Hindawi
Complexity
Volume 2022, Article ID 8089428, 18 pages
https://doi.org/10.1155/2022/8089428

Research Article

Predicting and Investigating the Permeability Coefficient of Soil with Aided Single Machine Learning Algorithm

Van Quan Tran

University of Transport Technology, No. 54 Trieu Khuc Street, Thanh Xuan District, Hanoi, Vietnam

Correspondence should be addressed to Van Quan Tran; quantv@utt.edu.vn

Received 19 July 2022; Revised 1 September 2022; Accepted 7 September 2022; Published 25 September 2022

Academic Editor: Xiaoan Yan

The permeability coefficient of soils is an essential measure for designing geotechnical construction. The aim of this paper was to select a highest performance and reliable machine learning (ML) model to predict the permeability coefficient of soil and quantify the feature importance on the predicted value of the soil permeability coefficient with aided machine learning-based SHapley Additive exPlanations (SHAP) and Partial Dependence Plot 1D (PDP 1D). To acquire this purpose, five single ML algorithms including K-nearest neighbors (KNN), support vector machine (SVM), light gradient boosting machine (LightGBM), random forest (RF), and gradient boosting (GB) are used to build ML models for predicting the permeability coefficient of soils. Performance criteria for ML models include the coefficient of correlation R^2, root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). The best performance and reliable single ML model for predicting the permeability coefficient of soil for the testing dataset is the gradient boosting (GB) model, which has $R^2 = 0.971$, $\mathrm{RMSE} = 0.199 \times 10^{-11}$ m/s, $\mathrm{MAE} = 0.161 \times 10^{-11}$ m/s, and $\mathrm{MAPE} = 0.185\%$. To identify and quantify the feature importance on the permeability coefficient of soil, sensitivity studies using permutation importance, SHapley Additive exPlanations (SHAP), and Partial Dependence Plot 1D (PDP 1D) are performed with the aided best performance and reliable ML model GB. Plasticity index, density > water content, liquid limit, and plastic limit > clay content > void ratio are the order effects on the predicted value of the permeability coefficient. The plasticity index and density of soil are the first priority soil properties to measure when assessing the permeability coefficient of soil.

1. Introduction

One of the most fundamental elements that governs the fluid flow properties of soil is its permeability. Permeability is defined as the quantity of water that passes through the interconnected voids of a soil mass in a given period, and it may be measured using field and laboratory techniques. The permeability of soil is an essential component in the construction of most civil engineering projects that are built on soil, such as highways, tunnels, and dams [1]. The permeability coefficient of soil k is a coefficient that assesses the capacity of liquids to flow through interconnected spaces in soil under hydraulic gradients ranging from high to low value [2].

It is worth noting that the permeability coefficient of soil k is one of the important properties of soil that need to be calculated. In fact, the permeability coefficient of soil k is used in many theoretical and practical applications. The permeability coefficient value is used to solve a variety of geotechnical issues, including slope stability, construction collapse due to ground, seepage, and leakage. The required permeability coefficient value varies greatly depending on the kind of soil and the service life of buildings. For instance, a high value of the soil permeability coefficient is necessary for filter layer and drainage construction, but roadbeds' and dams' construction requires a low value of the soil permeability coefficient. In general, many factors such as density, void (size and kind), distribution of grain size (clay content, for example), and the Atterberg limits influence on permeability coefficient values [3].

For that reason, several investigations were attempted to establish empirical correlations between the permeability coefficient and affecting variables [4–6]. A few researchers estimated soil permeability based on bulk density, grain size, and particle shape [7, 8]. Although determining the soil permeability coefficient is extremely important, but because this process is very complicated, time-consuming and expensive, the number of research articles on this topic is still limited. Therefore, there is still no exact formula to predict the permeability coefficient of soil.

Due to the rapid advancement of artificial intelligence technology in recent decades, artificial intelligence (AI) or machine learning (ML) models have become widely employed in many parts of life [9, 10] and many complex civil engineering subjects, such as structural engineering [11–13], geotechnical engineering [9, 14–18], and materials science [19, 20]. In geotechnical engineering, soft computer approaches such as fuzzy logic, artificial neural network (ANN), and support vector machines (SVM) are currently being technically utilized to artificially forecast soil compressive and shear strength, load bearing capacity of foundations, and other soil properties [16, 18, 21]. Gajurel et al. [22] proposed the k-nearest neighbors (KNN) and support vector machines (SVM) for estimating the unconfined compressive strength. Elevado et al. [23] used the KNN model to predict successfully the compressive strength of concrete mixed with fly ash and waste ceramics. Najafzadeh and Oliveto [24] developed support vector machine (SVM), multivariate adaptive regression spline (MARS), and random forest (RF) models for estimating the approach densimetric Froude number. Najafzadeh and Niazmardi [25] used the SVM model for estimating the water quality parameters. Quan Tran [20] proposed the gradient boosting (GB) model to predict the chloride diffusion coefficient. Moreover, the GB model with high performance is used to predict the unconfined compressive strength of stabilized soil treated with calcium-based additive blended [26]. Shariati et al. [27] proposed extreme learning machine (ELM) and genetic programming (GP) for designing steel-concrete composite floor systems at elevated temperatures. ELM was successfully applied to predict the compressive strength of lightweight foamed concrete [28]. Recently, Liang et al. [29] proposed the RF model and the light gradient boosting machine (LightGBM) model to predict the creep behavior of concrete. Tran and Do [30] used the light gradient boosting machine (LightGBM) model for predicting the California bearing ratio (CBR) of stabilized soil. Moreover, since 2016, Microsoft Corporation has created the light gradient boosting machine (LightGBM), an efficient gradient boosting framework implementation. Overall, numerous applications of ML models are performed in the civil engineering field. Soft computer-based models (AI or ML) have been found to be good tools for forecasting the permeability coefficient of soil. In particular, popular algorithms such as GB, RF, SVM, KNN, and LightGBM are usually used for developing ML models.

In regard to predicting the soil permeability coefficient by the ML approach, some investigations were performed in the literature. For instance, Singh et al. [31] showed that the popular ML algorithm named random forest (RF) could be used for predicting the permeability coefficient of stabilized soil containing fly ash; the performance of ML prediction was evaluated by the coefficient of determination $R^2 = 0.878$. In the investigation of Singh et al. [31], the input variables of the RF model were directly derived from the falling head permeability test [32] such as head (cm) and time of measurement (min). This ML model required the input variable derived from the measurement of permeability coefficient; therefore, this model is limited in the practical engineering applications. Other ML models by Pham et al. [33], Pham et al. [34], Bui et al. [35], and Ahmad et al. [36] are proposed to predict the permeability coefficient of soil with six input variables including water content, void ratio, specific density, liquid limit, plastic limit, and clay content, which include the easy measurable properties of soil such as the Atterberg limits (liquid limit, plastic limit, and water content) and grain size distribution (clay content). Based on 84 data samples, Pham et al. [33] revealed that the RF model could predict the permeability coefficient of soil with the performance metrics such as $R^2 = 0.724$, $\text{RMSE} = 0.840 \times 10^{-11}$ m/s, and $\text{MAE} = 0.490 \times 10^{-11}$ m/s. Using the same number of data, Pham et al. [34] improved the performance of the ML model for predicting the permeability coefficient of soil with $R^2 = 0.766$, $\text{RMSE} = 0.810 \times 10^{-11}$ m/s, and $\text{MAE} = 0.450 \times 10^{-11}$ m/s. Proposing the hybrid ML model ANN-TLBO (combination of ANN and metaheuristic algorithm named teaching learning-based optimization (TLBO)), Bui et al. [35] enhanced the ML performance for predicting the permeability coefficient of soil with $R^2 = 0.819$, $\text{RMSE} = 0.294 \times 10^{-11}$ m/s, and $\text{MAE} = 0.231 \times 10^{-11}$ m/s. Recently, Ahmad et al. [36] used the ML model Gaussian process regression (GPR) model based on the Pearson universal kernel (PUK) for predicting the permeability coefficient of soil with higher performance with $R^2 = 0.951$, $\text{RMSE} = 0.620 \times 10^{-11}$ m/s, and $\text{MAE} = 0.370 \times 10^{-11}$ m/s. Using the input variables as the easy measurable properties of soil makes the prediction of the permeability coefficient of soil by the ML model easier to access for engineers. However, the randomness was not taken into account in these ML models proposed in [33–36]. In fact, the reliability of these ML models was not verified by the validation technique such as K-fold cross-validation or Monte Carlo simulation (MCS), which should be applied to enhance the performance and reliability of the ML model [18, 37]. The K-fold cross-validation (CV) and Monte Carlo simulation (MCS) may be used to validate the predictability of ML models in order to confirm their dependability. In the Monte Carlo simulation, calculations are repeated at random while accounting for input space variability, and the associated output is then calculated using an ML model [38]. Although MCS computations take longer than K-fold CV computations, MCS findings are more dependable than K-fold CV results due to MCS's smaller variance, which is supported by the research of Fonseca-Delgado and Gomez-Gil [39]. Moreover, the effect of input variables on the predicted permeability coefficient of soil was not quantified in the investigations of [33–35].

Excepting Ahmad et al. [36], the author showed the relative importance of each input variable in the following order: water content > void ratio > liquid limit > plastic limit- > clay content > specific gravity, but the quantification of feature importance was not performed. Furthermore, the reliability and performance of the ML model influence strongly on the order or feature importance analysis. To improve the understanding of ML and AI models, SHapley Additive exPlanations (SHAP) was developed [40]. The game theory-based method SHAP calculates the Shapley value for each feature, which measures a feature's contribution to a prediction value [41]. It quantitatively displays how each characteristic influences the anticipated value and the average feature importance. Additionally, the Partial Dependence Plot (PDP) 1D method, one of the most often used ML techniques, might be helpful for calculating the influence of each feature on the forecasted value [42].

Therefore, the aim of this paper was to select a highest performance and reliable ML model to predict the permeability coefficient of soil and quantify the feature importance on the predicted value of the soil permeability coefficient with aided machine learning-based SHAP and PDP 1D. Obviously, using the metaheuristic algorithm increases the time consumption of the training process in building ML models. In order to improve the performance and reliability of ML models and simplify the training process of ML models, five popular single ML algorithms including K-nearest neighbors (KNN), support vector machine (SVM), light gradient boosting machine (LightGBM), random forest (RF), and gradient boosting (GB) are proposed in this study. The algorithms are accessible in the open-source library Sklearn [43]. The performance and reliability of ML models will be evaluated by four popular metrics: coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE) with verification by the Monte Carlo simulation (MCS) validation technique. In particular, an Excel file will be generated from the highest performance and reliable ML model to make ML models more approachable for the engineers in order to estimate the permeability coefficient of soil.

