Hindawi
Scientific Programming
Volume 2021, Article ID 3625289, 13 pages
https://doi.org/10.1155/2021/3625289

Research Article

A Comparison of Gaussian Process and M5P for Prediction of Soil Permeability Coefficient

Binh Thai Pham,[1] **Hai-Bang Ly**,[1] **Nadhir Al-Ansari**,[2] **and Lanh Si Ho**[1,3]

[1]*University of Transport Technology, Ha Noi 100000, Vietnam*
[2]*Department of Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 971 87 Lulea, Sweden*
[3]*Civil and Environmental Engineering Program, Graduate School of Advanced Science and Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8527, Japan*

Correspondence should be addressed to Binh Thai Pham; binhpt@utt.edu.vn, Nadhir Al-Ansari; nadhir.alansari@ltu.se, and Lanh Si Ho; lanhhs@utt.edu.vn

Received 7 May 2021; Revised 13 October 2021; Accepted 15 October 2021; Published 31 October 2021

Academic Editor: Shah Nazir

The permeability coefficient (k) of soil is one of the most important parameters affecting soil characteristics such as shear strength or settlement. Thus, determining soil permeability coefficient is very crucial; however, a field test for determining this parameter is difficult, time-consuming, and expensive. In this study, soft computing methods, namely, M5P and Gaussian process (GP), for estimating the permeability coefficient were constructed and compared. The results of this paper indicate that the two soft computing algorithms functioned well in predicting k. These two methods gave high accuracy of prediction capability. The determination coefficient of M5P ($R^2 = 0.766$) was higher than that ($R^2 = 0.700$) of GP. This implies that the M5P model is more reliable estimation than the GP model in predicting soils' permeability coefficient (k). This proves that applying these machine learning techniques can provide an alternative for predicting basic soil parameters, including the permeability coefficient of soil.

1. Introduction

Soil permeability is one of the most important characteristics when considering its construction applications. The soil permeability coefficient is a factor that shows how the fluid can flow through the interconnected voids in soil from the high energy to low energy location due to the hydraulic gradient [1]. It is an important input parameter for designing most of the geotechnical structures, including landfills, tailing dams, or earth dams [2]. Desired values of soil permeability coefficient typically vary depending on soil types and service life of structures. For example, a high permeability coefficient is necessary for filter and drain construction, whereas a low permeability coefficient is required in the case of landfill liners or the core of earth dams. Many factors such as density, void size, void type (interconnected void), mineral composition, particle shape, and surface roughness of soil are key factors, which govern the variety of soil permeability [3]. Therefore, the construction of canals, dams, or drainage structures can be affected due to the variety of soil permeability [4].

The soil permeability coefficient can be measured through field or laboratory tests. It is demonstrated that the soil permeability coefficient determination in the field is costly, complicated, fairly laborious, and time-consuming [5–7]. Meanwhile, it is difficult for laboratory measurement of soil permeability coefficient to obtain the undisturbed samples. In particular, the samples for laboratory measurements are usually reconstituted to be close to those from the field. Therefore, the laboratory measurement results might not reflect the real value of soil permeability in the field because of the devastation of soil fabric when sampling [1]. The combined measurements of field and laboratory data are also carried out to determine the soil permeability coefficient because of the individual advantages and disadvantages of each test [1, 5].

Ganjidoost et al. [2] reported that three-category factors remarkably affect the soil permeability coefficient, namely, permeable soil parameters (density, clay content, viscosity, etc.), inherent soil parameters (Atterberg limits, particle size distribution, etc.), and compacted soil factors (porosity, water content, density, etc.). Most of these factors have closely related to each other. It was reported that the soil permeability coefficient was decreased by over 100 times when the percentage of passing through sieve No. 100 increased in the range of 0 to 7% [8]. Conducting several experiments with the difference in percentages of granular and low-plastic marine soils, Shakoor and Cook [9] concluded that the soil permeability coefficient was noticeably grown up by increasing the percentage of granular material. Similarly, D'Appolonia [10] investigated soils combining bentonite and demonstrated that the increase in the fine particles of soils resulted in the reduction of their permeability coefficient. For the cohesive soils, the permeability coefficient can be reduced when increasing the plasticity index of clayey soils [11]. It was reported that the compaction energy and water content significantly influence the soil permeability coefficient [12]. Acar and Olivieri [13] verified that void size as well as the void ratio was remarkably decreased after increasing the compaction energy, leading to a reduction of soil permeability coefficient. Therefore, researchers intensively made efforts to estimate the soil permeability coefficient by suggesting many empirical formulae for not only cohesive soils [11, 14, 15], but also granular soils [16–19]. However, most of the empirical formulae were proposed based on the results obtained in the laboratory, which were conducted on either rebuilt or disturbed specimens. Moreover, samples for measurement in the laboratory are usually limited. The accuracy of empirical formulae depends on the quality of the preparation and experiment process of samples such as selecting samples, preparing homogeneous samples, and using appropriate methods. Therefore, these formulae can only apply to some specific cases but cannot be used to calculate the soil permeability coefficient for all cases.

Recently, several soft computing techniques have been proposed and applied to identify or predict parameters of soils such as artificial neural networks (ANNs), adaptive network-based fuzzy inference system (ANFIS), and hybrid optimization model of genetic algorithm adaptive network-based fuzzy inference system (GA-ANFIS) [6, 7, 18, 20–22]. For example, many researchers used machine learning models such as ANN, ANFIS, and SVM or hybrid machine learning models like PSO-MLP neural nets to estimate the compression coefficient of soil [23–25]. They indicated that these machine learning methods could predict compression coefficients with high accuracy. Besides, machine learning methods such as SVR, ELM, ANN, PANFIS, GANFIS, and other hybrid models have been applied successfully in predicting the shear strength of soil [26–29]. It was reported that ANN, SVM, and ANFIS have some advantages in predicting soil parameters. For example, ANN has a simple architecture and is easy in training and generalization; it can solve nonlinear problems with high accuracy [30–33]. Regarding SVMs, it was reported that they have some advantages such as the ability to provide good out-of-sample generalization and to be robust even when the training sample has some bias [34]. Based on experimental results of 55 different mixture proportions, Sinha and Wang [7] developed the ANN prediction models (with reliability of over 95%) including the permeability, maximum dry density, and moisture content to verify the properties of soil. Concerning particle shape and grain size distribution, the permeability of granular soil can be also predicted by using ANFIS [22]. Yilmaz et al. [20] compared the predicted permeability coefficient of coarse-grained soils between ANNs and ANFIS. Although both soft computing models can exhibit a high accuracy in estimating the soil permeability coefficient, the ANFIS model might outperform the ANNs model. Besides, some previous studies used multiple linear regression (MLR), ANN, SVM, and ANFIS to improve the prediction accuracy of hydraulic conductivity of soil [35]. Arshad et al. concluded that ANFIS has a better prediction ability compared to ANN and MLR in estimating the saturated hydraulic conductivity [36–40]. However, it was reported that these machine learning algorithms have some disadvantages such as having a greater computational resource, being time-consuming, having poor generalization, and being prone to overfitting [30, 31, 34, 41, 42].

Furthermore, there are several decision tree algorithms, which have been applied popularly in predicting soil parameters such as M5P and Gaussian process (GP). It was reported that M5P and GP have some advantages. For example, it was indicated that these algorithms require few user-defined parameters, can provide mathematical equations, offer more insight into the obtained equations, and also are more convenient to develop and implement [42–45]. M5P and GP have been employed in predicting structural numbers in flexible pavements, and it was indicated that these machine learning methods could be used in this problem [46]. Previous studies used M5P for predicting the compressive strength of normal concrete and high-performance concrete; they found that M5P could be a sufficient tool for estimating the compressive strength of concrete [42, 45, 47].

