Hindawi
Advances in Civil Engineering
Volume 2021, Article ID 8873993, 14 pages
https://doi.org/10.1155/2021/8873993

Research Article

Estimation of Soil Cohesion Using Machine Learning Method: A Random Forest Approach

Hai-Bang Ly, Thuy-Anh Nguyen, and Binh Thai Pham

University of Transport Technology, Hanoi 100000, Vietnam

Correspondence should be addressed to Binh Thai Pham; binhpt@utt.edu.vn

Received 30 September 2020; Revised 25 February 2021; Accepted 1 March 2021; Published 16 March 2021

Academic Editor: Guoyang Fu

Soil cohesion (C) is one of the critical soil properties and is closely related to basic soil properties such as particle size distribution, pore size, and shear strength. Hence, it is mainly determined by experimental methods. However, the experimental methods are often time-consuming and costly. Therefore, developing an alternative approach based on machine learning (ML) techniques to solve this problem is highly recommended. In this study, machine learning models, namely, support vector machine (SVM), Gaussian regression process (GPR), and random forest (RF), were built based on a data set of 145 soil samples collected from the Da Nang-Quang Ngai expressway project, Vietnam. The database also includes six input parameters, that is, clay content, moisture content, liquid limit, plastic limit, specific gravity, and void ratio. The performance of the model was assessed by three statistical criteria, namely, the correlation coefficient (R), mean absolute error (MAE), and root mean square error (RMSE). The results demonstrated that the proposed RF model could accurately predict soil cohesion with high accuracy ($R = 0.891$) and low error (RMSE = 3.323 and MAE = 2.511), and its predictive capability is better than SVM and GPR. Therefore, the RF model can be used as a cost-effective approach in predicting soil cohesion forces used in the design and inspection of constructions.

1. Introduction

The cohesion (C) of the soil is created by the bonds between the compounds, the particles, and the viscosity of the water-glue film that surrounds them. Along with the internal friction angle, the cohesion force is part of the shear resistance (slip resistance) of the cohesive soil, used to calculate the load capacity of the ground soil. Cohesion force is usually measured based on the Mohr–Coulomb theory. In the stress plane of the shear effect normal stress, the soil cohesion is the intercept on the shear axis of the Mohr–Coulomb shear resistance line [1–3]. The soil cohesion of the soil greatly depends on the composition of particles in the soil, soil texture, and moisture [4]. In the design of geotechnical constructions such as foundations, slopes, or open-pit pits, the precise determination of the soil cohesion is of great concern [5]. This important parameter can be determined in the field or laboratories [3]. Tests for soil cohesion determination are usually carried out as a direct shear test (slow cut, quick cut, and fast consolidation) or indirect soil shear test with a triaxial compressor [6]. However, the experiments to determine this parameter are often cumbersome, expensive, and time-consuming [7]. With field estimation, a team of skilled and experienced engineers is required [8–10]. To overcome the above difficulties, technical design models have been proposed based on useful correlations that exist between indicator properties obtained from field tests. Several studies have employed models to predict different soil properties and characteristics, for example, Masada's [11] study for clay and silt embankments, Mofiz and Rahman [12] for Barind soils, Cola and Cortellazo [13] for peaty soils, and Hajarwish and Shakor [14] for mudrock. However, soil is an extremely complex material, and the geological conditions in each region are different, so it is not possible to apply these models thoroughly to different regions [15]. This confirmed the need to propose a general method to be able to predict soil cohesion under different conditions.

More recently, machine learning (ML) or artificial intelligence (AI) based on computer science has gradually become popular and applied in many different fields [16–18]. The wide applications of ML have been applied in areas of

the construction industry, such as determining the critical force of steel [19]. Many dependent variables are affecting the critical force of steel [20] and the mechanical properties of the soil [21]. Therefore, the application of artificial intelligence to determine soil cohesion is completely feasible. Kovačević et al. [22] used a support vector machine (SVM) to estimate the chemical and physical properties of soil and classify soil types. Guo et al. [19] used Artificial Neural Network (ANN) and Generalized Linear Model (GLM) to predict soil aggregate stability. Moufiz and Rahman [12] used and compared different ML models, including Linear Regression (LR), ANN, SVM, random forest (RF), and M5 Tree (M5P) for prediction of Standard Penetration Test (SPT) based *N*-value of soil in the state of Haryana, India. In general, the ML models are proved as potential and highly accurate tools for the prediction of soil properties [23, 24].

In this study, the main aim of this study is to apply one of the most popular ML models, namely, random forest (RF) [25–27], for predicting the cohesion force of the soil quickly, avoiding costly and time-consuming experiments. Database of soil properties was constructed from the experimental results of the Da Nang-Quang Ngai expressway project, Vietnam. Two other ML models, namely, support vector machine (SVM) and Gaussian process regression (GPR), have been used for comparison.

2. Database Collection and Preparation

In this study, the testing results of 145 data of soil samples collected from Da Nang-Quang Ngai expressway project, located in the Central South part of Vietnam (Figure 1), were used to construct the database for modeling soil cohesion force prediction. In the modeling, we considered six input parameters, namely, clay content, moisture content, liquid limit, plastic limit, specific gravity, and void ratio, and one output parameter of soil cohesion force. The detailed determination of input and output parameters is calculated according to the formulas in the published works [28, 29].

The data in this study are randomly divided into two subsets using a uniform distribution, in which 70% of the data is used as a model training set, and 30% is used to test the performance of the model. All data are scaled to the range [0; 1] to reduce numeric error while processing with ML algorithms, as Witten et al. [30] recommended. This process ensures that the training phase of the AI models can be performed with functional generalization capabilities. Such proportions are represented by

$$x_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{1}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the considered variable and x_n is the normalized value of the variable x.

3. Modeling Approaches

3.1. Random Forest. Random forest (RF) is one of the most commonly used ML algorithms for its simplicity and variety. This is a supervised learning model used for classification and regression problems proposed by Breiman in 2001 [30]. RF is an integrated learning method that gathers results from single decision trees, thereby improving predictive efficiency through the form of majority voting or averaging results depending on each specific problem.

Suppose that there is an input data set $X = x_1, x_2, x_3, ..., x_n$ where n is the number of data dimensions or the number of predictive variables. An RF model would be a set of T trees $T_1(X)$, $T_2(X)$, $T_3(X)$,. . ., $T_n(X)$. The prediction result of these decision-making trees is $\widehat{Y}_1, \widehat{Y}_2, ..., \widehat{Y}_n$. For the regression problem, the final result of the RF model will be the average of all the prediction results of the above trees. The development of tree growing is done with the principle of dividing the initial training sets into smaller training sets, and in each split, only a few predictive variables are selected randomly. Decision trees are continuously developed without pruning to predetermined stopping criteria by the programmer. Commonly used tree growth stops are RMSE, Gini Diversity Index, or Mean Square Error. Trees with low predictive results are then discarded, and only plants with sufficient predictive value are selected in the final RF model. The random selection of predictor variables and the result set of decision trees eliminate the overfitting problem of the single decision tree model [30, 31]. The structure of the random forest is depicted in Figure 2. In this study, the RF model was trained and validated using the tools in MatLab application.

3.2. Support Vector Machine. Support vector machine (SVM), proposed by Vapnik since 1995 [32], is an effective and popular learning model for classification of linear and nonlinear regression problems. SVM machine learning model gives accurate prediction results and stable, good noise tolerance and is practical for high-dimensional feature spaces [33, 34]. Many successful SVM applications with classification and regression problems have been published in different fields [35–37]. The basic theory of SVM is summarized as follows.

A training dataset $\{(x_i, y_i), i = 1, 2, ..., N\}$ is selected for an SVM model as shown in Figure 3, where $x_i = [x_{1i}, x_{2i}, ..., x_{ni}] \in R^{n_h}$ is the input data, $y_i \in R^{n_m}$ is the output data corresponding to x_i, and N is the number of training samples. The SVM aims to find an optimal hyperplane function $f(x)$ (determined by the weight vector w and the offset b), passing through all the data elements with the insensitive loss coefficient ε (based on two supporting hyperplanes, $w.x - b = \varepsilon$ and $w.x - b = -\varepsilon$).

