3): 439−456
17. Khan S Z, Suman S, Pavani M, Das S K. Prediction of the residual strength of clay using functional networks. Geoscience Frontiers, 2016, 7(1): 67−74
18. Zhang W, Wu C, Zhong H, Li Y, Wang L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on bayesian optimization. Geoscience Frontiers, 2021, 12(1): 469−477
19. Mamudur K, Kattamuri M R. Application of boosting-based ensemble learning method for the prediction of compression index. Journal of The Institution of Engineers (India): Series A, 2020, 101: 409−419
20. Pham B T, Nguyen M D, Dao D V, Prakash I, Ly H B, Le T T, Ho L S, Nguyen K T, Ngo T Q, Hoang V, Son L H, Ngo H T T, Tran H T, Do N M, Van Le H, Ho H L, Tien Bui D. Development of artificial intelligence models for the prediction of compression coefficient of soil: An application of Monte Carlo sensitivity analysis. Science of the Total Environment, 2019, 679: 172−184
21. Bui D T, Nhu V H, Hoang N D. Prediction of soil compression coefficient for urban housing project using novel integration machine learning approach of swarm intelligence and multi-layer perceptron neural network. Advanced Engineering Informatics, 2018, 38: 593−604
22. Moayedi H, Gör M, Lyu Z, Bui D T. Herding behaviors of grasshopper and Harris Hawk for hybridizing the neural network in predicting the soil compression coefficient. Measurement, 2020, 152: 107389
23. Pham B T, Nguyen M D, Bui K T T, Prakash I, Chapi K, Bui D T. A novel artificial intelligence approach based on multi-layer perceptron neural network and biogeography-based optimization for predicting coefficient of consolidation of soil. Catena, 2019, 173: 302−311
24. Nguyen M D, Pham B T, Ho L S, Ly H B, Le T T, Qi C, Le V M, Le L M, Prakash I, Bui D T. Soft-computing techniques for prediction of soils consolidation coefficient. Catena, 2020, 195: 104802
25. Nguyen M D, Pham B T, Tuyen T T, Hai Yen H P, Prakash I, Vu T T, Chapi K, Shirzadi A, Shahabi H, Dou J, Quoc N K, Bui D T. Development of an artificial intelligence approach for prediction of consolidation coefficient of soft soil: A sensitivity analysis. Open Construction & Building Technology Journal, 2019, 13(1): 178−188
26. Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 2015, 71: 804−818
27. Trigila A, Iadanza C, Esposito C, Scarascia-Mugnozza G. Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). Geomorphology, 2015, 249: 119−136
28. Veronesi F, Hurni L. Random forest with semantic tie points for classifying landforms and creating rigorous shaded relief representations. Geomorphology, 2014, 224: 152−160
29. 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, Prakash I. A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil. Sustainability, 2020, 12(6): 2218
30. Windle M. Statistical Approaches to Gene X Environment Interactions for Complex Phenotypes. Cambridge, MA: MIT Press, 2016
31. Kononenko I, Sˇikonja M R. Non-Myopic Feature Quality Evaluation with (R) ReliefF. Oxford: Chapman and Hall/CRC, 2007
32. Kira K, Rendell L A. A practical approach to feature selection. In: Machine learning Proceedings 1992. Amsterdam: Elsevier, 1992: 249−256
33. Breiman L. Random forests. Machine Learning, 2001, 45(1): 5−32
34. Zhang P, Yin Z Y, Jin Y F, Chan T H. A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. Engineering Geology, 2020, 265: 105328
35. Ly H B, Thai Pham B. Soil unconfined compressive strength prediction using random forest (RF) machine learning model. Open Construction & Building Technology Journal, 2020, 14(Suppl 2): 278–285
36. Pham T D, Bui N D, Nguyen T T, Phan H C. Predicting the reduction of embankment pressure on the surface of the soft ground reinforced by sand drain with random forest regression. In: Proceedings of the IOP Conference Series: Materials Science and Engineering. Bristol: IOP Publishing, 2020: 072027
37. Durgabai R P L, YR B. Feature selection using ReliefF Algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2014: 8215−8218
38. Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm. AAAI, 1992, 2: 129−134
39. Ditterrich T G. Machine learning research: Four current directions. Artificial Intelligence Magazine, 1997, 18(4): 97−136
40. Sun Y. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1035−1051
41. Nagelkerke N J. A note on a general definition of the coefficient of

determination. Biometrika, 1991, 78(3): 691−692

42. Piepho H P. A coefficient of determination (R^2) for generalized linear mixed models. Biometrical Journal. Biometrische Zeitschrift, 2019, 61(4): 860−872

43. Wang W, Lu Y. Analysis of the mean absolute error (*MAE*) and the root mean square error (*RMSE*) in assessing rounding model. Materials Science and Engineering, 2018, 324: 012049

44. Ly H B, Le T T, Vu H L T, Tran V Q, Le L M, Pham B T. Computational hybrid machine learning based prediction of shear capacity for steel fiber reinforced concrete beams. Sustainability, 2020, 12(7): 2709

45. Willmott C J, Matsuura K. Advantages of the mean absolute error (*MAE*) over the root mean square error (*RMSE*) in assessing average model performance. Climate Research, 2005, 30: 79−82

46. Chai T, Draxler R R. Root mean square error (*RMSE*) or mean absolute error (*MAE*)?—Arguments against avoiding *RMSE* in the literature. Geoscientific Model Development, 2014, 7(3): 1525–1534

47. Le T T, Pham B T, Ly H B, Shirzadi A, Le L M. Development of 48-hour precipitation forecasting model using nonlinear autoregressive neural network. In: CIGOS 2019, Innovation for Sustainable Infrastructure. Hanoi: Springer, 2020: 1191−1196

48. Pham B T, Nguyen M D, Ly H B, Pham T A, Hoang V, Van Le H, Le T T, Nguyen H Q, Bui G L. Development of artificial neural networks for prediction of compression coefficient of soft soil. In: CIGOS 2019, Innovation for Sustainable Infrastructure. Hanoi: Springer, 2020: 1167−1172

49. Abualigah L M, Khader A T, Hanandeh E S. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. Journal of Computational Science, 2018, 25: 456−466

50. Wu Y L, Tang C Y, Hor M K, Wu P F. Feature selection using genetic algorithm and cluster validation. Expert Systems with Applications, 2011, 38(3): 2727−2732

51. Zhou Q, Zhou H, Li T. Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. Knowledge-Based Systems, 2016, 95: 1−11

52. Ly H B, Nguyen M H, Pham B T. Metaheuristic optimization of Levenberg–Marquardt-based artificial neural network using particle swarm optimization for prediction of foamed concrete compressive strength. Neural Computing & Applications, 2021, 33(24): 1−21

53. Qi C, Ly H B, Le L M, Yang X, Guo L, Pham B T. Improved strength prediction of cemented paste backfill using a novel model based on adaptive neuro fuzzy inference system and artificial bee colony. Construction & Building Materials, 2021, 284: 122857

54. Ly H B, Nguyen T A, Tran V Q. Development of deep neural network model to predict the compressive strength of rubber concrete. Construction & Building Materials, 2021, 301: 124081