sustainability

Article

A Novel Hybrid Soft Computing Model Using Random Forest and Particle Swarm Optimization for Estimation of Undrained Shear Strength of Soil

Binh Thai Pham [1,2], Chongchong Qi [3], Lanh Si Ho [4], Trung Nguyen-Thoi [1,2], Nadhir Al-Ansari [5,*], Manh Duc Nguyen [6], Huu Duy Nguyen [7], Hai-Bang Ly [8,*], Hiep Van Le [9,*] and Indra Prakash [10]

[1] Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh 700000, Vietnam; phamthaibinh@tdtu.edu.vn (B.T.P.); nguyenthoitrung@tdtu.edu.vn (T.N.-T.)
[2] Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh 700000, Vietnam
[3] School of Resources and Safety Engineering, Central South University, Changsha 410083, China; chongchong.qi@gmail.com
[4] Department of Civil and Environmental Engineering, Graduate School of Engineering, Hiroshima University, Hiroshima 739-527, Japan; hosilanh@hiroshima-u.ac.jp
[5] Department of Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 971 87 Lulea, Sweden
[6] University of Transport and Communications, Hanoi 100000, Vietnam; nguyenducmanh@utc.edu.vn
[7] Faculty of Geography, VNU University of Science, Vietnam National University, Hanoi 100000, Vietnam; huuduy151189@gmail.com
[8] University of Transport and Technology, Hanoi 100000, Vietnam
[9] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
[10] Department of Science & Technology, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Government of Gujarat, Gandhinagar 382007, India; indra52prakash@gmail.com
* Correspondence: nadhir.alansari@ltu.se (N.A.-A.); banglh@utt.edu.vn (H.-B.L.); levanhiep2@duytan.edu.vn (H.V.L.)

Received: 8 February 2020; Accepted: 11 March 2020; Published: 12 March 2020

Abstract: Determination of shear strength of soil is very important in civil engineering for foundation design, earth and rock fill dam design, highway and airfield design, stability of slopes and cuts, and in the design of coastal structures. In this study, a novel hybrid soft computing model (RF-PSO) of random forest (RF) and particle swarm optimization (PSO) was developed and used to estimate the undrained shear strength of soil based on the clay content (%), moisture content (%), specific gravity (%), void ratio (%), liquid limit (%), and plastic limit (%). In this study, the experimental results of 127 soil samples from national highway project Hai Phong-Thai Binh of Vietnam were used to generate datasets for training and validating models. Pearson correlation coefficient (R) method was used to evaluate and compare performance of the proposed model with single RF model. The results show that the proposed hybrid model (RF-PSO) achieved a high accuracy performance (R = 0.89) in the prediction of shear strength of soil. Validation of the models also indicated that RF-PSO model (R = 0.89 and Root Mean Square Error (RMSE) = 0.453) is superior to the single RF model without optimization (R = 0.87 and RMSE = 0.48). Thus, the proposed hybrid model (RF-PSO) can be used for accurate estimation of shear strength which can be used for the suitable designing of civil engineering structures.

Keywords: machine learning; random forest; particle swarm optimization; Vietnam

1. Introduction

In civil engineering, the shear strength of the soil is an essential engineering property in the foundation design and stability analysis of all major construction projects such as dams, bridges, highways and road, railway lines, jetties, underground structures, and high-rise buildings [1,2]. It is well-known that the shear strength of soil is governed by interlocking between soil particles, frictional resistance, and cohesion of soil particles. Soil is a complicated material containing soil particles of different sizes and minerals, water, air, and void. The shear strength of soil is influenced by soil constituents, specific gravity, void ratio, moisture content (liquid and plastic limits), clay content, stress history, and relative density. The shear strength parameters are usually determined in the laboratory using direct shear test, unconfined compression test, and triaxial compression test and also in the field by shear vane test, Standard Penetration Test (SPT) and in situ-shear test [3,4]. In addition to these tests, the estimation of the shear strength of soil from other indirect methods is needed for quick and reliable results. Many researchers have attempted to estimate the shear strength of soil using different alternative methods [3,5–10]. The shear strength of unsaturated soil can also be predicted using the empirical correlation function [10,11], using soil-water retention curve [10,12,13].