To acquire the aim of this study, a database needs to be created for training the ML algorithms proposed in this study. The database is generated from the investigation of Pham et al. [34] in which 84 data samples and 6 input variables such as easy measurable properties of soil such as water content, void ratio, the Atterberg limits (liquid limit and plastic limit), grain size distribution (clay content), and specific gravity; however, an Atterberg limit "plasticity index" is proposed to add in this database as a seventh input variable for attempting to improve the performance of ML models.

2. Database Description and Analysis

The database containing 84 samples and 6 input variables, such as density, volume content, liquid limit, clay content, void ratio, and plastic limit, is collected from the literature [34]. In order to improve the performance of ML models, a supplementary input variable "plasticity index = liquid limit–plastic limit" is added as a seventh input variable.

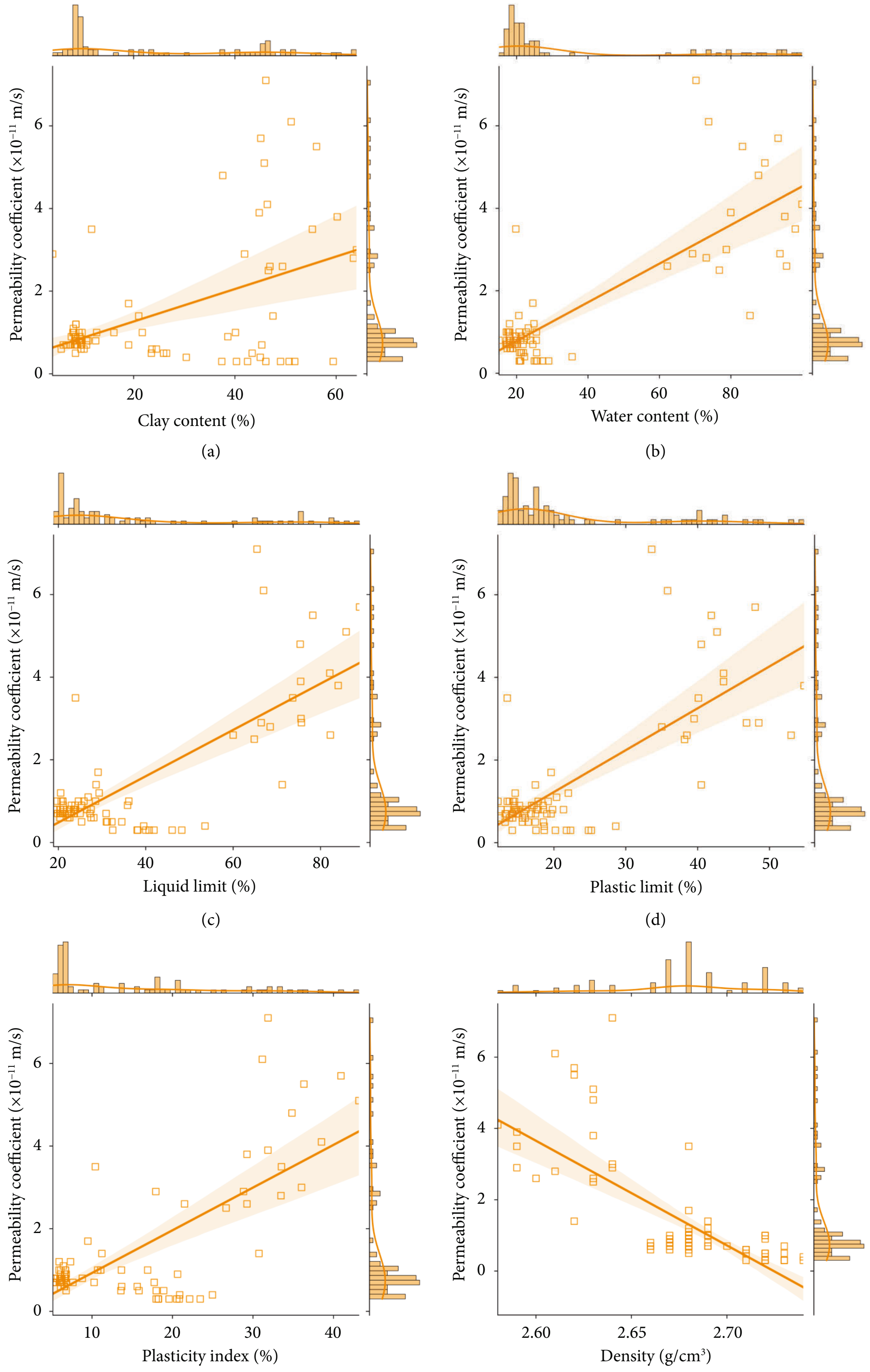
The simple relation between the input and output variables is shown in Figure 1 by a regression line that has the same direction and value as the Pearson correlation R. The gray region reflects the fitted linear regression line's 95 percent confidence range.

The Pearson correlation coefficient measures the linear relation of two variables and runs [−1.0; 1.0]. The scatter distribution figure in Figure 1 depicts linear interactions between input and output variables. This section evaluates the magnitude of the relations and distribution of data for reducing the dimension of ML models and enhancing the ML performance.

Table 1 describes the distribution value of all variables including input and output. Clay concentration ranges from 4 to 64% (mean value 24.694% and median value 12.6%). The water content fluctuates between 15.09 and 99.9% (mean value 34.228% and median value 21.135%). The liquid limit ranges between 18.9% and 88.93% (mean value 37.268% and median value 27.350%). The percentage of plastic allowed ranges from 12.2 to 54.8% (mean value 22.214% and median value 17.415%). The void ratio might range from 2.58 to 2.634% (mean value 0.968% and median value 2.634%). Density ranges from 2.58 g/cm^3 to 2.74 g/cm^3, with the content value of silica fume focusing on 2.675 g/cm^3, resulting in a median value of 2.68 g/cm^3. The permeability coefficient ranges from 0.3 10–11 m/s to 7.1 10–11 m/s, with the mean value being 1.45 10–11 m/s and the median value being 0.64 10–11 m/s, implying that the size of the coefficient is focused on the low end.

The correlations between the input and output parameters are then shown in Figure 2. The Pearson rank correlation coefficient (r_s) between each paired variable shown in Table 2 was used to create this map. In this map, the correlations between all parameters are shown simply and clearly, with various hues representing different correlation values.

Figure 2 shows the details of variable correlations, including inputs and outputs. The matrix of the Pearson correlation coefficient between input and output variables (Figure 2) demonstrates that practically all input factors have high and very strong correlations with the output variable permeability coefficient of soils in the value range from 0.48 to 0.0.83. The greatest association coefficient is between water content and permeability coefficient, and between void ratio and permeability coefficient with a correlation value of 0.83, implying that the higher the water content and void ratio, the higher the permeability coefficient. In correlation values between seven input variables, the greatest correlation value is between water content and void ratio. ML models are trained in this study using seven specified input variables. In feature selection, the Pearson correlation coefficient is a preferred approach [44]. The Pearson correlation matrix shows that the number of inputs can be considered in building ML models. As a result, seven variables could be useful for developing ML models.



(a)

(b)

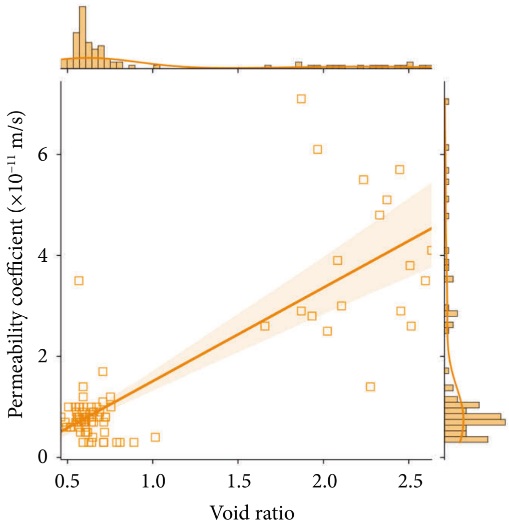
(c)

(d)

(e)

(f)

Figure 1: Continued.



(g)

Figure 1: Simplest linear correlation between permeability coefficient and each input variable.

Table 1: Distribution value of variables.

	Number of samples	Mean value	Min value	Max value	Q25%	Median	Q75%	Standard deviation	Skewness
Clay content (%)	84	24.694	4.000	64.000	8.775	12.600	44.850	18.577	0.606
Water content (%)	84	34.228	15.090	99.900	18.233	21.135	26.155	26.623	1.499
Liquid limit (%)	84	37.268	18.900	88.930	21.978	27.350	42.913	21.039	1.201
Plastic limit (%)	84	22.214	12.200	54.800	14.500	17.415	22.925	11.356	1.410
Plasticity index	84	15.054	5.140	43.160	6.408	10.335	20.853	10.657	0.968
Density (g/cm^3)	84	2.675	2.580	2.740	2.660	2.680	2.693	0.038	−0.572
Void ratio	84	0.968	0.462	2.634	0.581	0.640	0.794	0.673	1.497
Permeability coefficient × 10^{-11} (m/s)	84	1.450	0.300	7.100	0.700	0.800	1.250	1.511	1.996

3. Machine Learning Methods

3.1. Support Vector Machine Algorithm (SVM). Cortes and Vapnik [45] invented the SVM algorithm, which is a supervised artificial intelligence model. It is frequently used to evaluate data and find out what form it is in. This algorithm is widely used to anticipate and return. The SVM algorithm has a number of parameters; it is solved using optimized algorithms that pinpoint a close value to a limiting value on an experimental size. The SVM is more admirable intellectual thinking because of its benefits.

In the SVM algorithm, the estimated function is defined as follows:

$$q(x) = a\varphi(x) + c, \tag{1}$$

where $\varphi(x)$ is the higher dimensional feature space converted from the input vector x; a and c are the weights vector and a threshold, respectively, which is calculated by minimizing the following regularized risk function:

$$R(P) = P\frac{1}{m}\sum_{i=1}^{m} Z(e_i, y_i) + \frac{1}{2}\left\|a^2\right\|, \tag{2}$$

where P is the penalty parameter of the error, e_i is the desired value, m is the number of observations, $P(1/m)\sum_{i=1}^{m} Z(e_i, y_i)$ is the empirical error, and $(1/2)a^2$ is the so-called regularization term.

The detailed information and computation procedures of the SVM algorithm are presented in [45].