Regarding the prediction of soil properties, GP was employed to predict the ultimate load capacity as well as the effective stress parameter of unsaturated soil [48, 49]. It was revealed that the GP regression approach works well in the prediction of the load-bearing capacity of the pile in comparison with the SVM approach [48]. In addition, Samui and Jagan [49] indicated that GPR is a reliable model for predicting the effective stress parameter of unsaturated soil. A previous study used Random Forest (RF), GP, M5P, and ANN for estimating the strength of cement-treated dispersive soil, finding that these algorithms performed well [50]. Besides, it was reported that GP works better than SVR and MLR in the prediction of infiltration of sandy soil [43]. GP was also used to model the infiltration rate of the soil, and it was stated that GP could give a good estimation performance [51]. Furthermore, another study applied M5P in the prediction of cumulative infiltration of sandy soil and found that the bagged approach performed well with the M5P tree model than with the RF model [52]. GP was also

used for modeling the recharging rate of the stormwater filter system, and it was indicated that Pearson VII based GP regression approach works well compared to the other kernel functions based on GP and SVM models [53].

Based on the above literature, it is accepted that M5P and Gaussian process (GP) are the decision tree algorithms that are robust soft computing techniques and have the potential for predicting soil parameters [45, 47, 54–56]. They can provide understandable mathematical equations. Consequently, users can more easily know the parameters that affect the outputs of the modeling process. However, there is less application of these prominent techniques in the geotechnical field, for example, for the prediction of the soil permeability coefficient. Therefore, this study was extensively undertaken with the following objectives: to present the application of soft computing techniques, M5P and Gaussian process, for estimating the soil permeability coefficient and to compare the accuracy between the two techniques. In addition, this study will also fill the gap of literature using M5P and GP for predicting soil permeability coefficient. The used datasets were collected from a project in Vietnam; Da Nang–Quang Ngai expressway was employed for modeling.

2. Materials and Methods

2.1. Data Used. Indeed, the permeability coefficient of soil (k) is affected by many factors. However, this study will focus on the main factors which significantly govern soil permeability in order to reduce the model complexity. In the current research, data of 84 soil samples were collected from Da Nang–Quang Ngai expressway project as shown in Figure 1. Then, the soil samples were tested in the laboratory to determine the input parameters, namely, water content (w%), void ratio (e), specific density (γ), liquid limit (LL%), plastic limit (PL%), clay content, and permeability coefficient (k). The output of these parameters in modeling is the permeability coefficient of soils. The quantitative analysis of these input parameters is provided in Table 1. Figure 2 shows the distribution of input and output variables used in this study. Figure 3 shows the correlation between the input variables and the output variable. It can be seen that there is a high correlation between the input variables and the output variables of the data used in this study (in most of the cases, $R > 0.5$).

2.2. Methods Used

2.2.1. M5P. Wang and Witten [57] rebuilt and proposed the M5P algorithm from the M5 algorithm, which was originally proposed by Quinlan [21] with the addition of a linear regression function to the leaves nodes. By reducing tree size, M5P could perform better on datasets than M5. Normally, M5P has three main steps as follows:

(i) Building a model tree: the entered space was split into many subspaces using the dividing criterion. The standard deviation reduction factor (SDR) was used to minimize the expected error at the node. The equation of SDR is shown as follows:

$$\mathrm{SDR} = \mathrm{sd}(H) - \sum_i \frac{|H_i|}{|H|} \times \mathrm{sd}(H_i),$$

$$\mathrm{sd}(H) = \sqrt{\sum_1^N \frac{(H_i - \overline{H})^2}{N-1}}, \quad (1)$$

$$\overline{H} = \sum_i^N \frac{H_i}{N},$$

where H is the instances dataset that stretch the node, H_i is the set that is received from a divided node according to a given attribute, and sd is the standard deviation of H.

(ii) Pruning tree: a classifying and regression tree (CART) is introduced in each subspace to overfit the problem and increase the classification performance. The pruning step will delete the error which occurs in the learning data.

(iii) Smoothing step: the pruning tree can lead to sharp discontinuities among the neighboring linear models. Therefore, to solve this problem, all the leaf models will be combined from the leaf to the root to build the final model. Then, along the path back, the predicted value is filtered to the root. By regression, the final value is smoothed by combining the current value with the predicted value from the linear regression as the following equation:

$$T' = \frac{Nt + KA}{N + K}, \quad (2)$$

where T' is the predicted value shift to the higher level of the next node, N is the total number of training instances that shift to the next lower node, Nt is the predicted value shifted from the lower node to the present node, A is the predicted value by the node at this node, and K is a constant value.

2.2.2. Gaussian Process. The Gaussian process (GP) was firstly introduced by Rasmussen and Williams [58]. This is a popular method for nonparametric regression and classification problems. This method was also applied to predict the compressive strength of concrete [45, 47]. By combining the Bayesian learning and kernel machines, GP results in a principled and probabilistic approach for regression. The predicted value can be directly outputted with the uncertainty of a model prediction. Hence, this is a suitable model for dealing with time series data mining.

Generally, a GP can be measured by the mean and kernel function. It is assumed that GP is classified as an assemblage of random variables, which show the value of the function $f(t)$ at the location (t). It can be written as the following equations:

Figure 1: Location of Da Nang–Quang Ngai expressway project.

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

Parameters	Unit	Minimum	Maximum	Average	STD
Natural water content, w	%	15.1	99.9	34.23	16.5
Void ratio, e	—	0.46	2.63	0.97	0.42
Specific density, γ	g/cm^3	2.58	2.74	2.68	0.02
Liquid limit, LL	%	18.9	88.93	37.27	13.04
Plastic limit, PL	%	12.2	54.8	22.21	7.04
Clay content, cc	%	5.7	64	25.17	11.5
Permeability coefficient k	10^{-9} cm/s	0.3	7.1	1.45	0.94

STD = standard deviation.

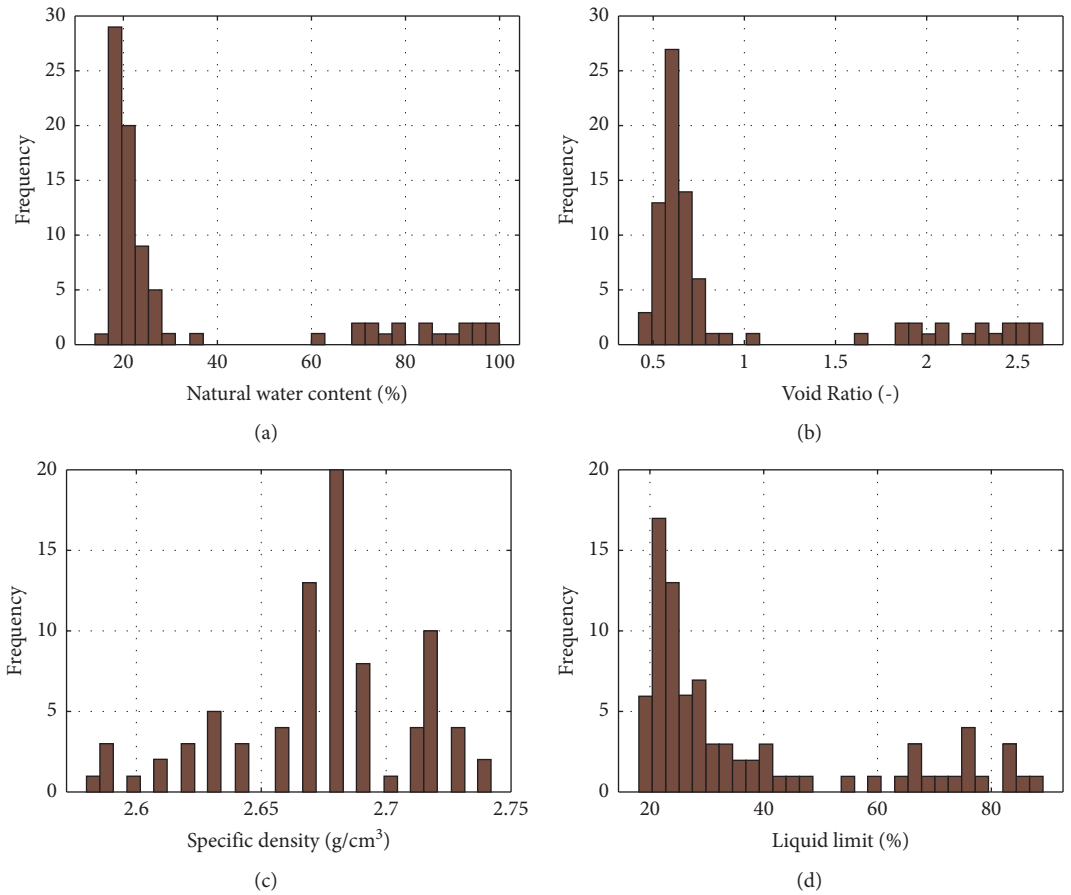

Figure 2: Continued.