Nowadays, soft computing techniques of machine learning (ML) or artificial intelligence (AI) have been widely used in many scientific, medical, and engineering fields including geotechnical engineering [14–26]. Sharma et al. [27] used Artificial Neuron Network (ANN) in estimating elasticity modulus of soil. Kalkan et al. [28] used Adaptive Neuro Fuzzy Inference System (ANFIS) and ANN methods for the prediction of compressive strength of compacted granular soils. They concluded that the ANFIS model gave a promising solution for predicting the compressive strength of the compacted soil. Other researchers have used several algorithms of ML namely Support Vector Regression (SVR) and its hybridization with Particle Swarm Optimization (PSO-SVR) [29] using some basic parameters such as water content, clay content, consistency limits, etc. [30]. Pham et al. [31] also developed and used Parsimonious Network based on a Fuzzy Inference System (PANFIS) hybrid ML model in the prediction of soil shear strength.

Random forest (RF) first introduced by Breiman for solving regression, unsupervised learning, and classification problems [32,33], is a powerful ML technique which has been applied successfully in many classification problems including geotechnical engineering [29]. However, for excellent RF modeling, fine-tuning of its hyperparameters is required by optimization algorithms. There are some good optimization algorithms that are usually employed in solving geotechnical engineering problems such as Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Ant Colony Optimization (ACO), Firefly Algorithm (FA), and Artificial Bee Colony (ABC). Out of these algorithms, the PSO is one of the most popular optimization techniques that has been mostly used in geotechnical engineering including slope stability and foundation engineering [27,34–37].

The main objective of the present study is to develop a novel hybrid soft computing model RF-PSO using goodness of individual models, namely RF and PSO, for the quick and better estimation of undrained shear strength of soil, based on the basic soil parameters such as moisture content, clay content, consistency limits, and Atterberg limit. The novelty of this study is that a hybrid model PSO-RF was developed the first time for better estimation of the shear strength of soil from basic parameters. For this, one of the national highway projects, Hai Phong-Thai Binh of Vietnam, was selected as a study area considering its importance and availability of the sufficient soil mechanics data for the development and validation of the model. Pearson correlation coefficient (R) and Root Mean Square Error (RMSE) methods were used to validate the model and sensitivity analysis was carried out to analyze the relationship between shear strength and influencing factors.

2. Case Study and Data Collection

2.1. Description of the Study Area

The study area is located along alignment of Hai Phong-Thai Binh coastal national highway, Vietnam. The total length of the highway project is approximately 29.7 km, which covers 20.782 km towards Hai Phong city side and 8.928 km towards Thai Binh province side (Figure 1). In this route, construction of eight bridges will be implemented. Two bridges on this alignment are large bridges with the length of 2 km and 1 km will cross over Van Uc and Thai Binh river, respectively. A total of 127 soil samples in this study were collected from the construction project of Hai Phong-Thai Binh coastal national highway for the estimation of soil shear strength and model study. The soil investigations carried out in this project included the following field tests: Standard Penetration Test (SPT), boring test, shear vane test, and laboratory tests (direct shear test and triaxial compression test) for the determination of engineering properties of soil [3,4]. However, only data of direct shear test with Undrained and Unconfined (UU) scheme [38] was used for this modeling study.

2.2. Data Used

2.2.1. Output (Undrained Shear Strength of Soil)

Output of this study is the undrained total normal shear strength parameter of soil. Shear strength of soil is defined as the maximum resistance per unit of soil that can mobilize to resist the shear stress causing the sliding failure in any plane of a soil mass. Shear strength of soil is known as an important parameter which is used in the design and analysis of stability problems of civil engineering structures. The sliding failure is related to both normal and shear stress, thus shear strength is considered as a linear function of normal and shear stress [38]. The undrained total normal shear strength (τ_f) is determined using the following equation.

$$\tau_f = c + \sigma \tan \varphi \tag{1}$$

where, c, φ, σ are the cohesion, internal friction angle, and normal stress, respectively.

In this study, direct shear tests with UU scheme was carried out to determine the values of undrained total normal shear strength for the modeling. Initial analysis of data used is presented in Table 1.

Table 1. Initial analysis of data used in this study.