3.2. K-Nearest Neighbors (KNN). The suggested novel mechanism for integrating heterogeneous data into the KNN framework, which consists of two key components: regression-based weighting approaches and probability voting diagrams, is the main contribution of this approach. The weighting of each data source is determined by the regression approach, which takes into account their proportional significance to their functional and linked predictions. The voting method makes statistical inference easier by combining function class nominations from the k closest neighbors and producing a sorted list of predictions with matching confidence scores [46]. A gene can also be classified into many function classes using this method. The local regression technique performs better with one or two

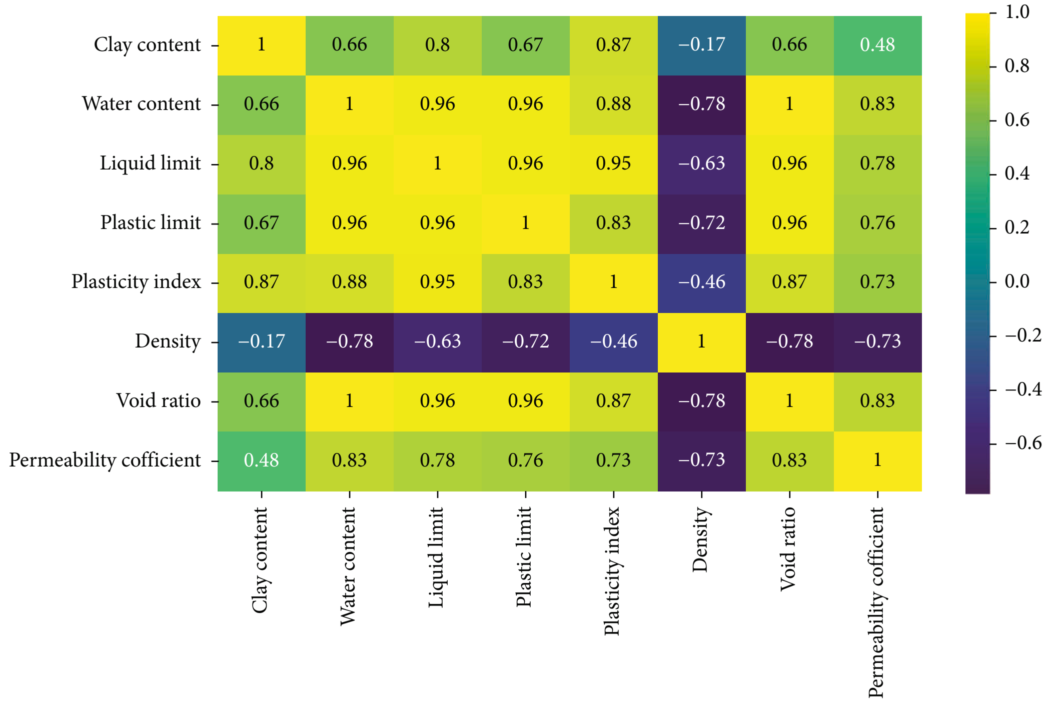


Figure 2: Pearson's correlation coefficient of input-output relations.

Table 2: Pearson's rank correlation coefficient.

r_s	Magnitude	Relation
0 ÷ 0.19	Very weak	Clay content vs density, water content vs density, liquid limit vs density, plastic limit vs density, plasticity index vs density, void ratio vs density, permeability coefficient vs density
0.2 ÷ 0.39	Weak	None
0.4 ÷ 0.59	Moderate	Permeability coefficient vs clay content
0.6 ÷ 0.79	Strong	Clay content vs water content, clay content vs plastic limit, clay content vs void ratio, liquid limit vs permeability coefficient, plastic limit vs permeability coefficient, plasticity index vs permeability coefficient, density vs void ratio, density vs permeability coefficient
0.8 ÷ 1.0	Very strong	Clay content vs liquid limit, clay content vs plasticity index, liquid limit vs water content, water content vs plastic limit, water content vs plasticity index, water content vs void ratio, water content vs permeability coefficient, liquid limit vs plastic limit, liquid limit vs plasticity index, liquid limit vs void ratio, plastic limit vs plasticity index, plastic limit vs void ratio, void ratio vs permeability coefficient

data sources, possibly due to more model flexibility, while the logistic regression method is more resilient and accurate.

The predictive power of any data source, however, may fluctuate between classes, as well as their weights. The abundance of training samples is one advantage of this present technique; nevertheless, one evident disadvantage is the lack of descriptive power. The one-layer/one-model strategy is the polar opposite. For classes with huge populations, an intermediate method is to create class-specific models. The regression model can be quite unstable when features are highly linked. To develop more robust models, principles such as component decomposition or conditioned regression techniques such as lasso regression or ridge regression might be used. The adoption of more complicated regression models is another possible expansion.

3.3. Light Gradient Boosting Machine Algorithm (LightGBM).

LightGBM is a tree-based steep frame that increases model efficiency while lowering memory use.

Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), two cutting-edge methods, are used by the LightGBM algorithm to speed up computation while preserving excellent accuracy.

3.3.1. Using a Gradient-Based Sample for LightGBM.

The various data formats play diverse roles in determining how to enhance information, with higher levels contributing more to information gathering. To maintain the accuracy of information, the SS store in cases of high slopes and merely at random in situations of short slopes. This approach can produce a more accurate advantage estimate than a single random sample, while keeping the target samples moving at the same speed, especially when the information value is high.

3.4. Random Forest (RF) and Gradient Boosting Algorithm (GB).

The random forest (RF) was created by Ho [47] and Breiman [48], and that is a strong machine learning algorithm regression problem. The practical applications of RF include bioinformatics [49], materials sciences [50], remote sensing [51], and land cover classification [9]. RF is a statistical algorithm that uses the process of extracting bootstrap samples to extract numerous samples from the original sample. When generating trees in the forest, a set of traits is used.

On the contrary, the tree number should be sufficient to ensure that all attributes are used several times. In most cases, 500 trees are used for classification and 1000 trees are used for regression.

The phases in the random forest algorithm modeling method are as follows:

Step 1: N estimator training sets were randomly selected from the original dataset (bootstrap sample). The training dataset is around two-thirds of the first dataset's dimensions. About a third of the data is considered out-of-bag since they are not involved in the tree-building process. These out-of-bag samples are used by the RF method to evaluate and quantify the attribute significance of the CART tree in the forest.

Step 2: For each bootstrap training set, a regression tree is created. A forest is created by combining the regression trees of N estimators, but these regression trees are not trimmed. Each tree's best qualities for branching are not decided throughout its growth phase. As a result, the RF technique enhances the difference between regression models by creating various training sets and boosting the combined regression model's extrapolation forecasting performance.

Step 3: For the classification problem, the value of the new sample was derived using the majority voting approach, while for the regression problem, the average of the projected values from trees was used.

The gradient boosting algorithm (GB) is a popular algorithm that has been demonstrated to be effective in a range of applications [52]. This approach uses continuous learning to create a more accurate response variable prediction, which is great for new models. The primary idea behind this technique is to create new elementary learners that are ideally correlated with the collection's negative gradient of the loss function.

3.5. Partial Dependence Plots. A partial dependence plot depicts the functional relationship between a small number of input parameters and forecasts (PDP). PDP demonstrates how the values of relevant input factors impact the predictions to some extent. The PDP also illustrates how characteristics have a little impact on the machine learning model's predicted outcomes [53].

PDP can indicate the linear relationship in a linear regression model, and the partial dependence function for regression may be constructed using the following formula:

$$\hat{q}_{x_A}(x_A) = E_{x_M}[\hat{q}(x_A, x_M)] = \int \hat{q}(x_A, x_M)\, dP(x_M). \quad (3)$$

Here, x_A represents the partial dependence function to be shown, and x_M represents the other features in the machine learning model q. A are the characteristics that the user wants to think about and forecast (only one or two features in set A). The features in set A impact the prediction outcomes that we wish to know about. The feature vectors x_A and x_M make up the whole feature space x. The function illustrates the connection between the features in set A by marginalizing the machine learning model output across the distribution of the features in set M.

The partial function q is calculated using the Monte Carlo approach by averaging all of the training data in set A so that PDP can handle multilayer issues by creating a line or plot for each layer as follows:

$$\hat{q}_{X_A}(x_A) = \frac{1}{m}\sum_{i=1}^{m}\hat{q}\left(x_A, x_M^{(i)}\right), \quad (4)$$

where $x_M^{(i)}$ is the dataset's actual feature values for the features we do not care about, and m is the dataset's total number of occurrences. PDP indicates that M's attributes are unconnected to A's. Without this assumption, the PDP average computation would contain exceedingly unlikely data points.

In addition, the PDP shows the likelihood of a certain layer in the various values for set A's attributes. PDP may therefore manage multilayer difficulties by constructing a line or plot for each layer. Because it analyzes all cases and reveals a worldwide link between a feature and the anticipated result, the PDP is a global technique.

3.6. Evaluation Criteria of the Machine Learning Model. In this study, four criteria were used, namely, correlation coefficient (R^2), RMSE (root mean square error), MAE (mean absolute error), and MAPE (mean absolute percentage error), to evaluate the accuracy of the developed model.

R^2 represents the rate of variation of the dependent variable caused by the total variation of the explanatory variables. The value of R^2 is closer to 1, and the predicted value is closer to the target value. MAE is a statistic that assesses the average amount of mistakes in a series of forecasts without taking into account their direction. The root mean square error (RMSE) is a widely used measure of the differences between values predicted by a model or estimator and the values observed. Because of its extremely obvious interpretation in terms of relative error, MAPE is often employed as a loss function for regression issues and model assessment. The MAE, RMSE, and MAPE criteria have one thing in common: they all indicate the mean model prediction error per unit of the desired output. The lower the value of RMSE, MAE, and MAPE, in contrast to higher R^2 scores, the better the model.

$$R^2 = \frac{\sum_{k=1}^{N}\left(\mathrm{val}_k^{ex} - \mathrm{val}_{\mathrm{avg}}^{ex}\right)^2 - \sum_{k=1}^{N}\left(\mathrm{val}_k^{ex} - \mathrm{val}_k^{\mathrm{pre}}\right)^2}{\sum_{k=1}^{N}\left(\mathrm{val}_k^{ex} - \mathrm{val}_{\mathrm{avg}}\right)^2},$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\mathrm{val}_k^{ex} - \mathrm{val}_k^{\mathrm{pre}}\right)^2}, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{k=1}^{N}\left|\mathrm{val}_k^{\mathrm{pre}} - \mathrm{val}_k^{ex}\right|,$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|\mathrm{val}_k^{ex} - \mathrm{val}_k^{\mathrm{pre}}\right|}{\mathrm{val}_k^{ex}} \times 100\%,$$

where N is the number of datasets, $\mathrm{val}^{\mathrm{ex}}$ and $\mathrm{val}_{\mathrm{avg}}^{ex}$ are the experimental value and mean experimental value, respectively, and $\mathrm{val}^{\mathrm{pre}}$ is the predicted value by the ML model.