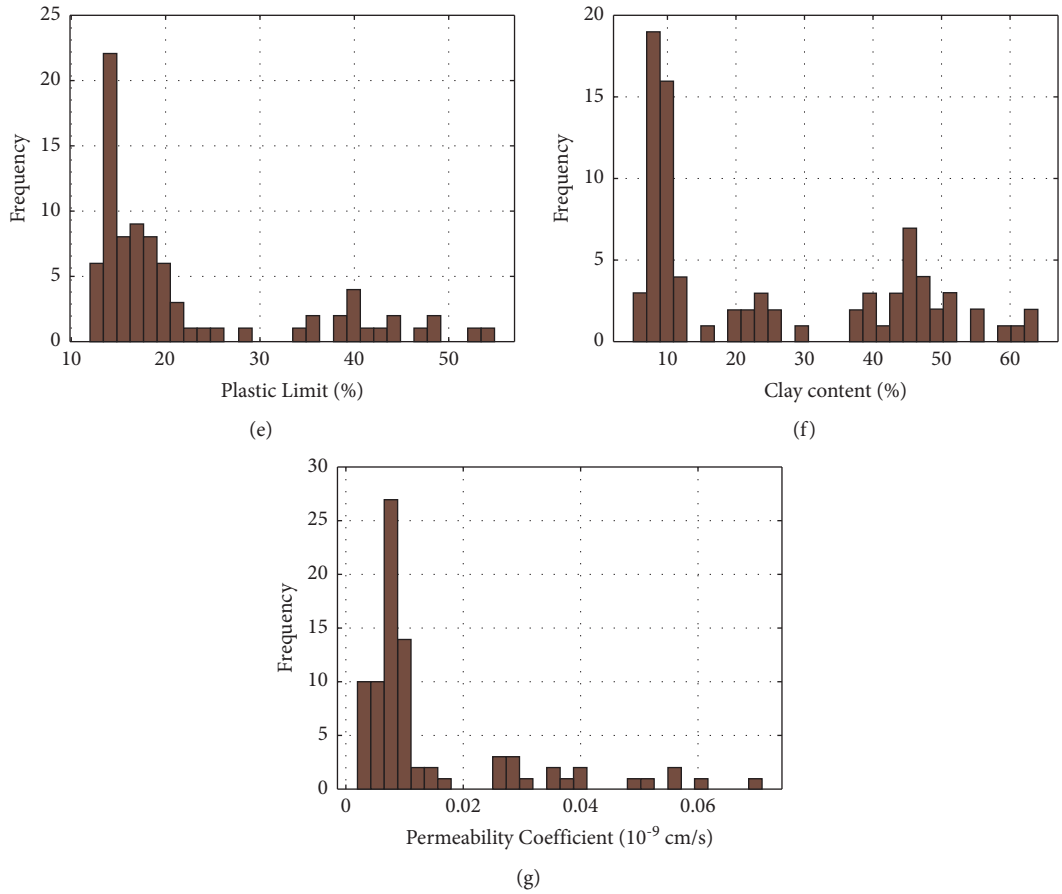

Figure 2: Histograms of the input and output variables used in this study.

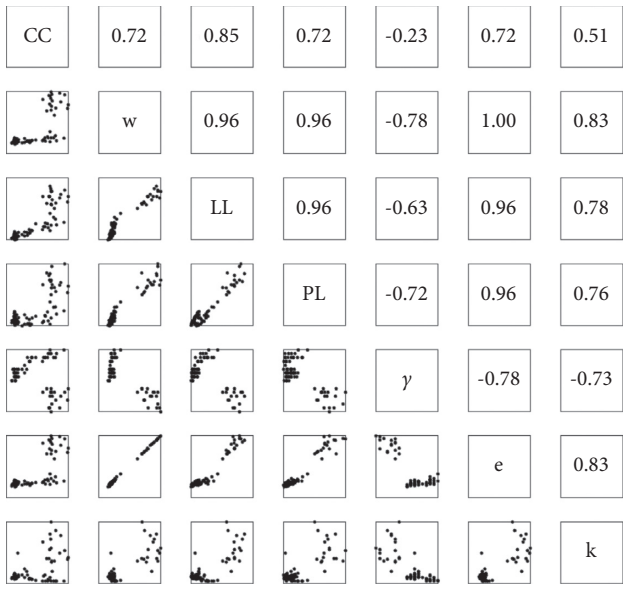

Figure 3: Scatter plots for each pair of variables with the permeability, including correlation analysis between variables used in this study.

$$f(t) \sim \mathrm{GP}(m(t), k(t, t')), \tag{3}$$

$$\begin{aligned} m(t) &= \mathbb{E}[f(t)], \\ k(t, t') &= \mathbb{E}[(f(t) - m(t))(f(t') - m(t'))], \end{aligned} \tag{4}$$

where $f(t)$ is the prior distribution of regression function and $m(t)$ and $k(t, t')$ are the mean and kernel function, respectively.

Suppose that a training set T consists of input finite numbers $t_1, \ldots, t_n$ in matrix form; GP will define a joint Gaussian distribution as follows:

$$p(f|T) = N(f|M, K), \tag{5}$$

where the mean function $M(T)$ is determined by the mean function $m(t)$, as follows:

$$M(T) = \begin{bmatrix} m(t_1) \\ \cdots \\ m(t_N) \end{bmatrix}. \tag{6}$$

Furthermore, the kernel function $K(T, T')$ is measured by the mean function $k(t, t')$, as follows:

$$K(T, T') = \begin{bmatrix} k(t_1, t_1') & \cdots & k(t_1, t_N') \\ \cdots & \cdots & \cdots \\ k(t_N, t_1') & \cdots & k(t_N, t_N') \end{bmatrix}. \quad (7)$$

For simplicity, the mean function in this study is set as zero in order to obtain a widely used GP prior. It was also applied in previous studies [56, 58]. Then, (3) can be rewritten as follows:

$$f(t) \sim \mathrm{GP}(m(t) = 0, k(t, t')). \quad (8)$$

2.2.3. Validation Indicators. In this research, several error indicators, namely, root mean square (RMSE), mean absolute error (MAE), and determination coefficient (R^2), were used to quantify the accuracy of the prediction of models. Particularly, RMSE shows the difference in the values between the actuality and prediction, as shown in (9). A lower value of RMSE means a higher accuracy of estimation. Meanwhile, MAE presents the average error of the actuality and prediction, as presented in (10). A lower MAE indicates a more precise model. In contrast, R is the standardized values observed from the model's prediction errors, as shown in (11). The value of R^2 varies from 0 to 1. A higher value of R^2 indicates a higher estimation ability, and a value close to 1 shows good accuracy.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(\dot{k} - \hat{k})^2}{N}}, \quad (9)$$

$$\mathrm{MAE} = \sum_{i=1}^{N} \frac{|\dot{k} - \hat{k}|}{N}, \quad (10)$$

$$R^2 = \frac{\sum_{i=1}^{N} (\dot{k} - \overline{\dot{k}})(\hat{k} - \overline{\hat{k}})}{\sqrt{\sum_{i=1}^{N} (\dot{k} - \overline{\dot{k}})^2 (\hat{k} - \overline{\hat{k}})^2}}, \quad (11)$$

where $\dot{k}$ and $\hat{k}$ are the computed and actual value, $\overline{\dot{k}}$ and $\overline{\hat{k}}$ are the average of computed and actual value, and N is the total number of samples.

2.3. Methodology. The present study is carried out based on the proposed methodology that comprises three main steps as follows: (1) data preparation, (2) construction of the models, and (3) validation of the proposed models (Figure 4):

(1) Data preparation: in this first step, the data of samples from the laboratory was employed to create the testing and training dataset. The training dataset was generated from 50% of the total data, and the testing dataset was built from the remaining 50%.

(2) Construction of the models: in this second step, the training dataset was employed for training the models based on M5P and GP algorithms. In the GP, we have used the RBF kernel, a batch size of 100, and 1 seed for training the model. In the M5P, we have used a batch size of 100 and a minimum number of instances to allow at the leaf node of 4. All training and validating processes were carried out in Weka software.