No	Parameters	Min Values	Max Values	Mean Values	Standard Deviation
1	Clay content (%)	1.00	47.5	25.72	10.172
2	Water content (%)	23.04	70.74	48.3	11.73
3	Specific gravity	2.67	2.72	2.69	0.01
4	Void ratio	0.63	1.92	1.36	0.31
5	Liquid limit (%)	26.08	79.76	53.34	13.39
6	Plastic limit (%)	15.36	40.48	28.38	5.01
7	Undrained total normal shear strength (kG/cm^2)	0.29	0.57	0.41	0.06

2.2.2. Input Variables

Input data used in the model study include physical properties of soil: clay content, specific gravity, void ratio, and Atterberg limit (liquid limit and plastic limit). According to Das and Sobhan [38], the size of the clay particles are less than 0.002 mm. Some researchers have considered clay size range to be less than 0.002 to 0.005 mm [31,38]. The amount of clay particles directly affects the Atterberg limits such as liquid and plastic limits. The soil that contains more clay content could result in high plasticity when it absorbs water, thus leading to a decrease in shear strength parameters such as cohesion c and internal friction angle φ. The amount of clay can be determined from the grain size test [39].

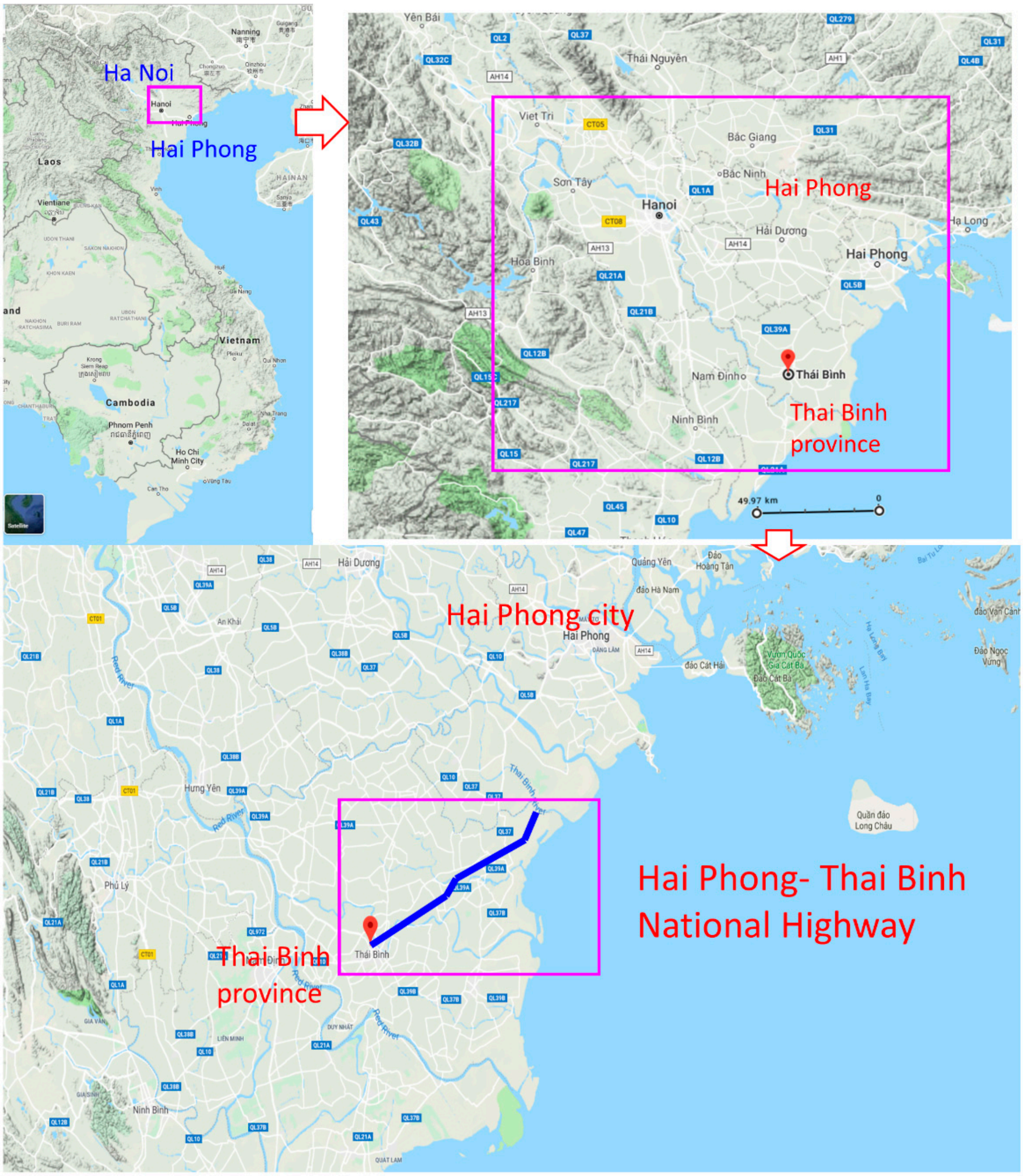

Figure 1. Location map of the study area.

Water content has an important role in influencing many soil properties such as strength, elasticity, and hydraulic conductivity [38]. Water content of the soil is governed by environment condition such as temperature, groundwater level, and humidity. With the increase of water content, cohesion between soil particles reduces; thus, the shear strength also decreases. It is well-known that the shear strength of soil is strongly affected by the moisture change, especially soil containing high clay minerals. Thus, in this research, we considered water content as one of the important input parameters for predicting the shear strength of soil. Water content is defined by the ratio between the mass (weight) of water divided to the mass (weight) of dried soil (i.e., soil particle) in a given volume of soil [39].

Specific gravity is a term that is used to compare how much lighter or heavier solid material (soil particles) is compared to water [38]. It is a ratio between the density of solid phase and water density. It is known that specific gravity is related to the density of minerals of soil: if the specific gravity is large, the soil will be denser with higher shear strength. Thus, the specific gravity must be considered as the main factor in predicting shear strength. In the laboratory, the specific gravity is directly determined by measuring density of soil samples using density bottle and Pycnometer methods [39]. Void ratio of soil not only influences the hydraulic conductivity but also affects significantly the shear strength of the soil. When soil has a large void ratio, the shear strength of soil will be small, because the higher void ratio could have a higher moisture content.