In the case of nonlinear regression, the function $f(x)$ is determined as follows:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b. \tag{2}$$

with

$$\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0, \quad C \geq \alpha_i, \alpha_i^* \geq 0, \forall i, \tag{3}$$

where C is the penalty constant used to control the penalty error, α_i, α_i^* are the Lagrange multipliers, and $K(x_i, x_j)$ is the kernel function defined as follows:

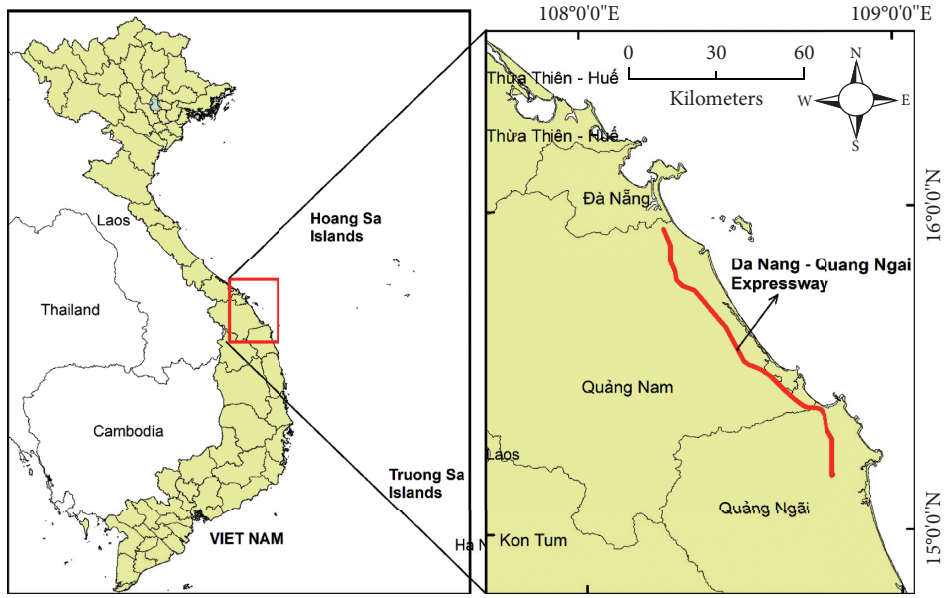


Figure 1: Location of Da Nang-Quang Ngai expressway project, Vietnam.

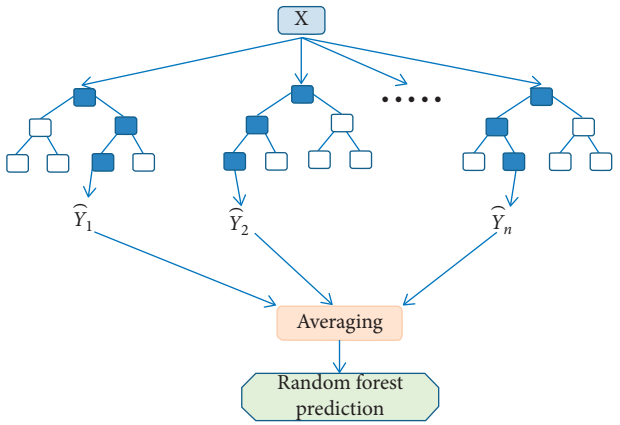


Figure 2: Random forest (RF) structure.

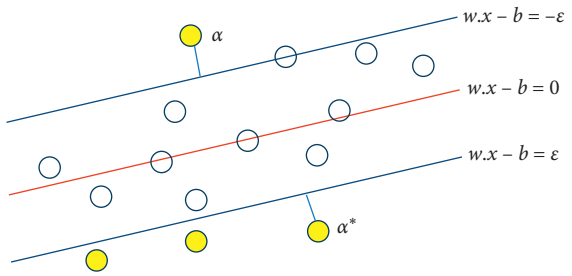


Figure 3: Support vector machine for a regression problem.

$$K\left(x_i, x_j\right) = \langle \Phi\left(x_i\right), \Phi\left(x_j\right) \rangle. \tag{4}$$

With F being a nonlinear mapping function. Linear, polynomial, sigmoid, and Gaussian functions are the most commonly used kernel functions:

$$\begin{aligned}
&\text{Linear kernel function}: K\left(x_i, x_j\right) = x_i.x_j,\\
&\text{Polynomial kernel function}: K\left(x_i, x_j\right) = \left(\gamma x_i.x_j + c\right)^d,\\
&\text{Gaussian kernel function}: K\left(x_i, x_j\right) = \exp\left(-\gamma\left(x_i - x_j\right)^2\right)^d,\\
&\text{Sigmoid kernel function}: K\left(x_i, x_j\right) = \tanh\left(\gamma x_i.x_j + c\right)^d.
\end{aligned} \tag{5}$$

3.3. Gaussian Process Regression. Gaussian process regression (GPR) is a nonparametric, Bayesian approach applied to regression problems. GPR has several advantages, working well on small datasets and having the ability to provide uncertainty measurements on the prediction values.

Given the training data set $D = \{(x_i, y_i)\}_{i=1}^N$, where N is the training set's dimension, $x_i \in R^D$ represent input data, and $y_i \in R$ is the corresponding output value. In data set D, random variables corresponding to input data set $\{x_i\}_{i=1}^N$ compose set $\{f(x_1), f(x_2), \ldots, f(x_N)\}$ and are subjected to the joint Gaussian distribution. For the simplest case, the relation between the latent function $f(x)$ and the observed target y is

$$y = f(x) + \varepsilon; \quad f(x) = x^T.w \text{ where } w \sim N(0, \Sigma_P); \varepsilon \sim N\left(0, \sigma_n^2\right), \tag{6}$$

where w denotes the weight, ε is the independent noise, σ_n^2 is the variance of the noise, and Σ_P is covariance. The distribution in the Gaussian process is represented by a mean function, denoted as $m(x)$, and a covariance kernel function, denoted as $K(x, x')$ [38]:

$$f(x) \sim GP\left[m(x), K\left(x, x'\right)\right], \tag{7}$$

where x and $x\prime \in R^D$ are random numbers of random variables. For the basic GPR, $m(x)$ is set to be zero, and formula (1) can be rewritten as

$$f(Xx) \sim GP[0, K(x, x')], \tag{8}$$

where x is the learning sample whose measure in the *GP* is the finite-dimensional distribution of the *GP*. As defined by the *GP*, the finite-dimensional distribution is a normal joint distribution as

$$[f(x_1), f(x_2), \ldots, f(x_n)]^T \sim N(m, K). \tag{9}$$

The noise e is free from $f(x)$, and it is subject to the Gaussian distribution. When $f(x)$ is an object of the Gaussian distribution, and y is also subjected to the Gaussian distribution. Then, the prior distribution of the observed target value y is inferred as:

$$y \sim N\left(0, K(x, x) + \sigma_n^2 I\right). \tag{10}$$

With given test sample points (x^*, y^*), the joint probability distribution of the observed target value y and prediction value y^* at test points is expressed as

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(x, x) + \sigma_n^2 I & K(x, x^*) \\ K(x^*, x) & K(x^*, x^*) \end{bmatrix}\right), \tag{11}$$

where $K(x, x) = (K_{ij})$ is a positive defined symmetry matrix of size $N \times N$; $K_{ij} = K(x_i, x_j)$ are the elements in the matrix, respectively, to measure the correlation of x_i and x_j; $K(x, x*)$ is the matrix of covariance of the training set and the testing set.

Applying the conditional distribution properties of the Gaussian distribution, an equation is proposed:

$$p(y^*|x, y, x^*) = N(y^*|\overline{y}^*, \operatorname{cov}(y^*)), \tag{12}$$

where

$$\begin{aligned} \overline{y}^* &= K(x, x^*)^T \left[K(x, x) + \sigma_n^2 I\right]^{-1} y, \\ \operatorname{cov}(y^*) &= K(x^*, x^*) - K(x, x^*)^T \left[K(x, x) + \sigma_n^2 I\right]^{-1} K(x, x^*). \end{aligned} \tag{13}$$

The mean value $\overline{y}^*$ is the estimation value of y^*; $\operatorname{cov}(y^*)$ is the variance matrix of test samples, which reflects the estimation value's reliability.

3.4. Model Evaluation. The application of modeling tools in the field of geotechnical engineering is increasingly popular and effective. However, to assess the ability of these models to make an accurate prediction still needs to be tested by appropriate model evaluation indicators. In this study, 3 indicators are used to evaluate the quality of the model compared to data collected from the experimental results, including mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (R) [39, 40].