4. Methodology Flowchart

The current study's approach flowchart for the permeability coefficient of soil contains three basic steps:

Step 1: Make a database.

As a first step, the dataset was compiled from 84 easily accessible literatures published in recognized journals. The database has seven input variables and one output variable. All data are split into two halves at random: a training dataset and a testing dataset, with 70% of the data (59 examples) being used to train the models and 30% being used to test the models.

Step 2: Create training models and select the suitable ML models.

The machine learning (ML) models were trained using training datasets and methodologies in this step, with the five single algorithms available in the Sklearn library of Python language programming [43], such as K-nearest Neighbors (KNN), support vector machine (SVM), light gradient boosting machine (LightGBM), random forest (RF), and gradient boosting (GB). The performance of five single ML models is evaluated with aided four metric criteria including R^2, RMSE, MAE, and MAPE. Based on the ML performance, the suitable ML models will be selected for the next step.

Step 3: Prediction of permeability coefficient and extensive sensitivity analyses are performed with the aided suitable ML models.

Machine Learning-based approaches such as GB- and RF-based SHapley Additive exPlanations and GB-based Partial Dependence Plot (PDP) 1D are used to comprehend the impact of each input variables on the permeability coefficient of soil.

A method schematic diagram is shown in Figure 3.

5. Results and Discussion

5.1. Evaluating Performance of ML Models. The performance of five single machine learning models is compared in this section. The Monte Carlo simulation is used to evaluate the performance and reliability of each ML model. Each ML model is subjected to a total of 10000 runs. The performance and reliability comparisons are carried out throughout four evaluation criteria R^2 value, RMSE value ($\times 10^{-11}$ m/s), MAE value ($\times 10^{-11}$ m/s), and MAPE (%), in Figures 4(a)–4(d), respectively.

It is worth noting that with the exception of the LightGBM model, the performance of the four ML models is an excellent fit for both training and testing datasets.

The LightGBM model has the lowest performance when compared to other variants. In both the training and testing datasets, the coefficient of determination R^2 is less than 0.6. Meanwhile, the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) values produced by the LightGBM model are larger than those obtained by other models. This approach appears to be ineffective for estimating the permeability coefficient of soils.

Table 3 presents the results of the models based on the four evaluation criteria mentioned above.

For the evaluation criterion of the coefficient of determinations R^2, the GB model gives the best results. The results obtained from training are asymptotically close to the value 1. According to the obtained results of the R^2 evaluation criteria in the training, the performance of ML models can be arranged in the sequence: LightGBM < SVM < KNN < RF < GB. In the testing, the performance of ML models can be ordered as follows: LightGBM < KNN < SVM < RF < GB.

For the two evaluation criteria RMSE and RMSE, the GB model also gives the best results compared with remaining models. According to the results obtained by the RMSE evaluation criteria in the training, the performance of ML models can be arranged in the sequence: LightGBM < SVM < KNN < RF < GB. In the testing, the performance of ML models can be ordered as follows: LightGBM < KNN < SVM < RF < GB. The performance of ML models can be set in the following order for the MAE assessment criteria in the training: LightGBM < KNN < SVM < RF < GB. The performance ML models in the testing may be ordered as follows: LightGBM < KNN < RF < SVM < GB.

In comparison with the other models in testing, the GB model produced the best results for the MAPE assessment criterion. In training, the GB model produces the best results compared with other models, but the results in testing are not as good. The performance of ML models may be organized in the following order for the MAPE assessment criteria in training: LightGBM < SVM = KNN < RF < GB. The performance of ML models in the testing may be put in the following order: LightGBM < RF < GB < KNN < SVM.

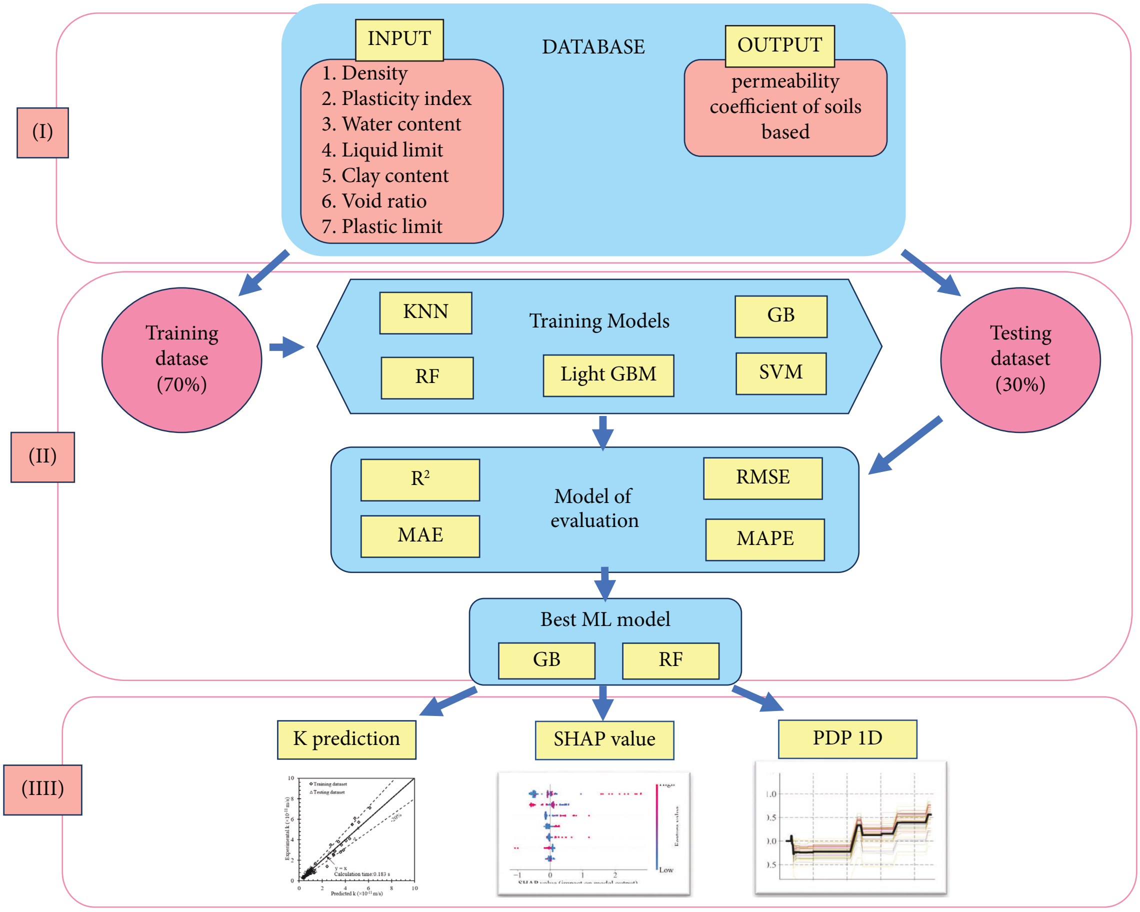


Figure 3: Methodology of investigation.

Table 4 shows the best performance of the ML model according to the highest value R^2 for the testing dataset. The results show that among the models, RF and GB models have the highest performance for predicting the permeability coefficient of soils, in which the GB model is the most dominant model in both training and testing datasets. The two ML models GB and RF are used to predict the permeability coefficient of soils and investigate the feature importance-sensitivity analyses in the next sections.

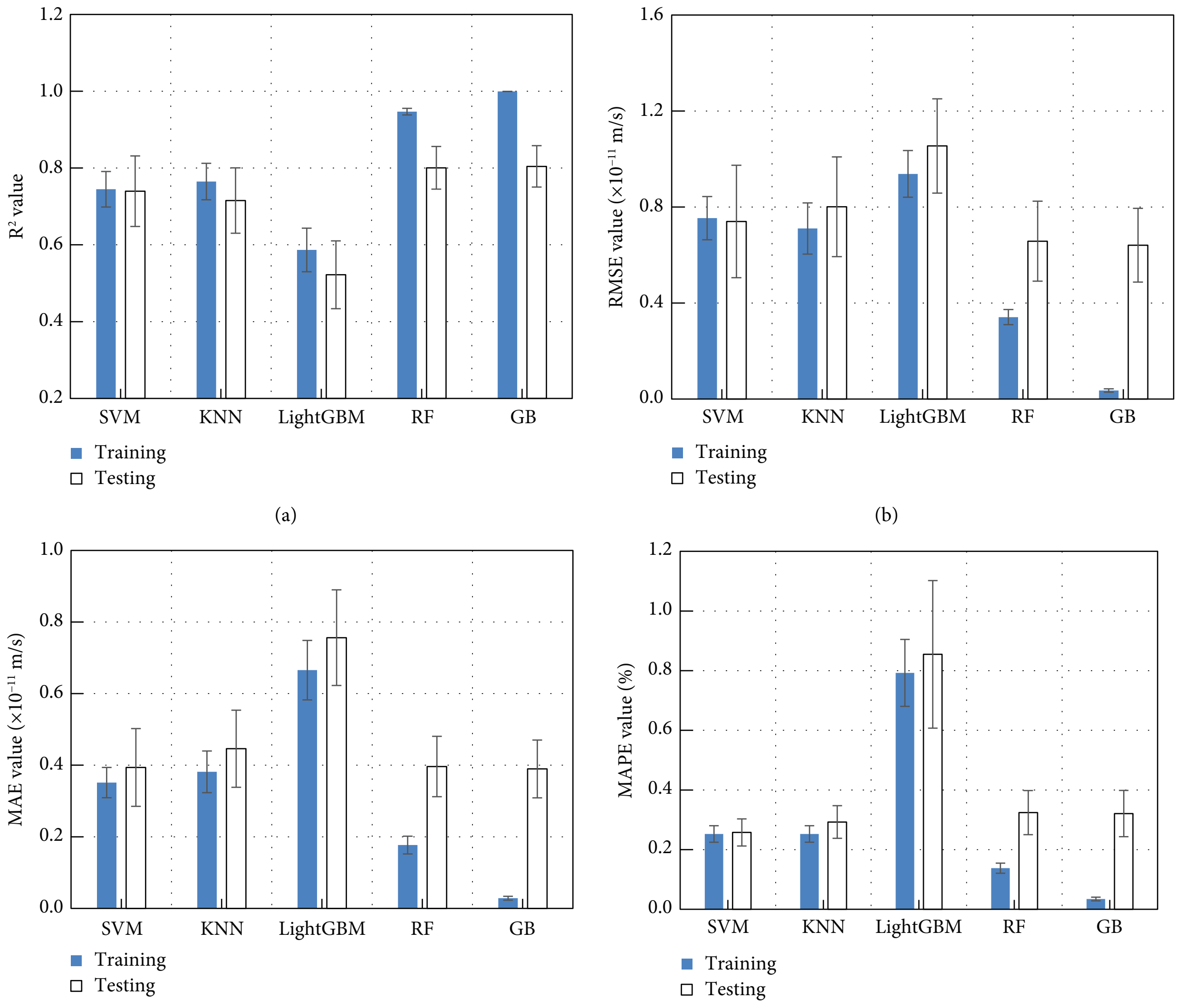
5.2. Predicting Permeability Coefficient by ML Models. Figure 5 shows a correlation graph between experimental permeability coefficient and permeability coefficient predicted by the (a) RF model and (b) GB model. For the RF model, the precision of both training and testing data results is lower than that of the GB model with the metric criteria R^2, RMSE, MAE, and MAPE to be equal to 0.960, 0.236, 0.196, and 0.238, and 0.971, 0.199, 0.161, and 0.185 for the testing dataset of the RF and GB model, respectively. In addition, the calculation time of the RF model is 0.183 s larger than that of the GB model (0.041 s). Using error line −20% and +20%, all correlation points including training and testing datasets in case of GB-based prediction are limited in a zone created by two lines; these correlation points approach very closely the perfect line $y = x$, where some correlation points including training and testing datasets in case of RF-based prediction exceeded the zone created by two lines. Therefore, the GB model is more suitable and gives the most accurate prediction where the four evaluation metrics of the GB model are better than those of the RF model.