Liquid limit is an important parameter which influences the shear strength. The shear strength decreases with the increase of liquid limit [38]. Liquid limit is determined as the moisture content; at that point soil transfers from the plastic to the liquid state. Plastic limit is one of the important factors affecting the soil strength. It is defined as the water content point at which soil transfers states from semisolid to plastic [38]. These Atterberg limits can be determined by Atterberg tests in laboratory [38].

3. Methods Used

3.1. Random Forest

Random forest (RF) is a powerful method that was first introduced by Breiman for solving regression, unsupervised learning, and classification problems [32,33]. It has been applied in many fields including geotechnical engineering with high performance results [29]. The RF has some advantages such as high accuracy performance with complicated datasets with small calibrating and variables with high noises [40,41]. For the classification problem, Bagging technique is employed in order to arbitrarily choose the variable candidates from the entire dataset for calibrating models [15]. In this research, an Out-Of-Bag (OOB) sample and two kinds of errors (namely decrease in precision and reduction in Gini) were calculated because these errors can be adopted to rank and select variables [42,43]. Regarding each variable, when the variable values are transposed across the OOB observations, the function decides the error of prediction model [44].

3.2. Particle Swarm Optimization (PSO)

The PSO is a computational method which is a form of evolution algorithm such as GA and ant colony algorithm used in the optimization problem, initially proposed by Eberhart and Kennedy [45]. This algorithm differs from the GA as it focuses more on the interaction between individuals in a population to explore search space. The PSO is a result of modeling bird flying to find food; thus, it uses swarm intelligence. This algorithm is a powerful technique that has been employed commonly for optimization problems in many fields, especially in geotechnical engineering [14,31,46]. The PSO works with a random population of particles, in which each particle is considered as a given approach to seek a solution for solving the problem. The PSO includes a group of particles, each particle in a group moves indiscriminately in a research space and is affected by surrounding position during movement [47,48]. The result of each particle position is affected by its knowledge and the knowledge of its neighbors. Thus, it can be said that in a swarm the knowledge of other particles can influence the searching method of a particle. For each iteration, the position of each particle is upgraded considering its current position and velocity [35,49]. The next swarm was established in accordance with the updated particle positions considering their own best position (P_{best}) in search space and the whole swarm best position (G_{best}). The particle position and velocities are computed as follows:

$$V_i^{t+1} = wV_i^t + m_1 n_1 (p_{best,i}^t - Y_i^t) + m_2 n_2 (g_{best,i}^t - Y_i^t) \tag{2}$$

$$Y_i^{t+1} = Y_i^t + V_i^{t+1} \tag{3}$$

where V_i^t and V_i^{t+1} denote velocities of particle *i* at iteration *t* and *t+1*, respectively; whereas, Y_i^t and Y_i^{t+1} represent positions of particle *i* at repetition *t* and *t+1*; w, $m_{1,}$ *and* m_2 correspond to the cognitive, social effect, and inertia parameters, respectively; n_1 and n_2 indicate arbitrary numbers with the range of [0, 1]; $p_{best,i}^t$ and $g_{best,i}^t$ symbolize the best position of particle *i* and swarm, respectively.

The best position of particle and swarm in the following iteration is defined as follows:

$$p_{best,i}^{t+1} = \begin{cases} Y_i^{t+1}, h\left(Y_i^{t+1}\right) < h\left(p_{best,i}^{t}\right) \\ p_{best,i}^{t}, h\left(Y_i^{t+1}\right) \geq h\left(p_{best,i}^{t}\right) \end{cases} \tag{4}$$

$$g_{best}^{t+1} = \mathrm{argmin}\left\{h\left(p_{best,0}^{t+1}\right), \ldots, h\left(p_{best,ns}^{t+1}\right), h\left(g_{best}^{t}\right)\right\} \tag{5}$$

where *ns* indicate the summation of particles in a swarm.

3.3. Dataset Splitting

In the modeling, clay content, moisture content, specific gravity, void ratio, liquid limit, and plastic limit were used as input, whereas the undrained total normal shear strength of soil was used as an output. In other words, there were six inputs and one output for each instance in the dataset. All variables were normalized into (0, 1) range based on their maximum and minimum values.

For supervised learning, the dataset of 127 soil samples needs to be split into two parts. The first part is used for model training and hyperparameter tuning, which is known as the training set. The second part is known as the testing set used for model verification. As the size of the training set will have an important influence on the performance of ML modeling [49], to ascertain the best training set size (TSS) for the preparation of shear strength dataset, the TSS was changed from 30% to 90%. The RF performance was calculated with the default hyperparameters from Scikit-learn [50]. For each TSS, 100 RF models were built and the average performance was calculated to reduce the randomness in random splitting.

It is important to note that the performance on the training set was calculated using 5-fold CV validation. In terms of the performance on the testing set, the whole training set was first used to train the RF model, with which the prediction on the testing set was obtained. In other words, the prediction on the training set was obtained using the RF model trained with part of the training set, compared with the whole training set during the prediction on the testing set. Thus, the prediction on the testing set is usually better than that on the training set if the same prediction method is used [49].

3.4. Modeling and Hyperparameters Tuning

In the current study, the RF is utilized to model the nonlinear relationship from the inputs to the output. In order to obtain the expert performance, the hyperparameters of RF were tuned using the PSO. Five important hyperparameters were tuned as suggested in the literature [51]. Table 2 summarizes the tuned hyperparameters, the definition, and their tuning ranges. For each set of hyperparameters, 10 RF models were built to reduce the randomness of random splitting. The number of particles was set to be 100 in the PSO. Moreover, the maximum iteration, inertia parameter, the cognitive influence parameter, and the social influence parameter were set to be 50, 0.7298, 1.49618, and 1.49618, respectively. All parameter setting in the PSO was determined by trial tests [52–55].

Table 2. Hyperparameters description and their tuning range.

No	Hyperparameters	Explanation	Range
1	Max_depth	The maximum depth of DTs.	1–20
2	Min_samples_split	The minimum number of samples for the split.	2–10
3	Min_samples_leaf	The minimum number of samples at the leaf node.	1–10
4	Max_DT	The maximum number of RT models in the ensemble	1–1000
5	Max_features	The number of features considered during the selection of the best splitting	0.4–1

During the hyperparameters tuning, the training performance from 5-fold CV was used as the fitness function of the PSO. Each set of hyperparameters was represented by a particle in the PSO. With the iteration of PSO, particle positions would be updated to maximize the fitness value and the hyperparameters were optimized accordingly. The optimum hyperparameters were selected after the PSO. Finally, the RF model with the optimum hyperparameters was verified on the testing set.