MAE is calculated by Equation (2), which evaluates the difference between actual data and is calculated from the model [28]. However, it does not tell the bias trend of the predicted and experimental values. When MAE = 0, the value of the model completely coincides with the actual value, and the model is considered "ideal." MAE value is in the range (0, +∞).

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| y_{0,i} - y_{t,i} \right|. \tag{14}$$

RMSE is one of the basic quantities and is commonly used for evaluating the results of predictive models [41]. RMSE is often used to denote the mean magnitude of the error. In particular, the RMSE is extremely sensitive to large error values. Therefore, the closer the RMSE is to the MAE, the more stable the model error is. Just like MAE, RMSE also does not indicate the deviation between forecast value and actual value. RMSE is determined by formula (3), and the value of RMSE is in the range (0, +∞).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_{0,i} - y_{t,i} \right)^2}. \tag{15}$$

R is the correlation coefficient representing the data's suitability with the algorithm, a measure commonly used in ML algorithms [42]. The equation for calculating the value of R is presented in equation (4). The R values range from -1 to 1. The absolute value of R equal to 1 represents a perfect distribution between the simulated and real values, while a value of 0 indicates no correlation.

$$R = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{0,i} - \overline{y}_0 \right) \left(y_{t,i} - \overline{y}_t \right)}{\sqrt{\sum_{i=1}^{n} \left(y_{0,j} - \overline{y}_0 \right)^2 \sum_{i=1}^{n} \left(y_{t,i} - \overline{y}_t \right)^2}}}, \tag{16}$$

where n is the number of database, y_0 and $\overline{y}_0$ are the actual experimental value and the average real experimental value, and y_t and $\overline{y}_t$ are the predicted value and the average predicted value, calculated according to the model forecast.

3.5. Methodological Flowchart. The process of implementing the methodology is depicted in Figure 4, including the following basic steps:

(i) Data acquisition: in this step, soil sample data collected from the Da Nang-Quang Ngai expressway project is used to build the model. On the basis of the data set collected, determine the input and output parameters to be defined.

(ii) Database preprocessing: this is one of the most critical steps in ML to help build a more accurate ML model. Some techniques are used to process data, such as transforming data, ignoring missing values, and filling in missing values. After that, the data set is randomly divided into two parts: the training part and the testing part.

(iii) Select the model best suited to the data type: in this study, a random forest (RF) algorithm is used to estimate soil cohesion. The results of RF model are also compared with the support vector machine

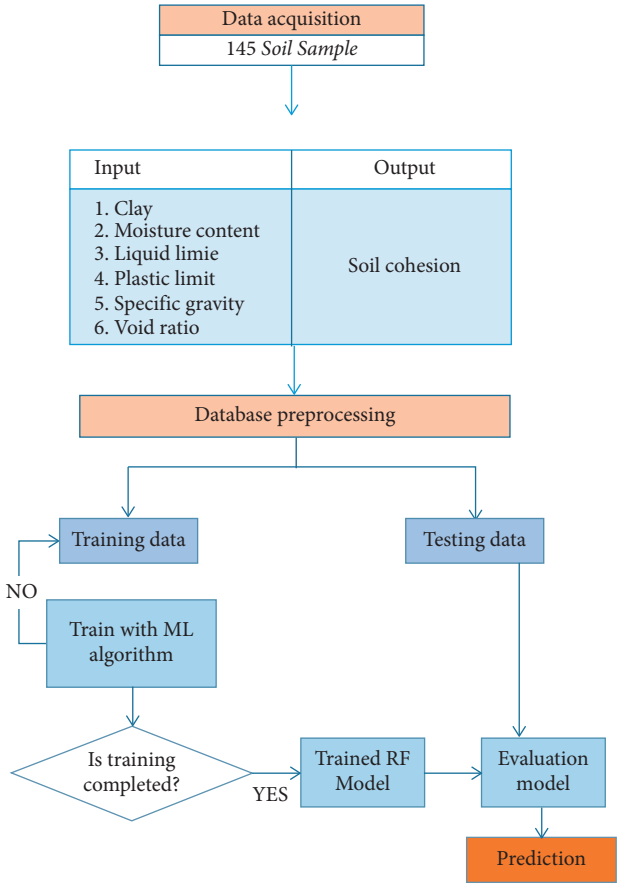


Figure 4: Methodology flow chart of the present study.

(SVM) [32] and Gaussian regression process (GPR) [43].

(iv) Train and test the model on data: in this step, train the tuple and tune the parameters using the "training database," and then test the performance on the unseen "testing database." An important point to note is that the test dataset is not used in the training process.

(v) Model evaluation: model evaluation is an indispensable part of the model development process, helping find the model to predict the best results.

4. Results and Discussion

4.1. Descriptive Statistics Analysis. The statistical analysis of the data was performed (Table 1 and Figure 5). In the database, the value of the clay content varies in the range of 4.09–47.96%, the natural moisture content is in the range of 15.53–115.41%, the liquid limit varies from 20.8 to 154.12%, the plastic limit ranges between 13.42 and 63.96%, the specific density value varies from 2.59 to 2.75 g/cm, and the void ratio ranges from 0.58–3.25. Besides, the soil cohesion values are in the range of 0.29 to 30.39 kPa. The histograms of the corresponding variables are presented in Figure 5. Besides, the quantitative analysis of input and output parameters is detailed in Table 1.

4.2. Prediction Performance of RF. In this section, the effectiveness of the RF model is evaluated. The hyperparameters of RF model are selected using trial and error tests, presented in Table 2. The comparison results between the experimental values of soil cohesion with those obtained from the RF model for the training and testing dataset are shown in Figure 6. Observe that the line representing the cohesion value of the soil is predicted to be quite close to the line representing this value experimentally. This good correlation was confirmed by the error diagram between the predicted and experimental soil cohesion for the training set (Figure 7(a)) and the testing dataset (Figure 7(b)). Of the 102 data samples of the training dataset and 43 data samples of the testing dataset, only a very few samples have an error in the range of [-7; 11] kPa. These errors show that the predictability of the RF algorithm is feasible with small errors.

Finally, the relationship between the actual data value and the predicted value is given as a regression graph in Figure 8. The quantitative values of the three criteria evaluating model performance are shown in Table 3. As shown in Table 3, the RF model provides $R = 0.90$; RMSE = 3.56; MAE = 0.90 and SD = 3.58 for the training dataset. For the testing dataset, these values are $R = 0.84$; RMSE = 2.68; MAE = 2.11; SD = 2,71, respectively. When considering all the data, the model provides $R = 0.89$; RMSE = 3.32; MAE = 2.51 and SD = 3.33. It can be seen that the predictability of the model is relatively high. Therefore, the RF model application to predict soil cohesion is feasible with high accuracy and low error.

4.3. Analysis of Simulation Convergence of RF and Other ML Models. In this work, the performance of the proposed model is assessed by the number of simulation runs. Several studies [44, 45] have shown that the predictive performance of the algorithm depends on randomly dividing the data set into training and test sets. Therefore, analysis of the model's performance should be performed with a sufficient number of simulations to demonstrate the generality of the obtained results. In this study, a total of 200 simulations were conducted to study the performance of the proposed RF model. The hyperparameters of other models are selected using trial and error tests, presented in Table 2.

Figures 9(a), 9(c), and 9(e) represent the normalized convergence values of RMSE, MAE, and R, respectively. In contrast, Figures 9(b), 9(d), and 9(f) represent the convergence values of the three respective criteria. As observed, after about 50 simulations, the oscillation of RMSE and MAE was in the range of less than 1% with the training set (Solid Green Line). With the testing set (Red dashed line), the number of simulations after about 70 times, the RMSE and MAE values fluctuate within the 1% error range. Meanwhile, the correlation coefficient R with the training set converges immediately after the first simulations. The testing set takes about 75 simulations to ensure the convergence of errors in a small range. When the number of simulations reaches 200, all RMSE, MAE, and R values are converged. It turns out that the selection of 200 simulators is suitable to get optimized results for all R, RMSE, and MAE values.

Table 1: Initial statistical analysis of the database.