Furthermore, according to the investigation of Najafzadeh et al. [54, 55] and the investigation of Saberi-Movahed et al. [56], scatter index (SI) and BIAS are the important metrics in evaluating the performance of ML models. Therefore, scatter index (SI) and BIAS are also used for evaluating the performance of GB and RF models. The lower the value of SI and BIAS, the higher the ML performance. The formula of scatter index (SI) and BIAS is described as follows:

$$\mathrm{SI} = \frac{\sqrt{(1/N)\sum_{k=1}^{N}\left(\left(\mathrm{val}_k^{\mathrm{pre}} - \mathrm{val}_{\mathrm{avg}}^{\mathrm{pre}}\right) - \left(\mathrm{val}_k^{\mathrm{ex}} - \mathrm{val}_{\mathrm{avg}}^{\mathrm{ex}}\right)\right)^2}}{(1/N)\sum_{k=1}^{N} val_k^{ex}}, \qquad (6)$$

$$\mathrm{BIAS} = \frac{\sum_{k=1}^{N}\left(\mathrm{val}_k^{\mathrm{pre}} - \mathrm{val}_k^{\mathrm{ex}}\right)}{N}.$$

Table 5 summarizes the values of SI and BIAS for the training dataset and the testing dataset of the GB and RF models. The value of SI and BIAS of the GB model is significantly lower than that of the RF model in two cases of



(a)

(b)

(c)

(d)

FIGURE 4: Performance evaluation of 5 machine learning models after 10000 simulations: (a) R^2, (b) RMSE, (c) MAE, and (d) MAPE.

dataset including the training dataset and the testing dataset, which confirms the highest performance of a single GB model compared with that of other ML models in this study.

Moreover, the uncertainty and reliability analyses are performed for two specific ML models including the RF and GB models. According to Saberi-Movahed et al. [56], restricting the predicted range in which the real value of an experiment's outcome lies is the main objective of the uncertainty analysis. The uncertainty interval is an interval that represents this estimated range. Based on the computed errors for the measurement procedure of the experiment under consideration, it may be approximated. U95 is one of the methods used in the uncertainty analysis to calculate the uncertainty interval. If you execute the given experiment again, you will find that 95 times out of every 100 trials, the real value of the experiment's outcome will fall inside the provided uncertainty interval. This is what the value of U95 associated with a specific experiment result means. U95 can be described as follows:

$$\mathrm{SI} = \frac{\sqrt{(1/N)\sum_{k=1}^{N}\left(\left(\mathrm{val}_k^{\mathrm{pre}} - \mathrm{val}_{\mathrm{avg}}^{\mathrm{pre}}\right) - \left(\mathrm{val}_k^{\mathrm{ex}} - \mathrm{val}_{\mathrm{avg}}^{\mathrm{ex}}\right)\right)^2}}{(1/N)\sum_{k=1}^{N}\mathrm{val}_k^{ex}},$$

$$\mathrm{BIAS} = \frac{\sum_{k=1}^{N}\left(\mathrm{val}_k^{\mathrm{pre}} - \mathrm{val}_k^{ex}\right)}{N}. \tag{7}$$

According to Saberi-Movahed et al. [56], the reliability analysis's metric is described as follows:

$$\mathrm{Reliability} = \left(\frac{100\%}{N}\right)\sum_{k=1}^{N} u_i. \tag{8}$$

There are two stages required to attain u_i. An initial definition of the relative average error (RAE) is a vector with a k^{th} component, which is expressed as follows:

$$\mathrm{RAE}_k = \left|\frac{\mathrm{val}_k^{ex} - \mathrm{val}_k^{\mathrm{epre}}}{\mathrm{val}_k^{ex}}\right|, \tag{9}$$

Table 3: Min, max, average, and StD of performance values of 10000 performance values.

		R^2				RMSE ($\times 10^{-11}$ m/s)			
		StD	Min	Average	Max	StD	Min	Average	Max
Training	SVM	0.046	0.541	0.745	0.884	0.090	0.428	0.754	0.943
	KNN	0.047	0.613	0.765	0.935	0.107	0.252	0.711	0.927
	LightGBM	0.057	0.380	0.587	0.784	0.097	0.530	0.938	1.192
	RF	0.009	0.884	0.947	0.972	0.031	0.237	0.342	0.438
	GB	0.000	0.998	0.999	1.000	0.007	0.019	0.036	0.067
Testing	SVM	0.092	0.597	0.740	0.960	0.234	0.148	0.740	1.333
	KNN	0.085	0.549	0.715	0.961	0.208	0.173	0.801	1.403
	LightGBM	0.088	0.342	0.522	0.819	0.196	0.460	1.054	1.749
	RF	0.055	0.721	0.801	0.960	0.167	0.209	0.658	1.036
	GB	*0.054*	0.724	*0.804*	*0.971*	*0.153*	0.199	*0.641*	1.013
			MAE ($\times 10^{-11}$ m/s)				MAPE (%)		
Training	SVM	0.042	0.214	0.351	0.457	0.028	0.156	0.252	0.356
	KNN	0.058	0.180	0.381	0.526	0.028	0.156	0.252	0.356
	LightGBM	0.083	0.346	0.666	0.931	0.112	0.411	0.793	1.191
	RF	0.025	0.087	0.177	0.243	0.017	0.087	0.138	0.196
	GB	0.005	0.014	0.028	0.051	0.006	0.017	0.035	0.061
Testing	SVM	0.109	0.118	0.394	0.761	*0.045*	0.147	*0.258*	0.703
	KNN	0.108	0.137	0.446	0.876	0.055	0.155	0.293	0.602
	LightGBM	0.134	0.309	0.756	1.267	0.247	0.297	0.855	1.939
	RF	0.084	0.163	0.396	0.680	0.074	0.167	0.324	0.784
	GB	*0.081*	0.161	*0.390*	0.656	0.077	0.157	0.321	0.830

Table 4: Best performance of the ML model according to highest R^2 for the testing dataset.

ML model	Training dataset				Testing dataset			
	R^2	RMSE ($\times 10^{-11}$ m/s)	MAE ($\times 10^{-11}$ m/s)	MAPE	R^2	RMSE ($\times 10^{-11}$ m/s)	MAE ($\times 10^{-11}$ m/s)	MAPE
SVM	0.725	0.867	0.436	0.244	0.819	0.199	0.181	0.282
KNN	0.688	0.889	0.504	0.300	0.960	0.248	0.179	0.211
LightGBM	0.573	1.068	0.779	0.860	0.960	0.484	0.362	0.639
RF	*0.948*	*0.367*	*0.215*	*0.147*	*0.961*	*0.236*	*0.196*	*0.238*
GB	*0.999*	*0.047*	*0.037*	*0.043*	*0.971*	*0.199*	*0.161*	*0.185*

where delta is the threshold value of a permeability coefficient parameter, $u_i = 1$ if $\mathrm{RAE}_i <$ delta and $u_i = 0$. According to Chinese standards, the ideal value is 20%. Therefore, Table 6 summarizes the value of U95 and reliability of the GB and RF models.

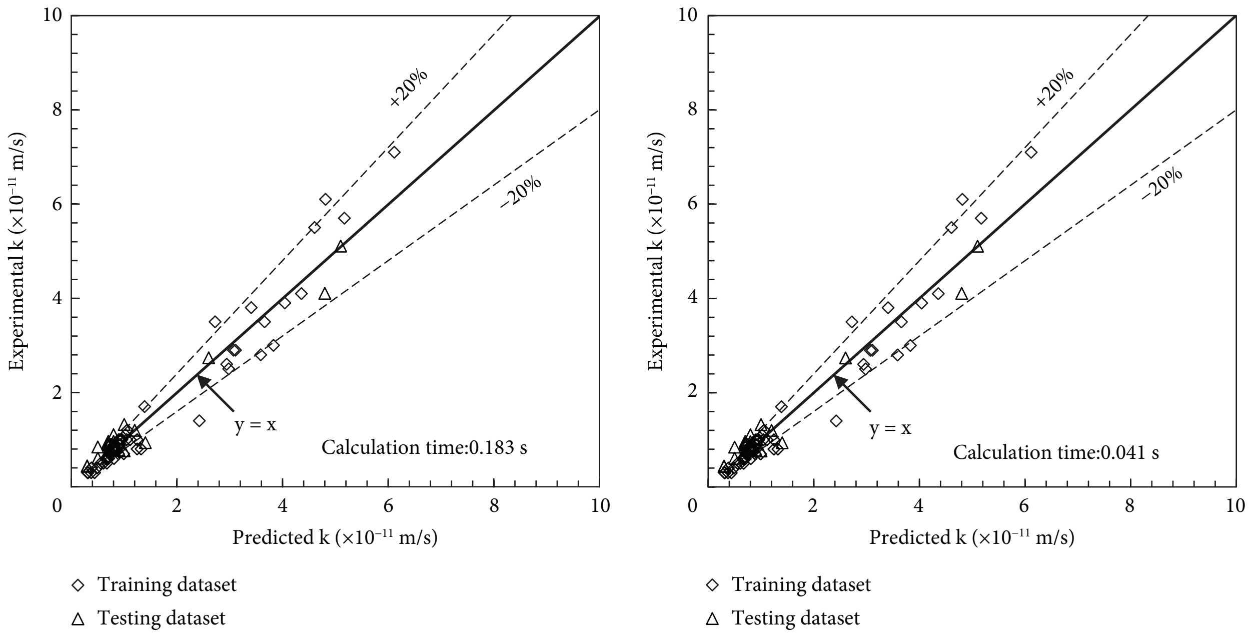
The results in Table 6 show the higher reliability and lower uncertainty of the GB model compared with those of the RF model where the U95 value and reliability value of the GB model for the testing dataset are, respectively, equal to 0.4569 and 61.5385% versus 0.4595 and 46.1538% for the testing dataset of the RF model. Overall, the single GB model of this study has highest reliability compared with other ML models for predicting the permeability coefficient of soil.