3.5. RF Model Assessment

The RF model assessment of this study was carried out using Pearson correlation coefficient (R) and Root Mean Square Error (RMSE) criteria, which are the most common indicators for validation and comparison of machine learning models [22,56–58]. Basically, R presents the correlation between the actual and predicted outputs. Values of R range from −1 to 1, and higher absolute values of R close to 1 indicate better prediction accuracy and vice versa. While RMSE measures the average squared difference between actual and predicted outputs [59–62]. Lower value of RMSE indicates better performance of the model. These indicators are expressed in equations as follows [63–67]:

$$R = \frac{\sum_{i=1}^{m}\left(SS_{coi} - \overline{SS_{co}}\right)\left(SS_{aci} - \overline{SS_{ac}}\right)}{\sqrt{\sum_{i=1}^{m}\left(SS_{coi} - \overline{SS_{co}}\right)^2\left(SS_{aci} - \overline{SS_{ac}}\right)^2}} \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(SS_{coi} - SS_{aci}\right)^2} \tag{7}$$

where SS_{coi} and $\overline{SS_{co}}$ denote the output value of the sample *i*th and the output average value of the sample calculated according to the model, respectively; SS_{aci} and $\overline{SS_{ac}}$ indicate the actual value of the sample *i*th and the actual average values, respectively; m is the summation of samples.

4. Results and Discussion

4.1. Influence of Training Set Size (TSS)

From Figure 2, it can be seen that training performance progressively increased with the increase of TSS. Moreover, the standard deviation from the RF models was decreased with increasing TSS. To be more specific, the average *R* value was increased from 0.71 to 0.84 when the TSS was increased from 30% to 90%. Instead, the *SD* was decreased from 0.11 (TSS = 30%) to 0.03 (TSS = 90%). Above results demonstrate that the training performance was improved and became more stable with the increase of TSS.

In terms of the testing performance, it was increased from 0.79 to 0.87 when the TSS was increased from 30% to 80%. After that, the testing performance was decreased to 0.86 when the TSS was further increased to 90%. Standard Deviation (SD) was evidently increased from 0.08 to 0.12 when the TSS was increased to 80% to 90%. The increase of *SD* indicates that the RF prediction was negatively influenced when the TSS surpassed 80%. Since the testing performance represents the generalization capability of ML models, 80% was selected to be the best TSS in this study.

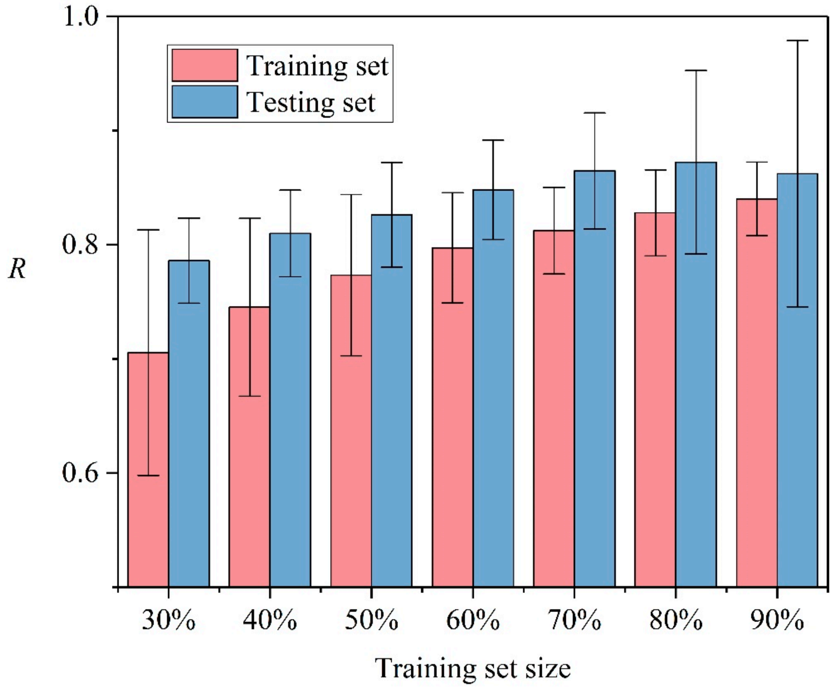

Figure 2. Sensitivity analysis of the model using different training set sizes.