Variable	Clay	Moisture content	Liquid limit	Plastic limit	Specific gravity	Void ratio	Soil cohesion
Role	Input	Input	Input	Input	Input	Input	Output
Symbol	C_l	M_c	LL	PL	δ	e	C
Unit	%	%	%	%	g/cm^3	—	kPa
Min	4.09	15.53	20.80	13.42	2.59	0.58	0.29
Median	18.73	40.67	47.35	25.35	2.68	1.25	8.33
Average	20.09	47.38	51.07	63.96	2.68	1.42	10.05
Max	47.96	115.41	154.12	63.96	2.75	3.25	30.39
SD	9.16	24.33	22.42	8.42	0.26	0.66	6.70
SK	0.69	0.88	2.09	1.72	−0.10	0.82	1.24

SD = standard deviation; SK = skewness.

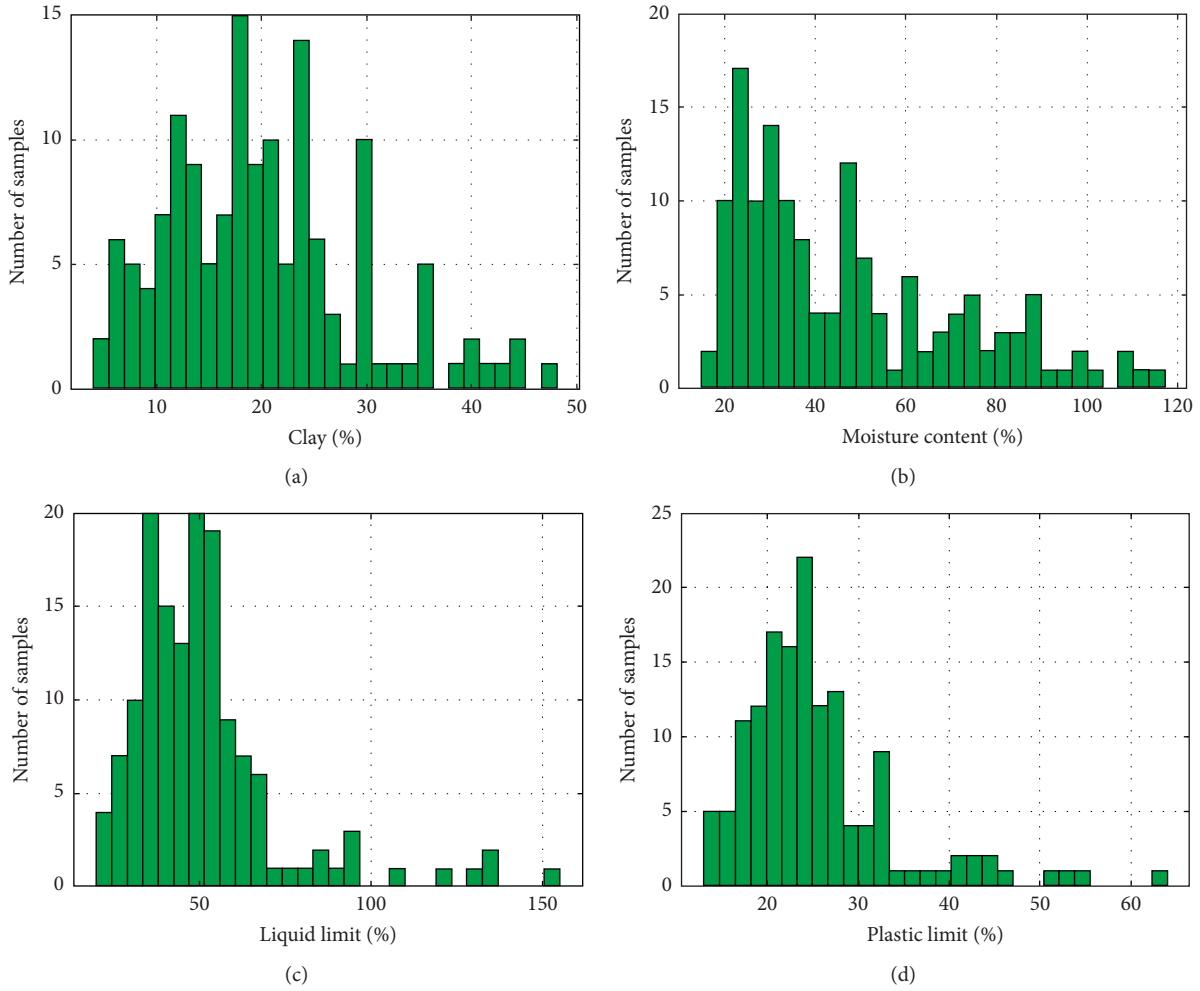


Figure 5: Continued.

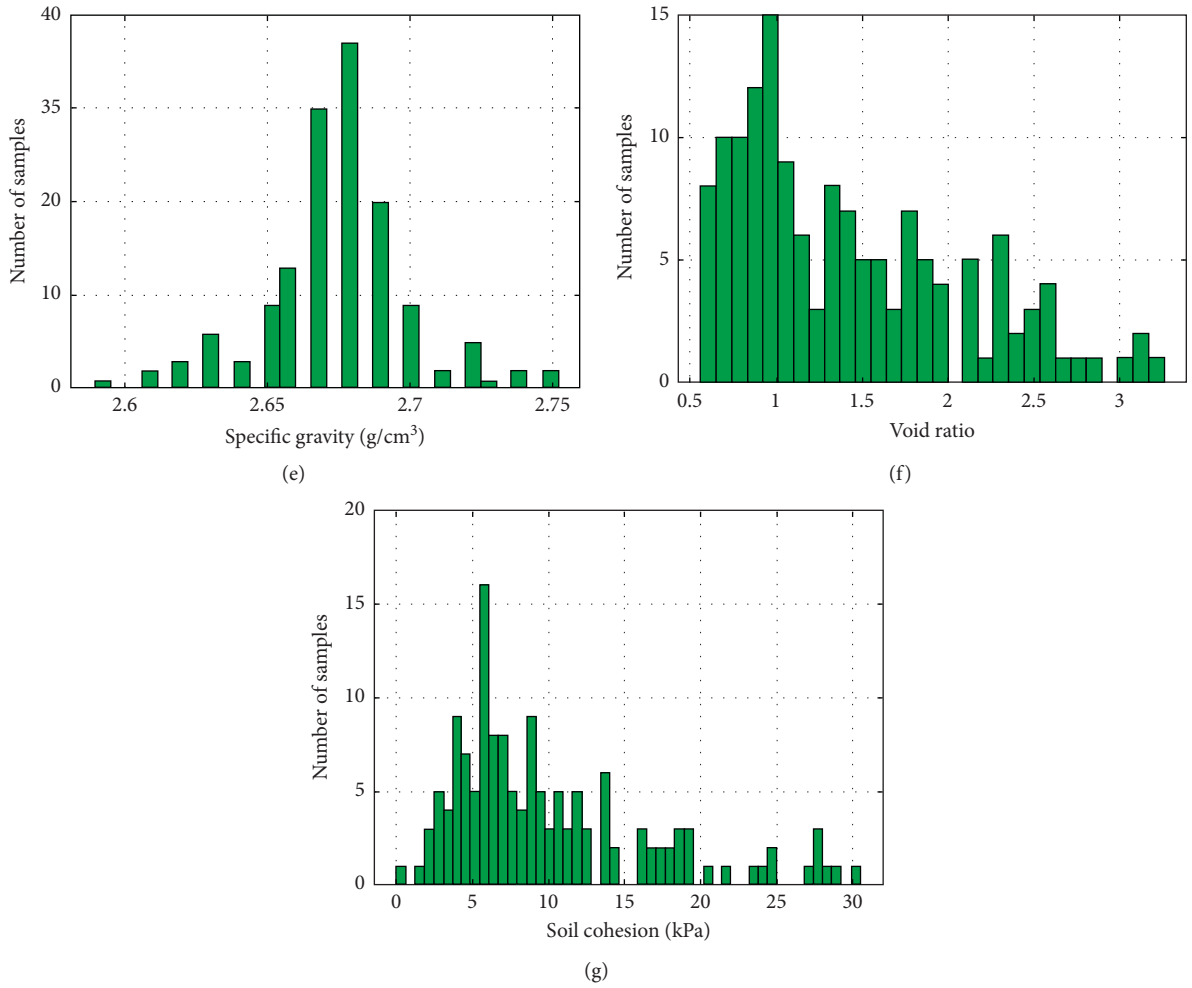


Figure 5: Histograms of the input and output variables used for the development of the RF algorithm: (a) clay content; (b) moisture content; (c) liquid limit; (d) plastic limit; (e) specific gravity; (f) void ratio; (g) soil cohesion.