(3) Validation of the proposed models: in this final step, the testing dataset was adopted for validating the proposed models. Statistical indicators including RMSE, MAE, and R^2 were employed to validate the models.

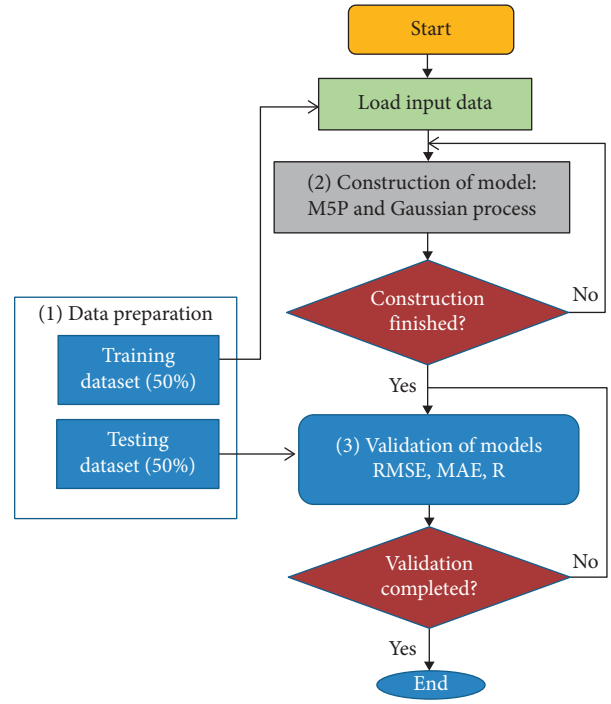

Figure 4: Flowchart of the proposed methodology.

3. Results and Discussion

Training and validation of the M5P model were done using training and testing datasets. In this study, the M5P was trained with the number of instances to allow at a leaf node of 4 and a batch size of 100. The training and validation results of the M5P model are presented in Figures 5–7. It can be seen from Figure 5 that the actual values (determined from the experiments) and predicted values (predicted by the M5P model) are very close with low error values (RMSE = 0.0064 and MAE = 0.004) (Figure 5) and a high determination coefficient ($R^2 = 0.792$) (Figure 7) in the case of the training dataset. This indicates that the M5P model has a great degree of goodness of fit with the data used. In the case of the testing dataset, there exists a good agreement between actual and predicted values, the error values of the M5P models are very low (RMSE = 0.0081 and MAE = 0.0045) (Figure 6), and the determination coefficient is high ($R^2 = 0.766$) (Figure 7). This indicates that the predictive capability of the M5P model is good.

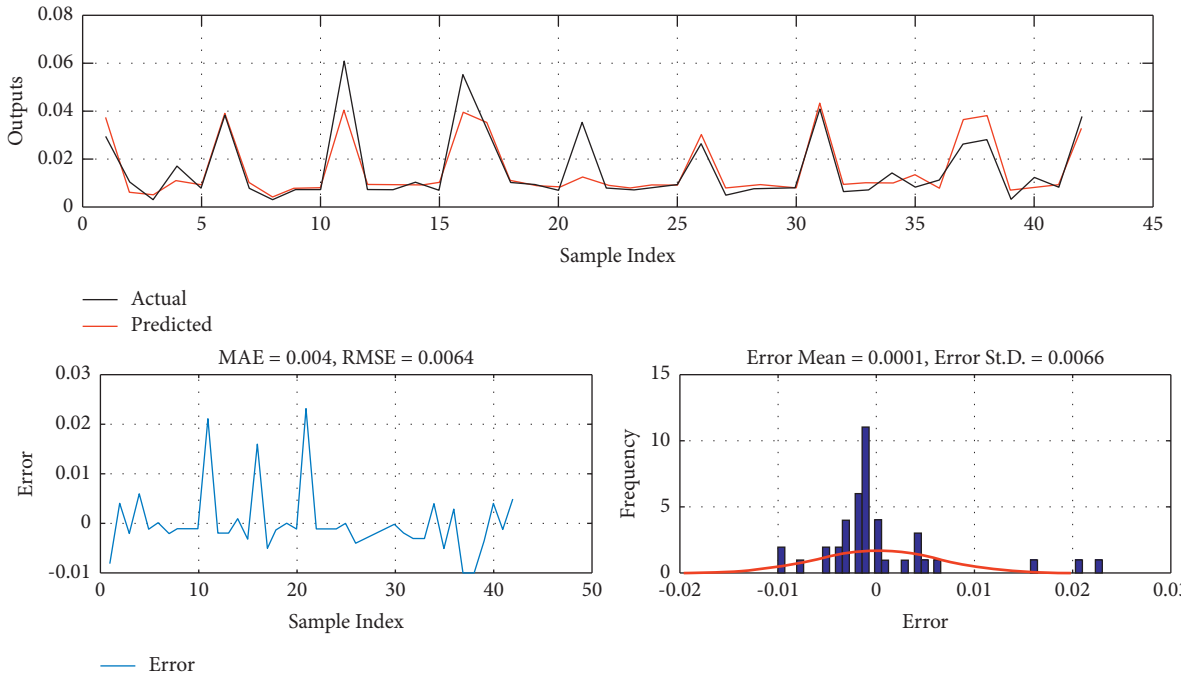

Figure 5: Error analysis of M5P using the training dataset.

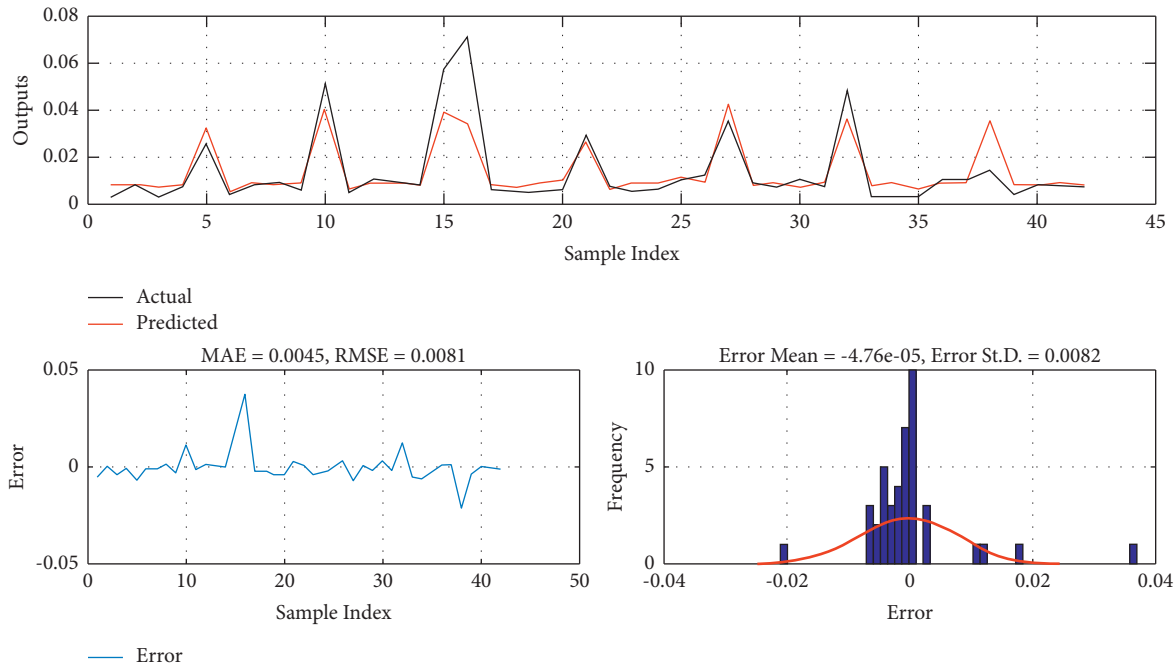

Figure 6: Error analysis of M5P using the testing dataset.