4.2. Hyperparameters Tuning

Figure 3 illustrates the highest *R* value ever found by the PSO with iterations. It can be seen that the highest *R* value was progressively increased with the iteration of PSO. The highest *R* was 0.83 at the first iteration, which was increased to 0.88 at the 50th iteration. The optimum RF hyperparameters were determined to be n_estimator = 935, max_depth = 14, min_sample_split = 2, min_samples_leaf = 1, max_features = 0.648.

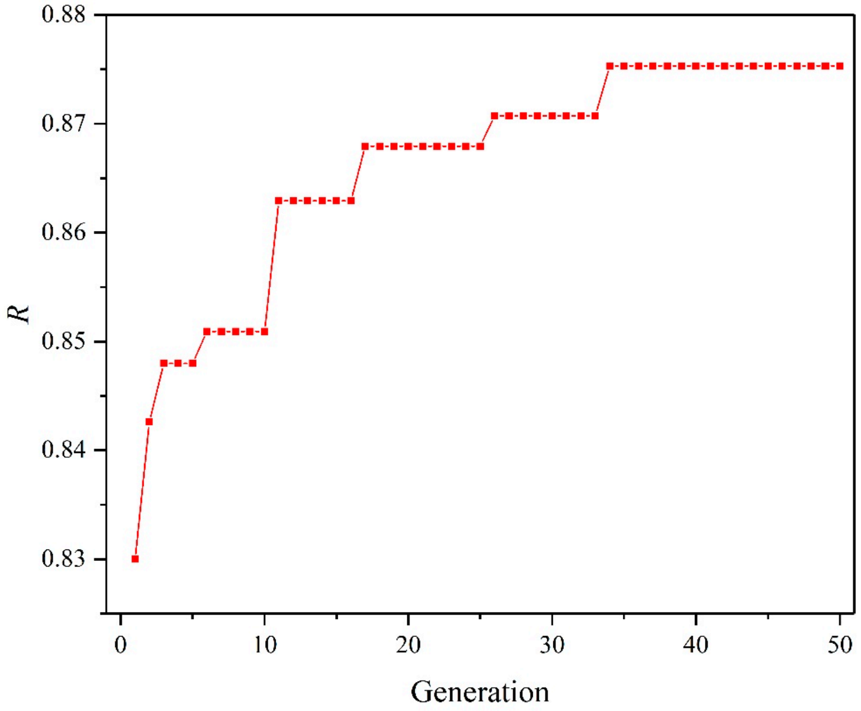

Figure 3. Hyperparameters tuning using the model.

4.3. Predictive Capability of the Models

Figure 4 presents a visual comparison of the UU based experimental and predicted results from a representative RF model. In this case, the representative RF model was selected since its performance was similar to the average performance from the RF models (Figure 3). The *R* value was 0.87 on the training set and 0.90 on the testing set. It can be seen that there was a good agreement between the experimental and predicted shear strength values, implying the robustness of RF modeling.