Table 2: Hyperparameters of ML methods used in this study.

ML methods	Hyperparameters description
RF	Using *TreeBagger* MatLab function A number of 500 trees Minimum leaf size of 5
SVM	Using *fitrsvm* MatLab function Using hyperparameter optimization that minimize 10-fold cross-validation The 6 hyperparameters are box constraint, kernel function, kernel scale parameter, polynomial kernel function order, half the width of the epsilon-insensitive band, standardize method for data
ANN	Using *fitrrgp* MatLab function Using hyperparameter optimization that minimizes 10-fold cross-validation The 5 hyperparameters are basis function, kernel function, kernel scale, sigma value, standardize method for data

Figure 10 shows a box plot illustration of RMSE, MAE, and R values after 200 runs corresponding to the training and testing sets simulated by RF algorithm. The mean and corresponding standard deviations of R are 0.90 and 0.01 for the training dataset. For the testing dataset, these values are 0.71 and 0.08, respectively. Considering the RMSE criterion, the mean and standard deviation are 3.25 and 0.16, respectively, for the training dataset, and 4.73 and 0.65 for the testing set. For MAE, these values are 2.37 and 0.13, respectively, corresponding to the training set, and 3.54 and 0.48 for the testing set. Besides, the minimum and maximum values of R, RMSE, and MAE for the two data sets are shown in Table 4. In addition, 200 simulations with SVM and GPR algorithms are performed and presented in Figure 10. It

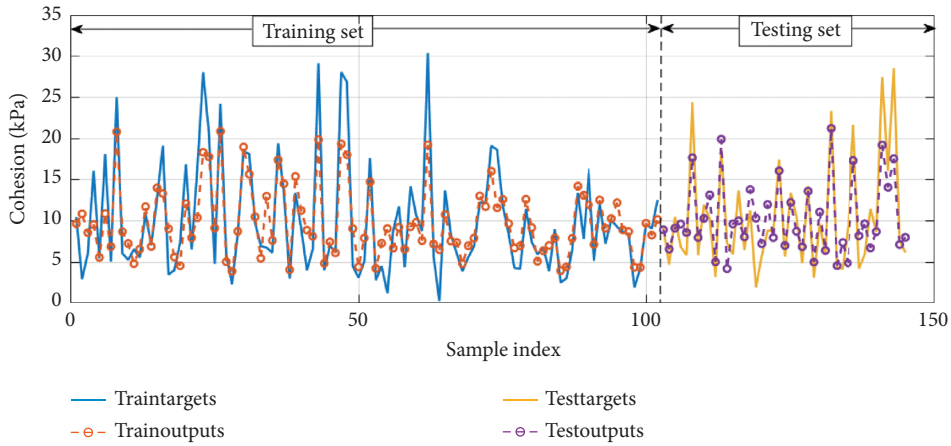


Figure 6: Comparisons of the training and testing datasets with the predicted and experimental values of soil cohesion.

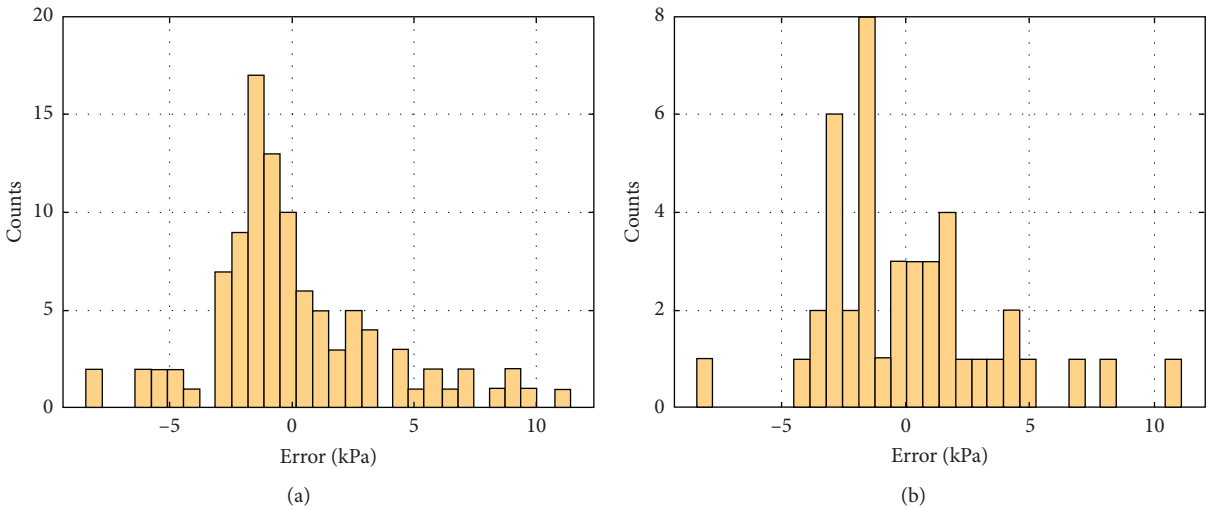


Figure 7: Error histogram analysis of RF using (a) training dataset and (b) testing dataset.

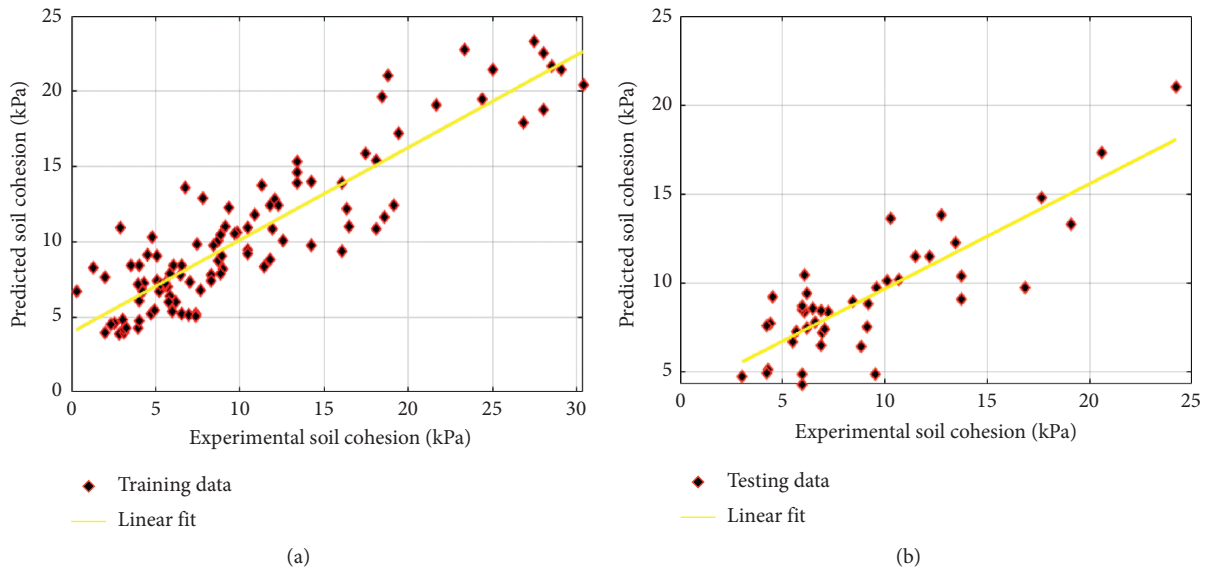
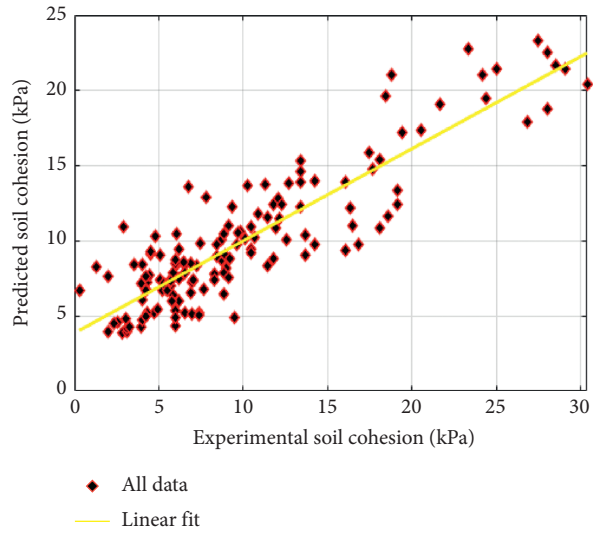


Figure 8: Continued.