Training and validation of the GP model were carried out using training and testing datasets. In this study, the GP was trained using polynomial kernel with the normalized training data. The level of Gaussian noise was determined as "1" and the batch size and number of seeds were set as "100" and "1." The training and validation results of the GP model are presented in Figures 8–10. It can be observed from Figure 8 that the actual values (determined from the experiments) and predicted values (predicted by the GP model) are very close with low error values (RMSE = 0.0077 and MAE = 0.0047) (Figure 9) and a high determination coefficient ($R^2 = 0.71$) (Figure 10) in the case of the training dataset. This indicates that the GP model has a good degree of goodness of fit with the data used. In the case of the testing dataset, there exists a good agreement between actual and predicted values, the error values of the GP model are very

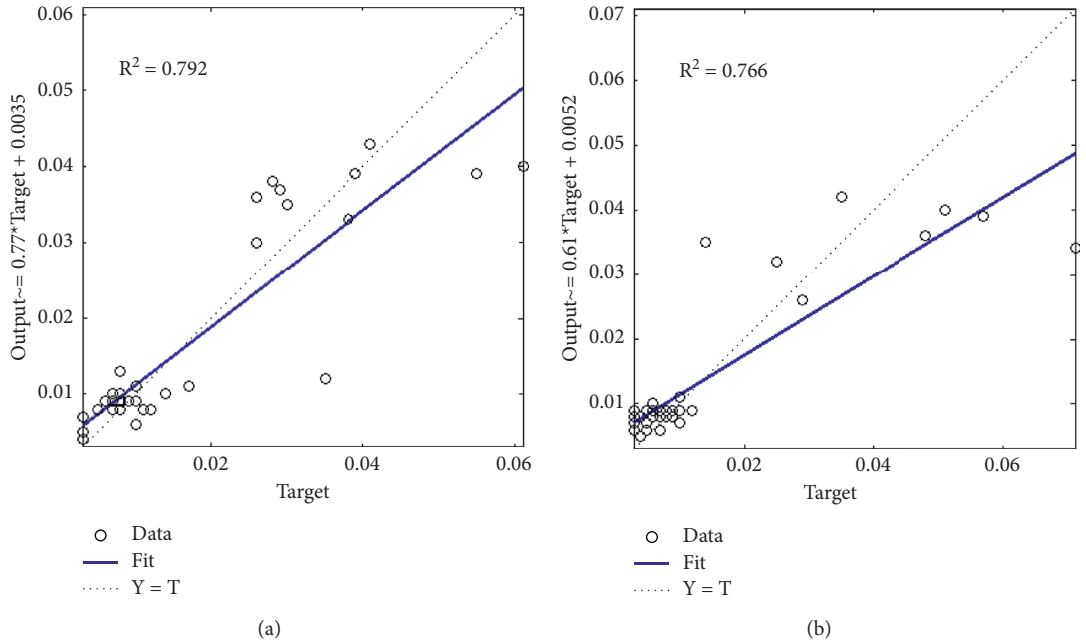

Figure 7: Correlation analysis of actual and predicted outputs using M5P: (a) training; (b) testing.

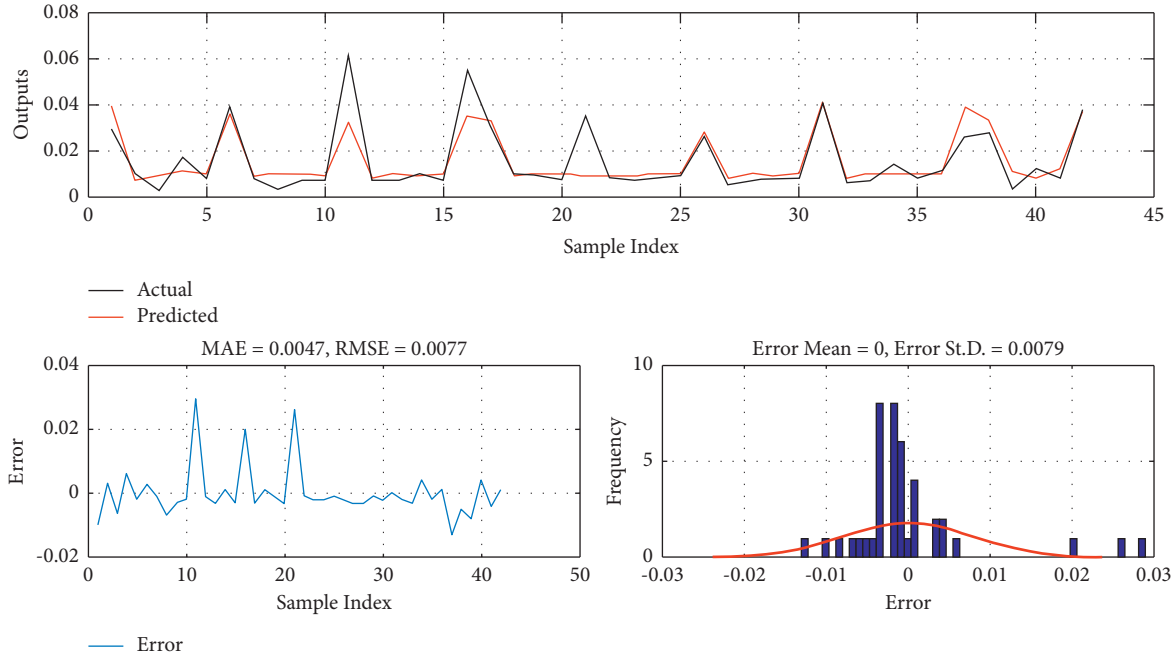

Figure 8: Error analysis of M5P using the testing dataset.

low (RMSE = 0.0093 and MAE = 0.0054) (Figure 9), and the determination coefficient is high ($R^2 = 0.700$) (Figure 10). This indicates that the predictive capability of the GP model is good.

The comparison results show that the error values of the M5P model are lower than those of the GP model in both the training (Figures 5 and 8) and testing (Figures 6 and 9) datasets. In the case of the correlation analysis, the values of the determination coefficient of the M5P model are higher than those of the GP model in both the training (Figure 7) and testing (Figure 10) datasets.

In general, it can be stated that both M5P and GP models are good for the prediction of the soil permeability coefficient, but the M5P model outperforms the GP model in this study. It is reasonable as the advantage of M5P compared with other models is that it is able to deal with both

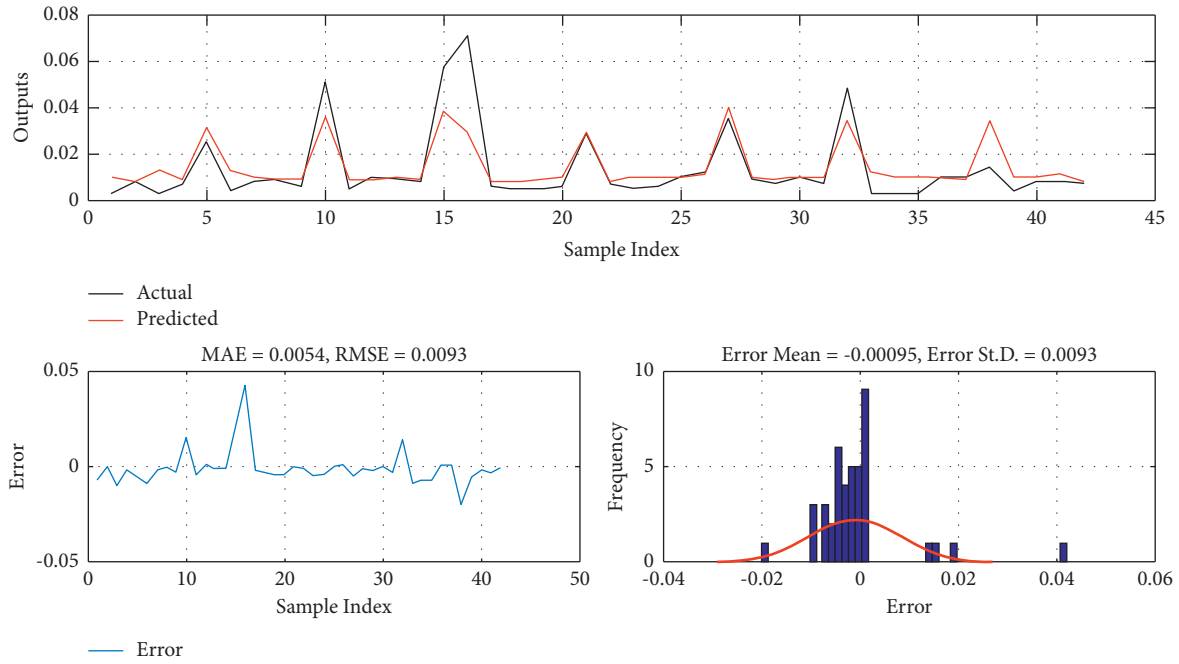

Figure 9: Error analysis of GP using the testing dataset.