Table 7 shows the comparison between the predicted permeability coefficient and the experimental permeability coefficient. Pham et al. [33], Pham et al. [34], Bui et al. [35], and Ahmad et al. [36] used the same number of 84 data samples for this investigation; however, the performance of the single ML model "gradient boosting" is significantly improved with $R^2 = 0.971$, RMSE = 0.199, MAE = 0.161, and MAPE = 0.185 for the testing dataset. With the ML model's sharply improved performance and reliability with R^2 (mean value of 10000 runs) = 0.804 for the testing dataset, the highest performance and the reliability of the GB model seem to come from the supplementary input variable "plasticity index" proposed in this study. The input variable effect on the predicted permeability coefficient is deeply investigated in the next section.

The single ML model gradient boosting appears to be an outstanding soft tool for forecasting soil permeability. An Excel file created from the highest performance gradient boosting is supplied to forecast the soil permeability coefficient for enhancing the use of ML models in engineering applications (please find the Excel file at this link: https://drive.google.com/file/d/1jjNs6qwR_BuFhCjjSzxt56DlNA9FplLl/view?usp=sharing).

5.3. *Investigation of Input Variable Effect on Permeability Coefficient.*

Figure 6 shows the feature importance analysis of the predicted value of the soil permeability coefficient including (a) RF model-based permutation and (c) GB model-based permutation analysis; and (b) RF model-based and (d) GB model-based SHAP value.



(a)

(b)

Figure 5: Comparison between the experimental and the predicted permeability coefficient of soil in two cases: (a) RF model and (b) GB model.

Table 5: SI and BIAS values of the GB and RF models.

Model	Data	SI	BIAS
GB	Training dataset	0.0299	0.0000
	Testing dataset	*0.1714*	*0.0192*
RF	Training dataset	0.2391	0.0346
	Testing dataset	*0.2000*	*0.0615*

Table 6: U95 and reliability values of the GB and RF models.

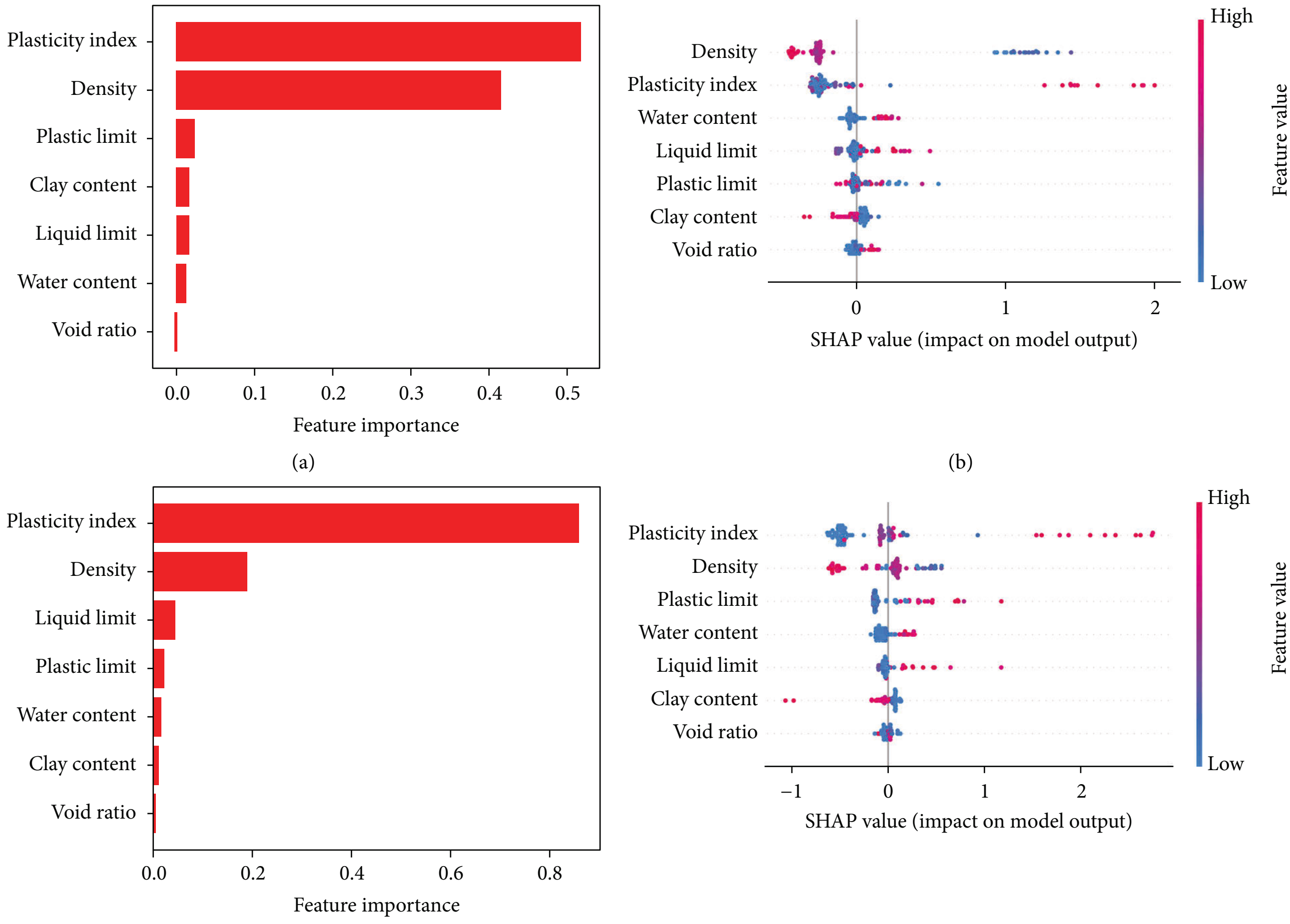
Model	Data	U95	Reliability (%)
GB	Training dataset	0.4149	98.2759
	Testing dataset	*0.4569*	*61.5385*
RF	Training dataset	0.4260	72.4138
	Testing dataset	*0.4598*	*46.1538*

Inspiring from sensitivity analysis of investigation performed by Najafzadeh and Saberi-Movahed [55], the feature importance analysis based on permutation importance is carried out in this section to reveal the most important feature for predicting the permeability coefficient of soil by the single GB model. The permutation analyses based on the Sklearn library of Python language programming [43] (cf. Figures 6(a) and 6(c)) show that the plasticity index is the most important input variable on the predicted value of permeability coefficient. Ranked 2nd in the importance is the input variable "density." According to the feature importance value, the sum of the relative importance value for two most importance feature "plasticity index and density" is equal to about 0.9 and 1.0 for the RF model and GB model, respectively. The other features, such as liquid limit, plastic limit, water content, clay content, and void ratio, have an insignificant effect on the predicted value of permeability coefficient. The results are cross-checked simultaneously by the GB model and the RF model. The obtained results (water content > void ratio > liquid limit > plastic limit > clay content > specific gravity) are different from those of the feature importance investigated by Ahmad et al. [36].

The SHapley Additive exPlanations (SHAP) is based on the two best performance metrics of ML models such as RF (cf. Figure 6(b)) and GB (cf. Figure 6(d)) to identify and quantify each of the input variables such as density, plasticity index, water content, liquid limit, plastic limit, clay content, and void ratio on the permeability coefficient of soil.

Table 7: Comparison between ML models in previous studies and the GB model proposed in this study for predicting the permeability coefficient of soil.

Reference	Best machine learning algorithm	Number of inputs	Dataset size	Best performance evaluation for testing part
Pham et al. [33]	RF	6 inputs: water content, void ratio, specific density, liquid limit, plastic limit, and clay content	84	$R^2 = 0.724$ $\text{RMSE} = 0.840 \times 10^{-11}$ m/s $\text{MAE} = 0.490 \times 10^{-11}$ m/s
Pham et al. [34]	M5P	6 inputs: water content, void ratio, specific density, liquid limit, plastic limit, and clay content	84	$R^2 = 0.766$ $\text{RMSE} = 0.810 \times 10^{-11}$ m/s $\text{MAE} = 0.450 \times 10^{-11}$ m/s
Bui et al. [35]	TLBO-ANN	6 inputs: water content, void ratio, specific density, liquid limit, plastic limit, and clay content	84	$R^2 = 0.819$ $\text{RMSE} = 0.294 \times 10^{-11}$ m/s $\text{MAE} = 0.231 \times 10^{-11}$ m/s
Ahmad et al. [36]	GPR-PUK	6 inputs: water content, void ratio, specific density, liquid limit, plastic limit, and clay content	84	$R^2 = 0.951$ $\text{RMSE} = 0.620 \times 10^{-11}$ m/s $\text{MAE} = 0.370 \times 10^{-11}$ m/s
This study	GB	7 inputs: water content, void ratio, specific density, liquid limit, plastic limit, and clay content + plasticity index	84	R^2 (mean value of 10000 runs) = *0.804* $R^2 = 0.971$ $\text{RMSE} = 0.199 \times 10^{-11}$ m/s $\text{MAE} = 0.161 \times 10^{-11}$ m/s $\text{MAPE} = 0.185$