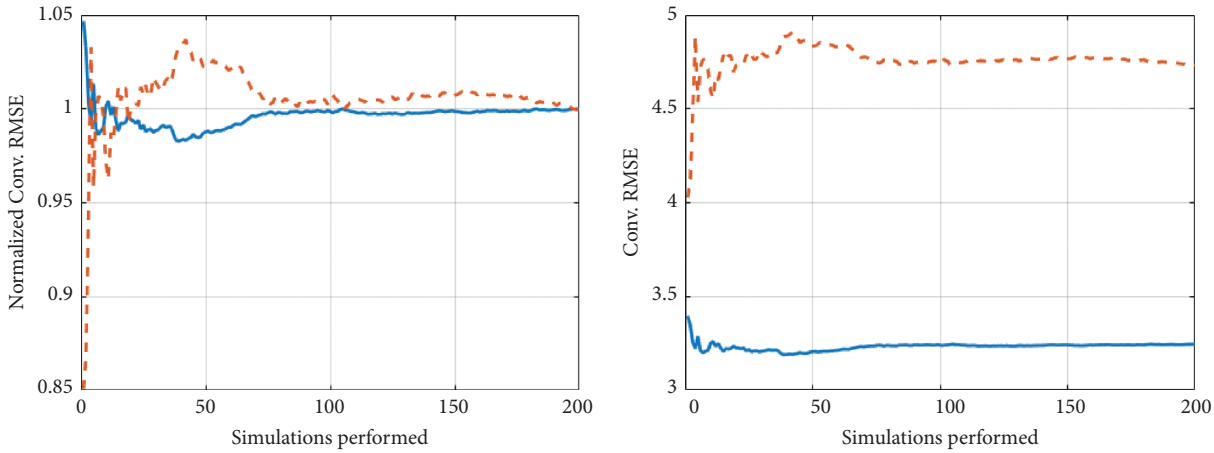
(c)

Figure 8: Regression analysis of RF with respect to (a) training dataset, (b) testing dataset, and (c) all dataset.

Table 3: Summary of prediction results of the RF model in terms of RMSE, MAE, and *R*.

Indicators	RMSE	MAE	SD	R
Training set	3.5585	0.8997	3.5757	0.8997
Testing set	2.6817	2.1097	2.7132	0.8370
All data	3.3227	2.5110	3.3341	0.8906

SD = standard deviation.



(a) (b)

Figure 9: Continued.

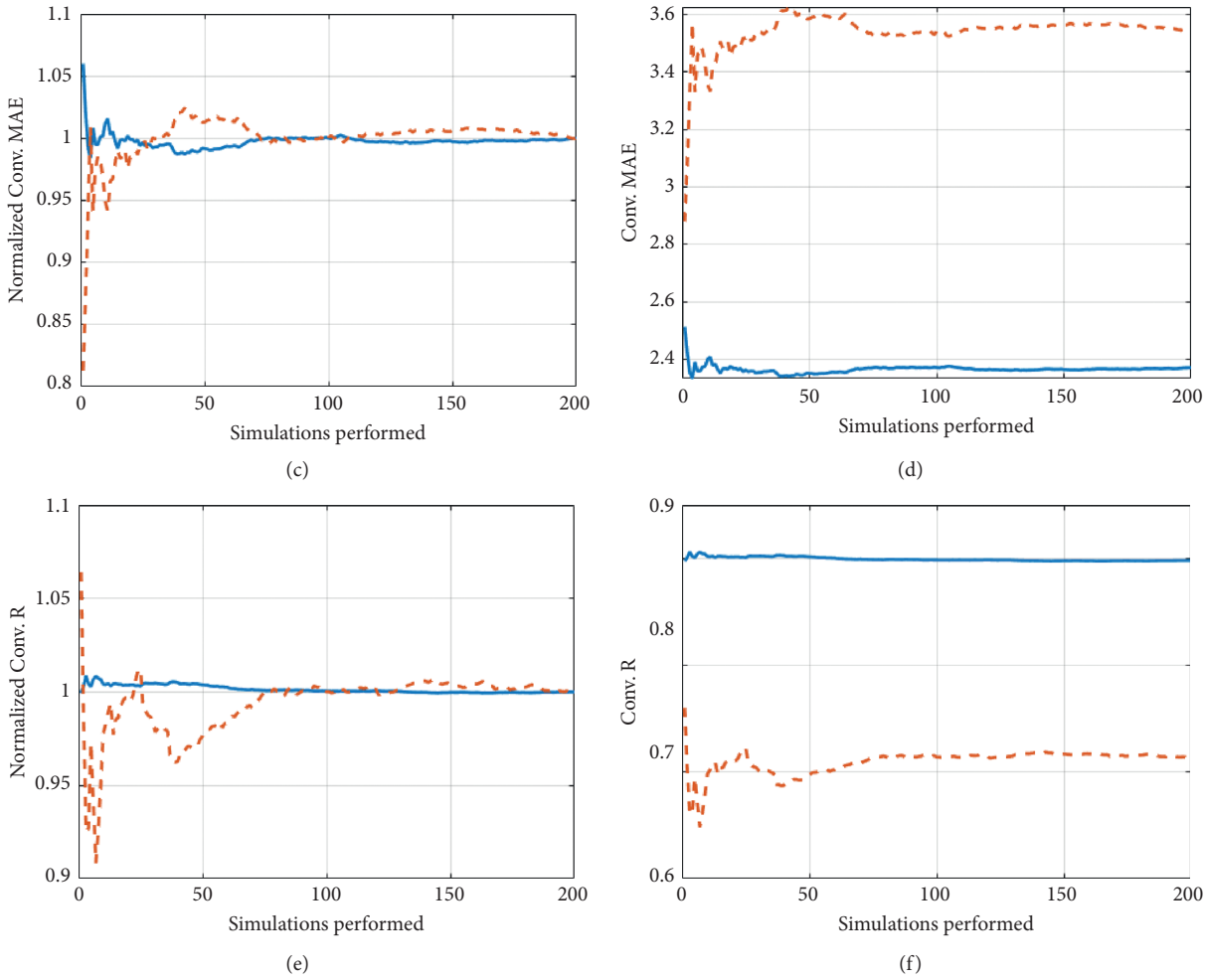


Figure 9: Analysis of simulation convergence over 200 runs with respect to (a) normalized convergence values of RMSE, (b) convergence values of RMSE, (c) normalized convergence values of MAE, (d) converged values of MAE, (e) normalized convergence values of R, and (f) converged values of R.

could be easily observed that RF model outperforms other algorithms on both the training and testing datasets. The average R values of RF are significantly higher than those of SVM ($R = 0.27$) and GPR ($R = 0.69$) for the training parts, whereas the average RMSE and MAE values of RF are lower than those of SVM (RMSE = 8.14, MAE = 7.18) and GPR (RMSE = 4.83, MAE = 3.60). Similar observations are noticed for the testing parts (RMSE = 5.46, MAE = 4.00 for SVM, and RMSE = 5.00, MAE = 3.72 for GPR), which reflect the prediction capability of the models.

Overall, the proposed RF algorithm is a better ML model compared with other ML models (SVM, GPR) in predicting soil cohesion. It is reasonable because RF has many advantages such as the following: (i) it can be effectively applied to large-scale datasets as it provides the facility for size reduction without deleting unwanted variables from the training dataset; (ii) it can handle thousands of input features and variables at a time; (ii) it has an embedded efficient technique for estimating missing or null values. Hence, it is possible to maintain a level of accuracy (i.e., consistent performance) even when a large portion of the data is missing; (iv) it is able to perform a good parallel simulation because the number of trees generated and computed is completely independent of each other; and (v) this model can minimize errors as the results are synthesized from different "learners" (random forest trees) [46]. The results of this study are also comparable with other previous published works [46–48].