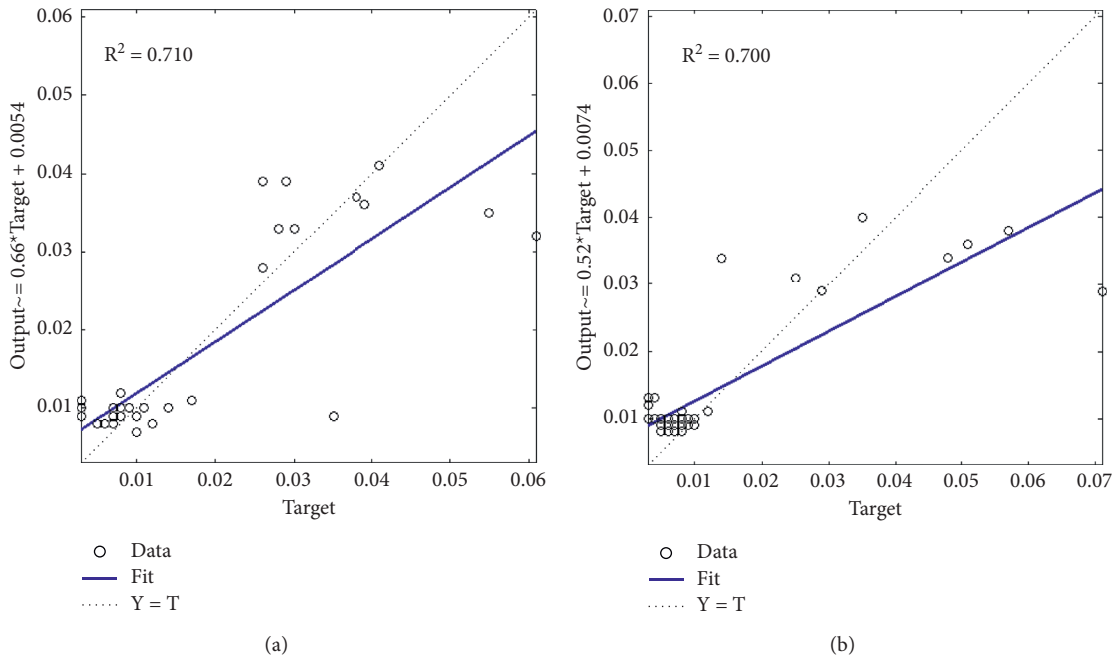

Figure 10: Correlation analysis of actual and predicted outputs using GP: (a) training; (b) testing.

continuous and categorical variables, and it has a great capability to handle the variables with missing values [45]. In addition, the M5P takes advantage of decision trees; in general, it requires no assumption related to the probability distribution of the used data [59]. Regarding the GP model, it has only few parameters to be tuned, and thus it can be trained with a small training dataset. However, the GP model is considered as a black-box model, and thus, it lacks transparency [60]. In comparison with previously published works [61–63], the results of this modeling study reveal that there are some data points associated with very large prediction errors of the models (Figures 5, 6, 8, and 9), which can be caused by the large deviations and high correlation between input and output variables of the data used in this study (Figure 3). Therefore, it is recommended that these models (M5P and GP) can be tested and validated with other larger datasets with lower variable correlation for better performance.

Table 2: Dataset used in this study to predict permeability coefficient, k.

Clay content, cc (%)	Water content, w (%)	Liquid limit, LL (%)	Plastic limit, PL (%)	Specific density, γ (g/cm^3)	Void ratio, e	Permeability coefficient, k (10^{-9} cm/s)
44	93.73	75.62	46.8	2.59	2.453	0.029
21.7	20.71	24.58	13.5	2.72	0.639	0.01
51.8	20.98	38.17	20.2	2.73	0.625	0.003
19	24.55	29.08	19.6	2.68	0.707	0.017
8.5	19.85	23.67	17.58	2.67	0.599	0.008
44.8	79.96	75.45	43.6	2.59	2.083	0.039
8.7	15.09	18.9	12.63	2.66	0.462	0.008
59.4	24.95	41.87	22.3	2.74	0.713	0.003
8.9	21.79	24.98	19	2.68	0.654	0.007
8.4	19.46	22.97	17.43	2.68	0.605	0.007
51.1	73.75	66.96	35.8	2.61	1.966	0.061
19	18.35	23.61	13.35	2.7	0.579	0.007
8.3	18.01	21	14.2	2.67	0.599	0.007
8.2	17.12	19.7	13.8	2.67	0.571	0.01
6.1	16.97	21.01	15.87	2.66	0.556	0.007
56.1	83.25	78.23	41.9	2.62	2.235	0.055
64	78.72	75.53	39.5	2.64	2.106	0.03
16.1	17.52	25.85	12.2	2.69	0.546	0.01
8.4	18.02	21.1	14.5	2.67	0.552	0.009
10.7	24.53	27.22	19.6	2.69	0.713	0.007
11.7	19.77	23.91	13.5	2.68	0.567	0.035
9.5	18.12	21.2	14.5	2.68	0.567	0.008
7.6	20.23	23.62	16.8	2.69	0.64	0.007
11	20.14	22.78	16.1	2.67	0.608	0.008
9.4	19.64	23.8	17.2	2.67	0.648	0.009
49.4	62.2	59.99	38.5	2.63	1.657	0.026
25.9	21.23	31.18	13.2	2.72	0.609	0.005
8.6	20.12	20.82	14.8	2.67	0.599	0.007
10.7	17.25	19.5	13.5	2.68	0.558	0.008
9.3	21.14	23.89	18.53	2.68	0.686	0.008
46.4	99.9	82.11	43.6	2.58	2.634	0.041
24.5	18.28	28.11	12.5	2.71	0.522	0.006
9.7	17.34	20.49	14.3	2.66	0.486	0.007
21	20.62	28.62	17.4	2.69	0.592	0.014
12.5	19.25	23.46	14.67	2.67	0.628	0.008
8.1	23.28	26.8	20.36	2.68	0.707	0.011
46.9	95.58	82.25	53	2.6	2.514	0.026
63.4	73.1	68.47	35	2.61	1.933	0.028
42.5	27.28	39.99	21.74	2.72	0.789	0.003
8.6	18.02	20.51	14.6	2.69	0.592	0.012
8.5	25.49	27.49	21.32	2.67	0.723	0.008
60.2	95.09	84.05	54.8	2.63	2.507	0.038
40.3	20.75	40.77	18.64	2.72	0.591	0.003
8.4	18.25	21.08	14.5	2.69	0.592	0.008
50.7	28.97	46.04	25.2	2.72	0.889	0.003
8.8	17.19	19.81	14.3	2.68	0.549	0.007
46.6	76.77	64.83	38.17	2.63	2.023	0.025
45	35.53	53.56	28.6	2.74	1.015	0.004
9.6	17.99	20.42	15	2.67	0.571	0.008
8.6	19.9	23	16.9	2.68	0.586	0.009
9.6	18.18	22.58	16	2.68	0.567	0.006
45.8	89.51	85.86	42.7	2.63	2.372	0.051
43.4	25.6	34.5	15.6	2.73	0.717	0.005
9.2	17.81	21	14.3	2.68	0.506	0.01
7.7	21.23	25.3	18.5	2.68	0.654	0.009
9.4	17.85	20.48	14.8	2.68	0.558	0.008
45.1	93.19	88.93	48	2.62	2.447	0.057
46.1	70.21	65.46	33.6	2.64	1.87	0.071
23.6	18.84	27.48	13.8	2.71	0.604	0.006

Table 2: Continued.