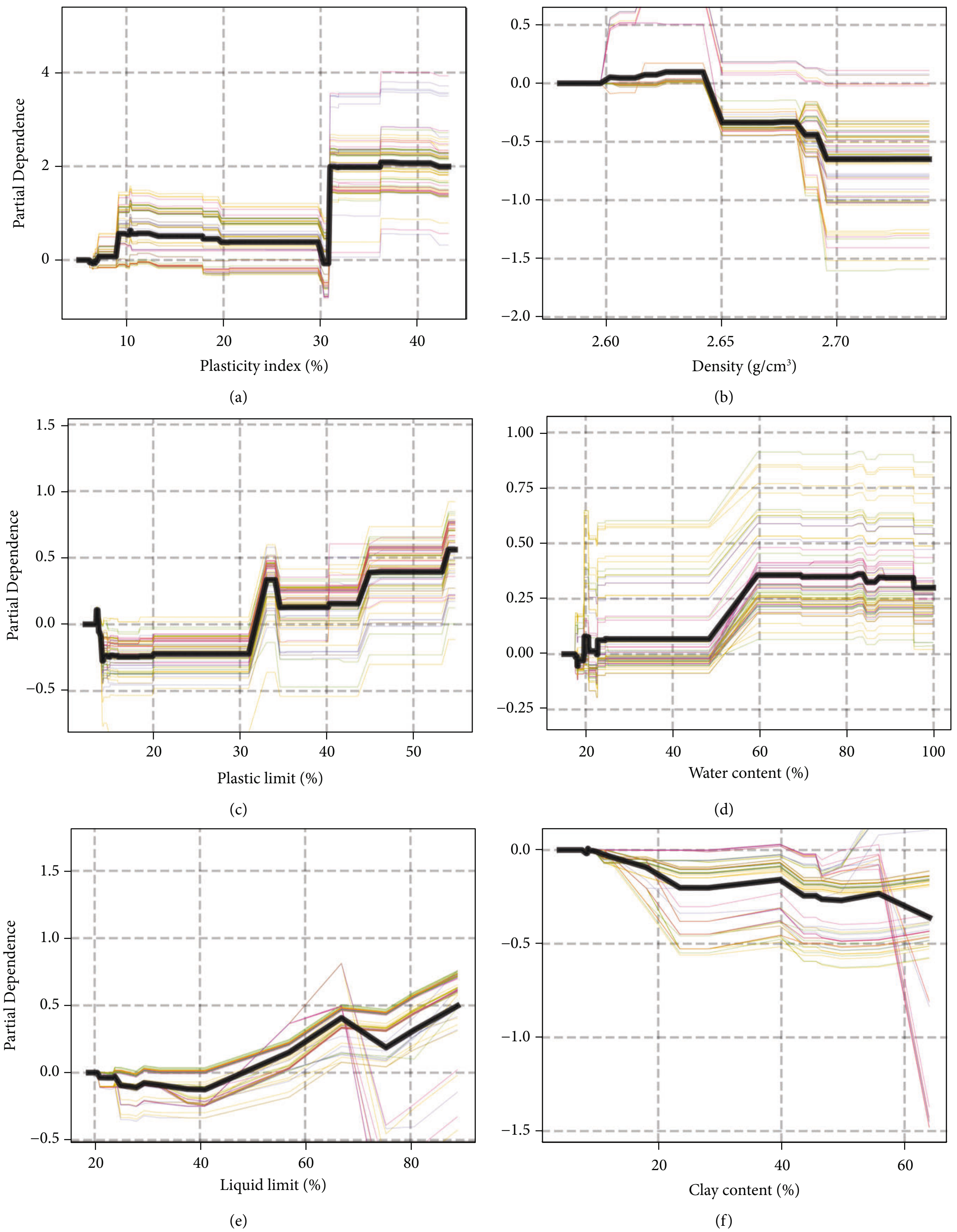
(a)

(b)

(c)

(d)

Figure 6: Feature importance analysis with (a) RF model-based permutation and (b) RF model-based and (c) GB model-based permutation; and global interpretation SHAP value with (d) GB model-based influence of input variable on the permeability coefficient of soil.



(a)

(b)

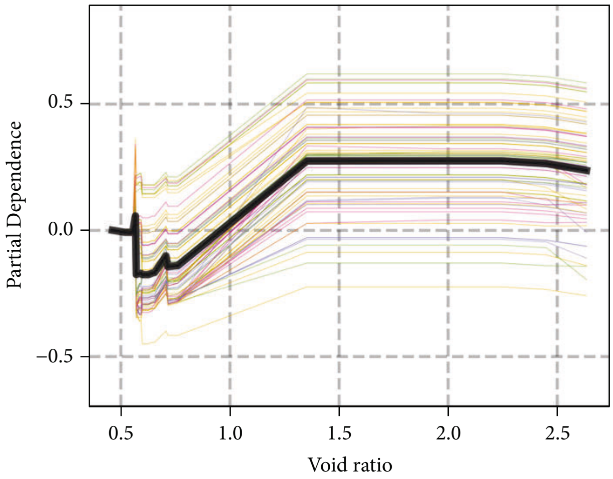
(c)

(d)

(e)

(f)

Figure 7: Continued.



(g)

Figure 7: Individual conditional expectation of each feature on the permeability coefficient of soil.

SHAP values based on the RF model clearly show that density, plastic limit, and clay content negatively affect soil permeability. When these inputs tend to increase, the permeability coefficient of soil will decrease. For the remaining factors, there will be a positive trend with the permeability of soil; when these factors increase, the permeability coefficient of soil will also increase. Density is the most essential input variable in the RF model, followed by plasticity index, water content, and liquid limit. The plasticity index, however, is the most essential input in the GB model, followed by density, plastic limit, and water content.

SHAP values based on the GB model clearly show that density, clay content, and void ratio negatively affect soil permeability. When these inputs tend to increase, the permeability coefficient of soil will decrease. For the remaining factors, there will be a positive trend with the permeability of soil; when these factors increase, the permeability coefficient of soil will also increase.

Figure 6 shows that the two most essential inputs are density and plasticity index, which have been cross-checked by the GB and RF machine learning models. Furthermore, the plasticity index was used in this study; the performance of the machine learning model is better than that of the model in the previous investigations (cf. Table 5). The density of soil can be considered as one of the most important parameters on the prediction of permeability coefficient because the higher value of density implies that the distribution of soil particles is denser, and the pore of soil is decreased. That induces the decreasing value of permeability coefficient. Overall, based on GB-RF, the feature effect on the permeability coefficient of soil can be sorted in order: (plasticity index, density) > (water content, liquid limit, plastic limit) > clay content > void ratio. Therefore, the plasticity index and density of soil are the first priority soil properties to measure when assessing the permeability coefficient of soil.

Figure 7 shows individual conditional expectation (ICE) of each feature on the coefficient of permeability. In the range from 8 to 30% (cf. Figure 7(a)), the plasticity index is directly proportional to the water permeability of soil. However, from the range of 32% onwards, in terms of average amplitude, there is the clearest jump in soil water permeability up to 2 to 3 times of the previous range and a wide range of fluctuations. With plasticity index varying from 30% to 40%, the magnitude value of permeability coefficient increases about from 0.2 to 2.0×10^{-11} (m/s). The input parameter density (cf. Figure 7(b)), although it is proportional to the soil's water permeability, is only in a very small, insignificant range. In general, it can be estimated that this parameter is inversely proportional to the permeability coefficient of soil. For the plastic limit parameter (cf. Figure 7(c)), the oscillation amplitude fluctuates unpredictably, sometimes being inverse with the permeability coefficient. However, in general, it can be seen that the main trend is proportional. In the range from 50% to 60%, the water content (cf. Figure 7(d)) that is directly proportional to the water permeability of soil is most pronounced. Liquid limit is inversely proportional to soil water permeability in the range from 20% to 42% and from 67% to 75% (cf. Figure 7(e)). However, the average amplitude of oscillation tends to go up (proportional) over a wider range, so overall, this parameter is directly proportional to the soil's water permeability. For clay content, the dominant trend is inversely proportional to soil water permeability (cf. Figure 7(f)). In the range from 0.75 to 1.3 (cf. Figure 7(g)), the void ratio that is directly proportional to the water permeability of soil is the most obvious. Overall, excepting density and clay content, the higher the value of the plasticity index, plastic limit, water content, liquid limit, and void ratio is, the higher the permeability coefficient of soil is.

6. Conclusions and Perspectives

In order to select the single machine learning (ML) model with high performance and reliability for predicting and investigating the permeability coefficient of soil, five single machine learning algorithms including support vector machine (SVM), K-nearest neighbors (KNN), light gradient boosting machine (LightGBM), random forest (RF), and gradient boosting (GB) are introduced in this study. Seven input variables such as plasticity index, plastic limit, liquid limit, clay content, density, water content, and void ratio are used to build ML models. R^2, RMSE, MAE, MAPE, SI, and BIAS are six measures used to assess the performance of machine learning models. 10000 Monte Carlo simulation runs are used to carefully verify ML model performance and reliability. The uncertainty and reliability analyses are also performed. The results of these analyses demonstrate that the single GB model is the most effective including highest performance and highest reliability ML model for predicting the soil permeability coefficient with the best performance metrics such as $\text{RMSE} = 0.971$, $\text{RMSE} = 0.199 \times 10^{-11}$ m/s, $\text{MAE} = 0.161 \times 10^{-11}$ m/s, $\text{MAPE} = 0.185$, $\text{SI} = 0.1714$, and $\text{BIAS} = 0.0192$ for the testing dataset.

Machine learning-based approaches such as GB- and RF-based SHAP and GB-based PDP 1D are used to comprehend the impact of input variable on the permeability coefficient of soil in extensive sensitivity studies. According to the results, the order influence on the value of the permeability coefficient may be ordered as plasticity index, density > water content, liquid limit, plastic limit > clay content > void ratio. The plasticity index and density of soil are the first priority soil properties to measure when assessing the permeability coefficient of soil.

Engineers can determine the soil permeability coefficient from the 7 variables with the aided Excel file generated by the GB model in real situations. Alternative machine learning techniques and a greater data number with out-of-range values may be used in the future to undertake a more in-depth examination.

Data Availability

The processed data are available from the corresponding author upon request.

Additional Points

(i) To estimate the soil permeability coefficient, five single machine learning algorithms with default hyperparameters are provided. (ii) The gradient boosting algorithm is the most suitable machine learning algorithm for predicting the permeability coefficient of soil. (iii) Factors' influence on the permeability coefficient of soil is quantified by SHapley Additive exPlanations (SHAP) and Partial Dependence Plot 1D. (iv) Plasticity index and density of soil are the first priority soil properties to measure when predicting the permeability coefficient of soil by the machine learning model.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] A. F. Elhakim, "Estimation of soil permeability," *Alexandria Engineering Journal*, vol. 55, no. 3, pp. 2631–2638, 2016.

[2] J. K. Mitchell, D. R. Hooper, and R. G. Campenella, "Permeability of compacted clay," *Journal of the Soil Mechanics and Foundations Division*, vol. 91, no. 4, pp. 41–65, 1965.

[3] B. Zhou and N. Lu, "Correlation between Atterberg limits and soil adsorptive water," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 147, no. 2, Article ID 04020162., 2021.