4.4. Sensitivity Analysis. In this section, the estimation of the feature importance of input variables is performed. For each simulation, the importance value is calculated by the sum of the difference taken by the splits of the given predictor and divided by the sum of the branch in RF. Figure 11 shows the out-of-bag feature importance over 200 simulations (by mean values) along with the standard deviation values. It can be seen that the void ratio is the most important variable in predicting soil cohesion. Besides, the moisture content is the second important input for the problem, followed by the plastic limit, liquid limit, specific gravity, and the clay

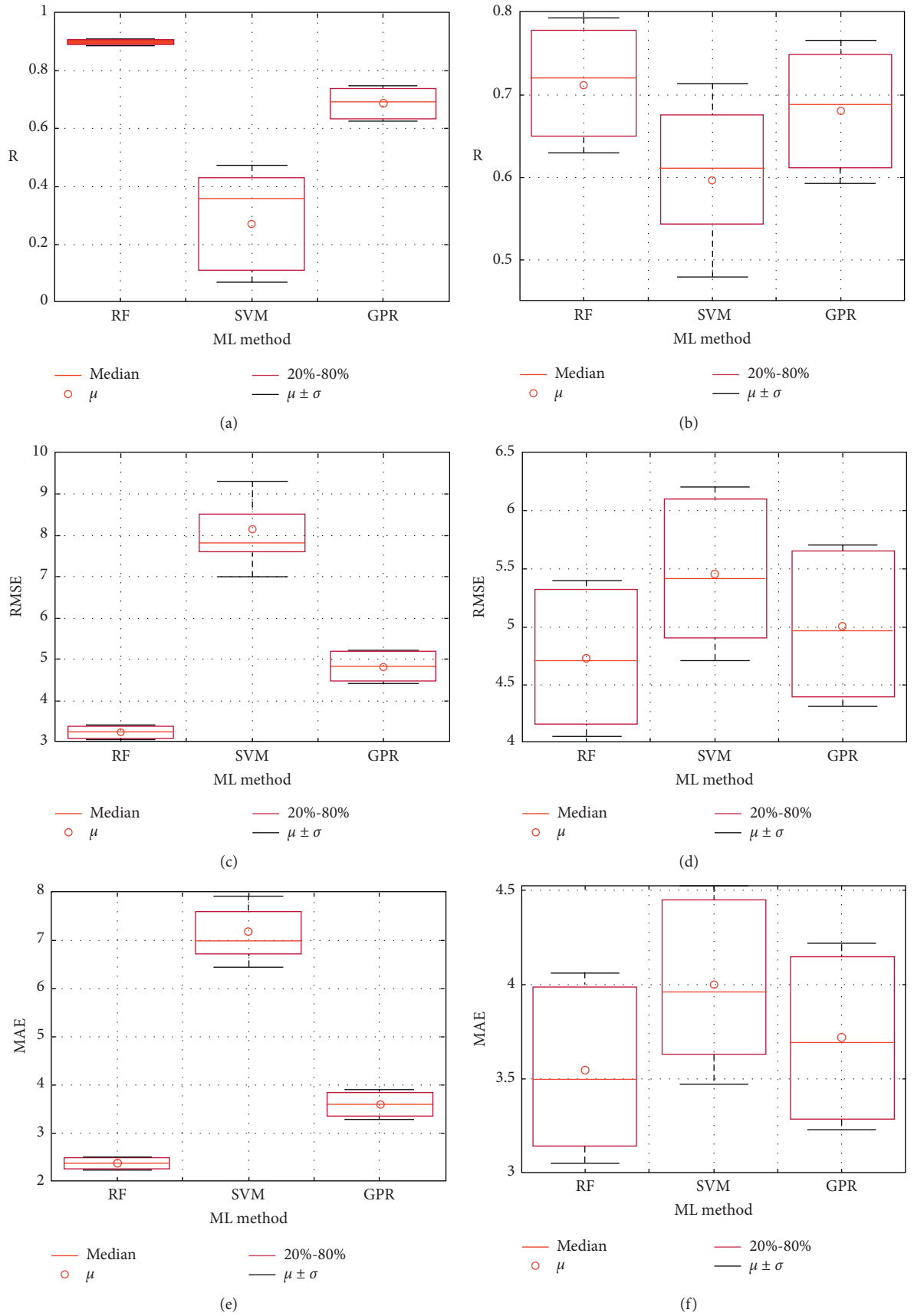


FIGURE 10: Box-plot of the prediction results over 200 simulations for the 3 ML algorithms (a) R for the training dataset, (b) R for the testing dataset, (c) RMSE for the training dataset, (d) RMSE for the testing dataset, (e) MAE for the training dataset, and (f) RMSE for the training dataset. The symbol μ denotes the mean values, σ denotes the standard deviation, and the 20%–80% denotes the quantile level of the results, respectively, from 20% to 80%.

Table 4: Statistical analysis of prediction results over 200 RF simulations in terms of RMSE, MAE, and R.

Datasets	Training set			Testing set		
Criteria	RMSE	MAE	R	RMSE	MAE	R
Min	2.7814	2.0009	0.8702	2.6817	2.1097	0.3084
Average	3.2450	2.3702	0.8982	4.7317	3.5404	0.7140
Median	3.2635	2.3770	0.8974	4.7337	3.5521	0.7264
Max	3.5852	2.6802	0.9333	6.3457	4.8180	0.8702
SD	0.1599	0.1314	0.0103	0.6498	0.4807	0.0844
SK	−0.3224	−0.0607	0.3887	−0.0614	−0.0437	−1.3655

SD = standard deviation; SK = skewness.

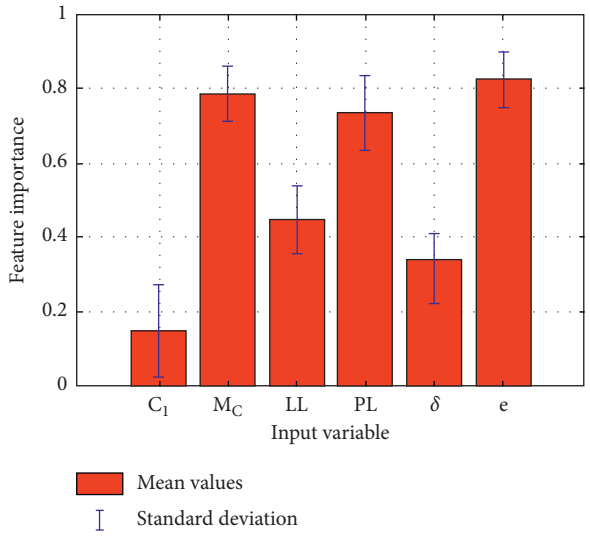


Figure 11: Out-of-bag feature importance calculated using RF algorithm.

content. These sensitivity results are reasonable and comparable with other published works [28, 49, 50].

5. Conclusion

In this study, a data set of 145 soil samples collected from the Da Nang-Quang Ngai expressway project was used to construct an RF model for the purpose of soil cohesion prediction. Input data for network training includes clay, moisture content, liquid limit, plastic limit, specific gravity, and void ratio. Three statistical criteria, namely, correlation coefficient (R), mean absolute error (MAE), and root mean square error (RMSE), are used to evaluate the correlation between the values predicted by the RF model and actual experimental values. The analysis results show that the built model can predict soil cohesion accurately and quickly, avoiding costly and difficult experiments that require complicated equipment.

However, in ML problems, data is the key factor in creating a reliable predictive tool. Therefore, the next research direction is to collect additional data to further improve the algorithm, making the prediction more accurate, avoiding costly on-field experiments.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Ministry of Transport, project titled "Building Big Data and Development of ML Models Integrated with Optimization Techniques for Prediction of Soil Shear Strength Parameters for Construction of Transportation Projects" under grant number DT 203029. We thank the ones who have supported us with the additional data for carrying out this research.

References

[1] V. N. S. Murthy, *Geotechnical Engineering: Principles and Practices of Soil Mechanics*, Taylor & Francis CRC Press, Florida, FL, USA, 2nd edition, 2008.

[2] M. Alsaleh, "Numerical modeling of strain localization in granular materials using Cosserat theory enhanced with microfabric properties," *Doctoral dissertations*, LA. ouisiana State University, Baton Rouge, LA, US, 2004.

[3] A. Mollahasani, A. H. Alavi, A. H. Gandomi, and A. rashed, "Nonlinear neural-based modeling of soil cohesion intercept," *KSCE Journal of Civil Engineering*, vol. 15, no. 5, pp. 831–840, 2011.