Clay content, cc (%)	Water content, w (%)	Liquid limit, LL (%)	Plastic limit, PL (%)	Specific density, γ (g/cm^3)	Void ratio, e	Permeability coefficient, k (10^{-9} cm/s)
26.5	21.89	30.98	17.4	2.72	0.619	0.005
23.5	21.32	32.23	16.4	2.71	0.604	0.005
5.7	17.35	20.34	14.25	2.66	0.494	0.006
41.9	69.26	66.42	48.5	2.64	1.87	0.029
45.3	19.6	30.92	13.2	2.73	0.569	0.007
8.5	20.81	25.31	18.53	2.68	0.576	0.005
10.2	18.15	22.14	15.6	2.67	0.517	0.006
12.7	22.71	28.5	17.8	2.69	0.671	0.01
8.6	24.84	29.32	22	2.68	0.752	0.012
55.3	98.01	73.63	40.1	2.59	2.597	0.035
38.6	22.79	35.83	15.2	2.72	0.689	0.009
9.7	18.02	20.51	14.2	2.68	0.605	0.007
40.1	25.53	36.11	19.2	2.72	0.755	0.01
9.7	18.01	20.3	14.2	2.67	0.599	0.007
37.6	87.71	75.34	40.5	2.63	2.329	0.048
49	25.45	48.24	24.8	2.72	0.711	0.003
37.4	21.13	32.44	14.2	2.71	0.642	0.003
46.1	25.78	38.03	17.5	2.73	0.808	0.003
9.8	18.07	20.62	14.5	2.68	0.567	0.01
8	18.05	20.99	14.3	2.68	0.595	0.01
47.5	85.35	71.24	40.5	2.62	2.275	0.014
30.4	22.23	39.53	18.64	2.72	0.648	0.004
9.8	22.03	23.92	17.8	2.68	0.644	0.008
9.2	23.97	26.52	19.8	2.67	0.723	0.008
6.7	18.91	21.49	15	2.69	0.582	0.007

4. Conclusion

In this paper, soft computing techniques comprising M5P and Gaussian process (GP) were used for estimating soil permeability coefficient and compared. The results indicated that it is possible to estimate the permeability coefficient (k) of cohesive soil from basic soil parameters (water content, density, liquid and plastic limits, void ratio, and clay content) by using the models of soft computing. Some conclusions of this study can be derived as follows:

(1) The results of M5P and GP indicated that the performance of models is high, with the determination coefficients of M5P and GP being 0.766 and 0.700, achieved from the correlation between actual and computed values of k, respectively.

(2) The M5P model's estimation of the permeability coefficient was found to be more reliable than that of the GP model.

(3) The results of this research also point out that these machine learning techniques can be a potential approach for estimating basic soil parameters like soil permeability coefficient.

This study reveals that M5P and GP can be applied for predicting the permeability coefficient of soil with high accuracy. However, the sample numbers as well as soil types are limited. Thus, this study should be extended to more sample numbers and other soil types such as granular soil or clay. Furthermore, future studies using other algorithms such as genetic programming, gene expression programming, and evolutionary polynomial regression should be carried out to evaluate the effectiveness of the used algorithms and to have a full picture of the techniques used for predicting the consolidation coefficient of soil (Table 2).

Data Availability

The dataset for simulation can be found in Table 2.

Conflicts of Interest

The authors declare that there are 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] H. Ganjidoost, S. J. Mousavi, and A. Soroush, "Adaptive network-based fuzzy inference systems coupled with genetic algorithms for predicting soil permeability coefficient," *Neural Processing Letters*, vol. 44, no. 1, pp. 53–79, 2016.

[3] S. Dong, Y. Guo, and X. Yu, "Method for quick prediction of hydraulic conductivity and soil-water retention of unsaturated soils," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 52, pp. 108–117, 2018.

[4] V. K. Singh, D. Kumar, P. S. Kashyap, P. K. Singh, A. Kumar, and S. K. Singh, "Modelling of soil permeability using different data driven algorithms based on physical properties of soil," *Journal of Hydrology*, vol. 580, Article ID 124223, 2020.

[5] T. Vienken and P. Dietrich, "Field evaluation of methods for determining hydraulic conductivity from grain size data," *Journal of Hydrology*, vol. 400, no. 1-2, pp. 58–71, 2011.

[6] K. R. Rehfeldt, J. M. Boggs, and L. W. Gelhar, "Field study of dispersion in a heterogeneous aquifer: 3. Geostatistical analysis of hydraulic conductivity," *Water Resources Research*, vol. 28, no. 12, pp. 3309–3324, 1992.

[7] S. K. Sinha and M. C. Wang, "Artificial neural network prediction models for soil compaction and permeability," *Geotechnical & Geological Engineering*, vol. 26, no. 1, pp. 47–64, 2008.

[8] H. R. Cedergren, Ed., *Seepage Drainage and Flow Nets*, Wiley, London, United Kingdom, Third edition, 1988.

[9] A. Shakoor and B. D. Cook, "The effect of stone content, size, and shape on the engineering properties of a compacted silty clay," *Environmental and Engineering Geoscience*, vol. xxvii, no. 2, pp. 245–253, 1990.

[10] D. J. D'Appolonia, "Soil-bentonite slurry trench cutoffs," *Journal of the Geotechnical Engineering Division*, vol. 106, no. 4, pp. 399–417, 1980.

[11] G. Mesri and R. E. Olson, "Mechanisms controlling the permeability of clays," *Clays and Clay Minerals*, vol. 19, no. 3, pp. 151–158, 1971.

[12] J. K. Mitchell and K. Soga, *Fundamentals of Soil Behavior*, Wiley, NJ, USA, Third edition, 2005.

[13] Y. B. Acar and I. Olivieri, *Pore Fluid Effects on the Fabric and Hydraulic Conductivity of Laboratory-Compacted clay*, pp. 144–159, Transportation Research Board, Washington D. C., USA, 1989.

[14] H. Pincus, P. R. Narasimha, N. Pandian, and T. Nagaraj, "Analysis and estimation of the coefficient of consolidation," *Geotechnical Testing Journal*, vol. 18, no. 2, p. 252, 1995.

[15] R. P. Chapuis, "Predicting the saturated hydraulic conductivity of soils: a review," *Bulletin of Engineering Geology and the Environment*, vol. 71, no. 3, pp. 401–434, 2012.

[16] R. P. Chapuis, D. E. Gill, and K. Baass, "Laboratory permeability tests on sand: influence of the compaction method on anisotropy," *Canadian Geotechnical Journal*, vol. 26, no. 4, pp. 614–622, 1989.

[17] J. Odong, "Evaluation of empirical formulae for determination of hydraulic conductivity based on grain-size analysis," *The Journal of American Science*, vol. 4, 2008.

[18] A. E. Cronican and M. M. Gribb, "Hydraulic conductivity prediction for sandy soils," *Ground Water*, vol. 42, no. 3, pp. 459–464, 2004.

[19] R. P. Chapuis, "Predicting the saturated hydraulic conductivity of sand and gravel using effective diameter and void ratio," *Canadian Geotechnical Journal*, vol. 41, no. 5, pp. 787–795, 2004.

[20] I. Yilmaz, M. Marschalko, M. Bednarik, O. Kaynar, and L. Fojtova, "Neural computing models for prediction of permeability coefficient of coarse-grained soils," *Neural Computing & Applications*, vol. 21, no. 5, pp. 957–968, 2012.

[21] J. R. B. Quinlan, *Learning with Continuous Classes*, World Scientiic, Singapore, 1992.

[22] A. Sezer, A. B. Göktepe, and S. Altun, "Estimation of the permeability of granular soils using neuro-fuzzy system," in *Proceedings of the CEUR Workshop Proceedings*, pp. 333–342, Thessaloniki, Greece, April 2009.

[23] 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," *The Science of the Total Environment*, vol. 679, pp. 172–184, 2019.

[24] D. T. Bui, V. H. Nhu, and N. D. Hoang, "Prediction of soil compression coefficient for urban housing project using novel integration machine learning approach of swarm intelligence and multi-layer perceptron neural network," *Advanced Engineering Informatics*, vol. 38, pp. 593–604, 2018.

[25] S. Kirts, O. P. Panagopoulos, P. Xanthopoulos, and B. H. Nam, "Soil-compressibility prediction models using machine learning," *Journal of Computing in Civil Engineering*, vol. 32, Article ID 04017067, 2018.

[26] B. T. Pham, T. A. Hoang, D. M. Nguyen, and D. T. Bui, "Prediction of shear strength of soft soil using machine learning methods," *Catena*, vol. 166, pp. 181–191, 2018.