[4] I. Garcia-Bengochea, A. G. Altschaeffl, and C. W. Lovell, "Pore distribution and permeability of silty clays," *Journal of the Geotechnical Engineering Division*, vol. 105, no. 7, pp. 839–856, 1979.

[5] H. Beushausen, R. Torrent, and M. G. Alexander, "Performance-based approaches for concrete durability: state of the art and future research needs," *Cement and Concrete Research*, vol. 119, pp. 11–20, 2019.

[6] R. E. Olson, "Effective stress theory of soil compaction," *Journal of the Soil Mechanics and Foundations Division*, vol. 89, no. 2, pp. 27–45, 1963.

[7] J. M. Sperry and J. J. Peirce, "A model for estimating the hydraulic conductivity of granular material based on grain shape, grain size, and porosity," *Ground Water*, vol. 33, no. 6, pp. 892–898, 1995.

[8] I. Lebron, M. G. Schaap, and D. L. Suarez, "Saturated hydraulic conductivity prediction from microscopic pore geometry measurements and neural network analysis," *Water Resources Research*, vol. 35, no. 10, pp. 3149–3158, 1999.

[9] T. A. Pham, V. Q. Tran, H.-L. T. Vu, and H.-B. Ly, "Design deep neural network architecture using a genetic algorithm for estimation of pile bearing capacity," *PLoS One*, vol. 15, no. 12, Article ID e0243030, 2020.

[10] Q. H. Nguyen, H.-B. Ly, T.-A. Nguyen, V.-H. Phan, L. K. Nguyen, and V. Q. Tran, "Investigation of ANN architecture for predicting shear strength of fiber reinforcement bars concrete beams," *PLoS One*, vol. 16, no. 4, Article ID e0247391, 2021.

[11] Q. H. Nguyen, H.-B. Ly, V. Q. Tran et al., "A novel hybrid model based on a feedforward neural network and one step secant algorithm for prediction of load-bearing capacity of rectangular concrete-filled steel tube columns," *Molecules*, vol. 25, no. 15, p. 3486, 2020.

[12] H. Q. Nguyen, H.-B. Ly, V. Q. Tran, T.-A. Nguyen, T.-T. Le, and B. T. Pham, "Optimization of artificial intelligence system by evolutionary algorithm for prediction of axial capacity of rectangular concrete filled steel tubes under compression," *Materials*, vol. 13, no. 5, p. 1205, 2020.

[13] V. Q. Tran and T. A. Pham, *Applied Sciences | Free Full-Text | Prediction of Pile Axial Bearing Capacity Using Artificial Neural Network and Random Forest*, 2022.

[14] V. Quan and H. Q. Do, "Prediction of California bearing ratio (CBR) of stabilized expansive soils with agricultural and industrial waste using Light gradient boosting machine," *J. Sci. Transp. Technol.*, pp. 1–9, 2021.

[15] V. Q. Tran and L. Q. Nguyen, "Using machine learning technique for designing reinforced lightweight soil," *Journal of Intelligent and Fuzzy Systems*, vol. 43, no. 1, pp. 1633–1650, 2022.

[16] V. Q. Tran, "Compressive strength prediction of stabilized dredged sediments using artificial neural network," *Advances in Civil Engineering*, vol. 2021, pp. 1–8, 2021.

[17] T. A. Pham and V. Q. Tran, "Developing random forest hybridization models for estimating the axial bearing capacity of pile," *PLoS One*, vol. 17, no. 3, Article ID e0265747, 2022.

[18] V. Q. Tran, "Hybrid gradient boosting with meta-heuristic algorithms prediction of unconfined compressive strength of stabilized soil based on initial soil properties, mix design and effective compaction," *Journal of Cleaner Production*, vol. 355, Article ID 131683, 2022.

[19] V. Quan Tran, V. Quoc Dang, and L. Si Ho, "Evaluating compressive strength of concrete made with recycled concrete aggregates using machine learning approach," *Construction and Building Materials*, vol. 323, Article ID 126578, 2022.

[20] V. Quan Tran, "Machine learning approach for investigating chloride diffusion coefficient of concrete containing supplementary cementitious materials," *Construction and Building Materials*, vol. 328, Article ID 127103, 2022.

[21] M. Marjanović, M. Kovačević, B. Bajat, and V. Voženílek, "Landslide susceptibility assessment using SVM machine learning algorithm," *Engineering Geology*, vol. 123, no. 3, pp. 225–234, 2011.

[22] A. Gajurel, P. S. Mukherjee, and B. Chittoori, *Estimating Optimal Additive Content for Soil Stabilization Using Machine Learning Methods*, pp. 662–672, 2019.

[23] K. J. T. Elevado, J. Galupino, and R. Gallardo, "Compressive strength modelling of concrete mixed with fly ash and waste ceramics using K-Nearest Neighbor Algorithm," *International Journal of GEOMATE*, vol. 15, no. 48, pp. 169–174, 2018.

[24] M. Najafzadeh and G. Oliveto, "Riprap incipient motion for overtopping flows with machine learning models," *Journal of Hydroinformatics*, vol. 22, no. 4, pp. 749–767, 2020.

[25] M. Najafzadeh and S. Niazmardi, "A novel multiple-kernel support vector regression algorithm for estimation of water quality parameters," *Natural Resources Research*, vol. 30, no. 5, pp. 3761–3775, 2021.

[26] E. U. Eyo, S. J. Abbey, and C. A. Booth, "Strength predictive modelling of soils treated with calcium-based additives blended with eco-friendly pozzolans-A machine learning approach," *Materials*, vol. 15, no. 13, p. 4575, 2022.

[27] M. Shariati, M. S. Mafipour, P. Mehrabi et al., "Application of Extreme Learning Machine (ELM) and Genetic Programming (GP) to design steel-concrete composite floor systems at elevated temperatures," *Steel and Composite Structures*, vol. 33, pp. 319–332, 2019.

[28] Z. M. Yaseen, R. C. Deo, A. Hilal et al., "Predicting compressive strength of lightweight foamed concrete using extreme learning machine model," *Advances in Engineering Software*, vol. 115, pp. 112–125, 2018.

[29] M. Liang, Z. Chang, Z. Wan, Y. Gan, E. Schlangen, and B. Šavija, "Interpretable Ensemble-Machine-Learning models for predicting creep behavior of concrete," *Cement and Concrete Composites*, vol. 125, Article ID 104295, 2022.

[30] V. Q. Tran and H. Q. Do, "Prediction of California bearing ratio (CBR) of stabilized expansive soils with agricultural and industrial waste using Light gradient boosting machine," *J. Sci. Transp. Technol.*, 2021.

[31] B. Singh, P. Sihag, S. M. Pandhiani, S. Debnath, and S. Gautam, "Estimation of permeability of soil using easy measured soil parameters: assessing the artificial intelligence-based models," *ISH Journal of Hydraulic Engineering*, vol. 27, no. sup1, pp. 38–48, 2021.

[32] IS 2720-17 (1986), "Methods of Test for Soils, Part 17," *Laboratory determination of permeability*, 1986.

[33] B. T. Pham, M. D. Nguyen, N. Al-Ansari et al., "A comparative study of soft computing models for prediction of permeability coefficient of soil," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–11, 2021.

[34] B. T. Pham, H.-B. Ly, N. Al-Ansari, and L. S. Ho, "A comparison of Gaussian process and M5P for prediction of soil permeability coefficient," *Scientific Programming*, vol. 2021, pp. 1–13, 2021.

[35] Q.-A. T. Bui, N. Al-Ansari, H. V. Le, I. Prakash, and B. T. Pham, "Hybrid model: teaching learning-based optimization of artificial neural network (TLBO-ANN) for the prediction of soil permeability coefficient," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–9, 2022.

[36] M. Ahmad, S. Keawsawasvong, M. R. Bin Ibrahim, M. Waseem, K. R. Kashyzadeh, and M. M. S. Sabri, "Novel approach to predicting soil permeability coefficient using Gaussian process regression," *Sustainability*, vol. 14, no. 14, p. 8781, 2022.

[37] V. Q. Tran, H. T. Mai, Q. T. To, and M. H. Nguyen, "Machine learning approach in investigating carbonation depth of concrete containing Fly ash," *Structural Concrete*, 2020.

[38] S. Mordechai, *Applications of Monte Carlo Method in Science and Engineering*, InTech, United States, 2011.

[39] R. Fonseca-Delgado and P. Gomez-Gil, "An Assessment of Ten-fold and Monte Carlo Cross Validations for Time Series Forecasting," in *Proceedings of the 2013 10th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, IEEE, Mexico City, Mexico, October 2013.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 2017Adv. Neural Inf. Process. Syst., Curran Associates, Inc*, IEEE, December 2017.

[41] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014.

[42] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] E. C. Blessie and E. Karthikeyan, "Sigmis: a feature selection algorithm using correlation based method," *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385–394, 2012.

[45] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 6, pp. 237–297, 1995.

[46] Z. Yao and W. L. Ruzzo, "A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data," *BMC Bioinformatics*, vol. 7, no. S1, p. S11, 2006.

[47] T. K. Ho, "Random decision forests,"vol. 1, pp. 278–282, in *Proceedings of the 3rd International Conference on Document Analysis and Recognition Proc*, vol. 1, pp. 278–282, IEEE Computer Society, Montreal, QC, Canada, August 1995.

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

[49] A. Boulesteix, S. Janitza, J. Kruppa, and I. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics,"

Wiley Interdiscip. RevData Mining and Knowledge Discovery, vol. 2, 2012.

[50] H.-V. T. Mai, T.-A. Nguyen, H.-B. Ly, and V. Q. Tran, "Prediction compressive strength of concrete containing GGBFS using random forest model," *Advances in Civil Engineering*, vol. 2021, pp. 1–12, 2021.

[51] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[52] A. Natekin and A. Knoll, "Gradient boosting machines, A tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013.

[53] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[54] M. Najafzadeh, F. Saberi-Movahed, and S. Sarkamaryan, "NF-GMDH-Based self-organized systems to predict bridge pier scour depth under debris flow effects," *Marine Georesources & Geotechnology*, vol. 36, no. 5, pp. 589–602, 2018.

[55] M. Najafzadeh and F. Saberi-Movahed, "GMDH-GEP to predict free span expansion rates below pipelines under waves," *Marine Georesources & Geotechnology*, vol. 37, no. 3, pp. 375–392, 2019.

[56] F. Saberi-Movahed, M. Najafzadeh, and A. Mehrpooya, "Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: training group method of data handling using extreme learning machine conceptions," *Water Resources Management*, vol. 34, no. 2, pp. 529–561, 2020.