[4] M. Hosseini, S. A. R. Movahedi Naeini, A. A. Dehghani et al., "Modeling of soil mechanical resistance using intelligent methods," *Journal of Soil Science and Plant Nutrition*, vol. 18, no. ahead, pp. 939–951, 2018.

[5] S. M. Mousavi, A. H. Alavi, A. H. Gandomi, and A. Mollahasani, "Nonlinear genetic-based simulation of soil shear strength parameters," *Journal of Earth System Science*, vol. 120, no. 6, p. 1001, 2012.

[6] S. Havaee, M. R. Mosaddeghi, and S. Ayoubi, "In situ surface shear strength as affected by soil characteristics and land use in calcareous soils of Central Iran," *Geoderma*, vol. 237-238, pp. 137–148, 2015.

[7] A. Besalatpour, M. A. Hajabbasi, S. Ayoubi, M. Afyuni, A. Jalalian, and R. Schulin, "Soil shear strength prediction using intelligent systems: artificial neural networks and an adaptive neuro-fuzzy inference system," *Soil Science and Plant Nutrition*, vol. 58, no. 2, pp. 149–160, 2012.

[8] S. Roy and G. Dass, "Statistical models for the prediction of shear strength parameters at sirsa, India," *International Journal of Civil and Structural Engineering*, vol. 4, pp. 483–498, 2014.

[9] H. Ersoy, M. B. Karsli, S. Çellek, B. Kul, İ. Baykan, and R. L. Parsons, "Estimation of the soil strength parameters in tertiary volcanic regolith (NE Turkey) using analytical hierarchy process," *Journal of Earth System Science*, vol. 122, no. 6, pp. 1545–1555, 2013.

[10] J. Arvidsson and T. Keller, "Comparing penetrometer and shear vane measurements with measured and predicted mouldboard plough draught in a range of Swedish soils," *Soil and Tillage Research*, vol. 111, no. 2, pp. 219–223, 2011.

[11] T. Masada, *Shear Strength of Clay and Silt Embankments*, U.S. Department of Transportation, Federal Highway Administration. State Job Number 134319 (0), , p. 134319, 2009 .

[12] S. A. Mofiz and M. M. Rahman, *Shear Strength Behavior of Barind Soil on Triaxial Extension Stress Path Tests*, 2010.

[13] S. Cola and G. Cortellazzo, "The shear strength behavior of two peaty soils," *Geotechnical and Geological Engineering*, vol. 23, no. 6, pp. 679–695, 2005.

[14] A. Hajdarwish and A. Shakoor, *Predicting the Shear Strength Parameters of Mudrocks*, 2006.

[15] N. Puri, H. D. Prasad, and A. Jain, "Prediction of geotechnical parameters using machine learning techniques," *Procedia Computer Science*, vol. 125, pp. 509–517, 2018.

[16] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, p. 2020, 2020.

[17] H. Storm, K. Baylis, and T. Heckelei, "Machine learning in agricultural and applied economics," *European Review of Agricultural Economics*, vol. 47, no. 3, pp. 849–892, 2020.

[18] K. M. F. Elsayed, T. Ismail, and N. S. Ouf, "A review on the relevant applications of machine learning in agriculture," *Ijireeice*, vol. 6, no. 8, pp. 1–17, 2018.

[19] S. Guo, J. Yu, X. Liu, C. Wang, and Q. Jiang, "A predicting model for properties of steel using the industrial Big data based on machine learning," *Computational Materials Science*, vol. 160, pp. 95–104, 2019.

[20] H.-B. Ly, L. M. Le, H. T. Duong et al., "Hybrid artificial intelligence approaches for predicting critical buckling load of structural members under compression considering the influence of initial geometric imperfections," *Applied Sciences*, vol. 9, no. 11, p. 2258, 2019.

[21] S. Kiran, B. Lal, and S. S. Tripathy, "Shear strength prediction of soil based on probabilistic neural network," *Indian Journal of Science and Technology*, vol. 9, no. 41, 2016.

[22] M. Kovačević, B. Bajat, and B. Gajić, "Soil type classification and estimation of soil properties using support vector machines," *Geoderma*, vol. 154, pp. 340–347, 2010.

[23] M. Zeraatpisheh, S. Ayoubi, A. Jafari, and P. Finke, "Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran," *Geomorphology*, vol. 285, pp. 186–204, 2017.

[24] M. Zeraatpisheh, S. Ayoubi, A. Jafari, S. Tajik, and P. Finke, "Digital mapping of soil properties using multiple machine learning in a semi-arid region, Central Iran," *Geoderma*, vol. 338, pp. 445–452, 2019.

[25] T. Phạm, H.-B. Ly, T. Van Quan et al., "Prediction of pile axial bearing capacity using artificial neural network and random forest," *Applied Sciences*, vol. 10, p. 1871, 2020.

[26] A. Shaqadan, "Prediction of concrete mix strength using random forest model," vol. 11, pp. 11024–11029, 2016.

[27] T. Singh, M. Pal, and V. K. Arora, "Modeling oblique load carrying capacity of batter pile groups using neural network, random forest regression and M5 model tree," *Frontiers of Structural and Civil Engineering*, vol. 13, no. 3, pp. 674–685, 2019.

[28] H.-B. Ly and B. T. Pham, "Prediction of shear strength of soil using direct shear test and support vector machine model," *The Open Construction and Building Technology Journal*, vol. 14, no. 1, p. 41, 2020.

[29] B. Pham, M. Nguyen, H.-B. Ly et al., "Development of artificial neural networks for prediction of compression coefficient of soft soil," pp. 2366–2557, 2019.

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

[31] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77–89, 2017.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[33] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classification*, Taipei, China, 2003.

[34] B. Fowler, "A sociological analysis of the satanic verses affair," *Theory, Culture and Society*, vol. 17, no. 1, pp. 39–61, 2000.

[35] N. Barakat and A. P. Bradley, "Rule extraction from support vector machines: a review," *Neurocomputing*, vol. 74, no. 1-3, pp. 178–190, 2010.

[36] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, and B. Baesens, "Rule extraction from support vector machines: an overview of issues and application in credit scoring," *Rule Extraction from Support Vector Machines*, pp. 33–63, 2008.

[37] V. Uslan and H. Seker, "Support vector-based takagi-sugeno fuzzy system for the prediction of binding affinity of peptides," in *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4062–4065, IEEE, Osaka, Japan, July 2013.

[38] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang, "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 4, pp. 825–848, 2008.

[39] C. Qi, L. Guo, H.-B. Ly, H. V. Le, and B. T. Pham, "Improving pressure drops estimation of fresh cemented paste backfill slurry using a hybrid machine learning method," *Minerals Engineering*, vol. 163, Article ID 106790, 2021.

[40] T.-A. Nguyen, H.-B. Ly, and B. T. Pham, "Backpropagation neural network-based machine learning model for prediction of soil friction angle," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8845768, 11 pages, 2020.

[41] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[42] Z. M. Yaseen, I. Ebtehaj, H. Bonakdari et al., "Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model," *Journal of Hydrology*, vol. 554, pp. 263–276, 2017.

[43] J. Bernardo, J. Berger, A. Dawid, and A. Smith, "Regression and classification using Gaussian process priors," *Bayesian Statistics*, vol. 6, p. 475, 1998.

[44] B. T. Pham, M. D. Nguyen, D. V. Dao et al., "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*, vol. 679, pp. 172–184, 2019.

[45] D. Dao, H.-B. Ly, S. Trinh, T.-T. Le, and B. Pham, "Artificial intelligence approaches for prediction of compressive strength of geopolymer concrete," *Materials*, vol. 12, no. 6, p. 983, 2019.

[46] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," *Journal of Petroleum Science and Engineering*, vol. 174, pp. 776–789, 2019.

[47] H.-M. Lu, J.-S. Chen, and W.-C. Liao, "Nonparametric regression via variance-adjusted gradient boosting Gaussian process regression," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2020.

[48] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geology Reviews*, vol. 71, pp. 804–818, 2015.

[49] N. A. Al-Shayea, "The combined effect of clay and moisture content on the behavior of remolded unsaturated soils," *Engineering Geology*, vol. 62, no. 4, pp. 319–342, 2001.

[50] A. M. Mouazen, H. Ramon, and J. D. Baerdemaeker, "SW-soil and water," *Biosystems Engineering*, vol. 83, no. 2, pp. 217–224, 2002.