[27] Q. H. Nguyen, H. B. Ly, L. S. Ho et al., "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Mathematical Problems in Engineering*, vol. 2021, Article ID 4832864, 15 pages, 2021.

[28] D. T. Bui, N. D. Hoang, and V. H. Nhu, "A swarm intelligence-based machine learning approach for predicting soil shear strength for road construction: a case study at Trung Luong National Expressway Project (Vietnam)," *Engineering with Computers*, vol. 35, pp. 955–965, 2019.

[29] V. H. Nhu, N. D. Hoang, V. B. Duong, H. D. Vu, and D. T. Bui, "A hybrid computational intelligence approach for predicting soil shear strength for urban housing construction: a case study at Vinhomes Imperia project, Hai Phong city (Vietnam)," *Engineering with Computers*, vol. 36, pp. 603–616, 2020.

[30] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah et al., "A review on applications of ANN and SVM for building electrical energy consumption forecasting," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 102–109, 2014.

[31] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of Clinical Epidemiology*, vol. 49, pp. 1225–1231, 1996.

[32] D. Boudana, L. Nezli, A. Tlemçani, M. Mahmoudi, and M. Tadjine, "Robust DTC based on adaptive fuzzy control of double star synchronous machine drive with fixed switching frequency," *Journal of Electrical Engineering*, vol. 63, 2012.

[33] E. Benyoussef, A. Meroufel, and S. Barkat, "Three-level DTC based on fuzzy logic and neural network of sensorless DSSM using extended kalman filter," *International Journal of Power Electronics and Drive Systems*, vol. 5, 2015.

[34] L. Auria and R. A. Moro, *Support Vector Machines (SVM) as a Technique for Solvency Analysis*, DIW Discussion Papers, No. 811, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin, Germany, 2008.

[35] P. Sihag, S. M. Karimi, and A. Angelaki, "Random forest, M5P and regression analysis to estimate the field unsaturated hydraulic conductivity," *Applied Water Science*, vol. 9, pp. 1–9, 2019.

[36] W. A. Agyare, S. J. Park, and P. L. G. Vlek, "Artificial neural network estimation of saturated hydraulic conductivity," *Vadose Zone Journal*, vol. 6, pp. 423–431, 2007.

[37] Y. Erzin, S. D. Gumaste, A. K. Gupta, and D. N. Singh, "Artificial neural network (ANN) models for determining hydraulic conductivity of compacted fine-grained soils," *Canadian Geotechnical Journal*, vol. 46, pp. 955–968, 2009.

[38] B. Rogiers, D. Mallants, O. Batelaan, M. Gedeon, M. Huysmans, and A. Dassargues, "Estimation of hydraulic conductivity and its uncertainty from grain-size data using GLUE and artificial neural networks," *Mathematical Geosciences*, vol. 44, pp. 739–763, 2012.

[39] S. K. Das, P. Samui, and A. K. Sabat, "Prediction of field hydraulic conductivity of clay liners using an artificial neural network and support vector machine," *International Journal of Geomechanics*, vol. 12, pp. 606–611, 2012.

[40] P. Sihag, "Prediction of unsaturated hydraulic conductivity using fuzzy logic and artificial neural network," *Modeling Earth Systems and Environment*, vol. 4, pp. 189–198, 2018.

[41] K. I. Wong, C. M. Vong, P. K. Wong, and J. Luo, "Sparse Bayesian extreme learning machine and its application to biofuel engine performance prediction," *Neurocomputing*, vol. 149, pp. 397–404, 2015.

[42] C. Deepa, K. SathiyaKumari, and V. P. Sudha, "Prediction of the compressive strength of high performance concrete mix using tree based modeling," *International Journal of Computer Applications*, vol. 6, pp. 18–24, 2010.

[43] P. Sihag, N. K. Tiwari, and S. Ranjan, "Modelling of infiltration of sandy soil using Gaussian process regression," *Modeling Earth Systems and Environment*, vol. 3, pp. 1091–1100, 2017.

[44] J. Yuan, K. Wang, T. Yu, and M. Fang, "Reliable multi-objective optimization of high-speed WEDM process based on Gaussian process regression," *International Journal of Machine Tools and Manufacture*, vol. 48, pp. 47–60, 2008.

[45] A. Behnood, V. Behnood, M. M. Gharehveran, and K. E. Alyamac, "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm," *Construction and Building Materials*, vol. 142, pp. 199–207, 2017.

[46] N. Karballaeezadeh, H. G. Tehrani, D. M. Shadmehri, and S. Shamshirband, "Estimation of flexible pavement structural capacity using machine learning techniques," *Frontiers of Structural and Civil Engineering*, vol. 14, pp. 1083–1096, 2020.

[47] Y. Ayaz, A. F. Kocamaz, and M. B. Karakoç, "Modeling of compressive strength and UPV of high-volume mineral-admixtured concrete using rule-based M5 rule and tree model M5P classifiers," *Construction and Building Materials*, vol. 94, pp. 235–240, 2015.

[48] M. Pal and S. Deswal, "Modelling pile capacity using Gaussian process regression," *Computers and Geotechnics*, vol. 37, pp. 942–947, 2010.

[49] P. Samui and J. Jagan, "Determination of effective stress parameter of unsaturated soils: a Gaussian process regression approach," *Frontiers of Structural and Civil Engineering*, vol. 7, pp. 133–136, 2013.

[50] S. Mohanty, N. Roy, S. P. Singh, and P. Sihag, "Estimating the strength of stabilized dispersive soil with cement clinker and fly ash," *Geotechnical & Geological Engineering*, vol. 37, pp. 2915–2926, 2019.

[51] P. Sihag, B. Singh, A. S. Vand, and V. Mehdipour, "Modeling the infiltration process with soft computing techniques," *ISH Journal of Hydraulic Engineering*, vol. 26, pp. 138–152, 2020.

[52] P. Sihag, N. K. Tiwari, and S. Ranjan, "Prediction of cumulative infiltration of sandy soil using random forest approach," *Journal of Applied Water Engineering and Research*, vol. 7, pp. 118–142, 2018.

[53] P. Sihag, P. Jain, and M. Kumar, "Modelling of impact of water quality on recharging rate of storm water filter system using various kernel function based regression," *Modeling Earth Systems and Environment*, vol. 4, pp. 61–68, 2018.

[54] M. Rostam, R. Nagamune, and V. Grebenyuk, "A hybrid Gaussian process approach to robust economic model predictive control," *Journal of Process Control*, vol. 92, pp. 149–160, 2020.

[55] A. Pretorius, H. Kamper, and S. Kroon, "On the expected behaviour of noise regularised deep neural networks as Gaussian processes," *Pattern Recognition Letters*, vol. 138, pp. 75–81, 2019.

[56] Y. Wang and B. D. Chaib, "An online Bayesian filtering framework for Gaussian process regression: application to global surface temperature analysis," *Expert Systems with Applications*, vol. 67, pp. 285–295, 2017.

[57] Y. Wang and I. H. Witten, *Induction of Model Trees for Predicting Continuous Classes*, University of Waikato, Hamilton, New Zealand, 1996.

[58] C. E. Rasmussen and K. C. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, MA, USA, 2006.

[59] C. Zhan, A. Gan, and M. Hadi, "Prediction of lane clearance time of freeway incidents using the M5P tree algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, pp. 1549–1557, 2011.

[60] S. Faul, G. Gregorčič, G. Boylan, W. Marnane, G. Lightbody, and S. Connolly, "Gaussian process modeling of EEG for the detection of neonatal seizures," *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 2151–2162, 2007.

[61] R. Venkata Rao and J. Taler, *Advanced Engineering Optimization through Intelligent Techniques: Select Proceedings of AEOTIT*, V. Rao and R. Taler Jan, Eds., Springer, Berlin, Germany, 2018.

[62] V. H. Nhu, H. Shahabi, E. Nohani et al., "Geo-information daily water level prediction of zrebar lake (Iran): a comparison between M5P, random forest, random tree and reduced error pruning trees algorithms," *ISPRS International Journal of Geo-Information*, vol. 9, 2020.

[63] M. Yakhchi, J. Alonso, M. Fazeli, A. A. Bitaraf, and A. Patooqhy, "Neural network based approach for time to crash prediction to cope with software aging," *Journal of Systems Engineering and Electronics*, vol. 26, pp. 407–414, 2015.