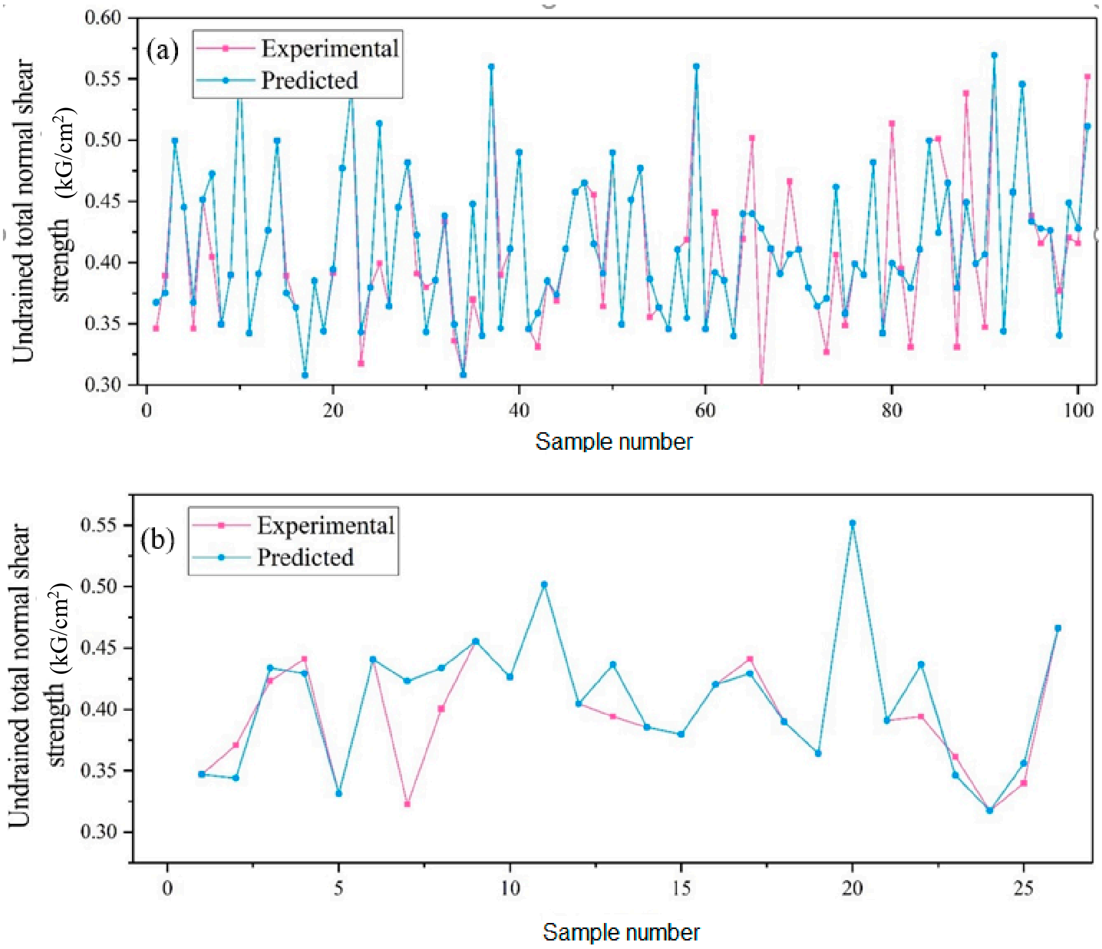

Figure 4. Experimental and predicted values of shear strength using the model: (**a**) training set, (**b**) testing set.

Figure 5 demonstrates performance comparison between the RF model with default hyperparameters and the optimum hyperparameters from the PSO (RF-PSO). To obtain more representative results, 100 RF models were constructed and the average performance was compared. It can be seen that the performance of RF modeling was improved after PSO hyperparameters tuning. On the training set, the *R* value was increased from 0.86 to 0.87 after the PSO hyperparameters tuning. At the same time, *SD* was decreased from 0.028 to 0.027, indicating more stable RF modeling. Similar results can be observed on the testing set, where the *R* value was increased (0.87 to 0.89) and *SD* was decreased (0.060 to 0.057). Similarly, the values of RMSE of RF-PSO on both training (0.487) and testing (0.453) is lower than those of sing RF model (0.517 for training and 0.48 for testing) (Table 3). These results confirmed the feasibility of PSO in improving the performance of RF modeling.

Table 3. Predictive capability of the models using RMSE criteria.

No	Models	RMSE	
		Training	Testing
1	RF	0.517	0.480
2	RF-PSO	0.487	0.453

In general, both hybrid model RF-PSO and single RF model performed well for the prediction of undrained shear strength of soil but hybrid model RF-PSO outperforms single RF model. The results are suitable as the RF can measure data structures and classify data, which help focus important variables and remove similar variables. It is not sensitive to unit differences, pointing out that there is no need for a preprocessing process [68]. In addition, the PSO is effectively applied to address the problem of complex optimization as it is automatically to search for optimization solutions and it can easily perform with good efficiency [69]. Thus, it is confirmed that the PSO is an effective optimization technique in improving performance of the RF model.

In general, the proposed RF-PSO model can be used for quick, better prediction of the undrained total normal shear strength of different types of soil. Performance of this model might be different and improved depending on type of soil. One of the advantages of application of hybrid machine learning model (RF-PSO) is that it can handle the big and complicated data. Thus, large or big data can also be analyzed with this model. It is proposed that researchers use large data, depending on the availability, for future model studies.

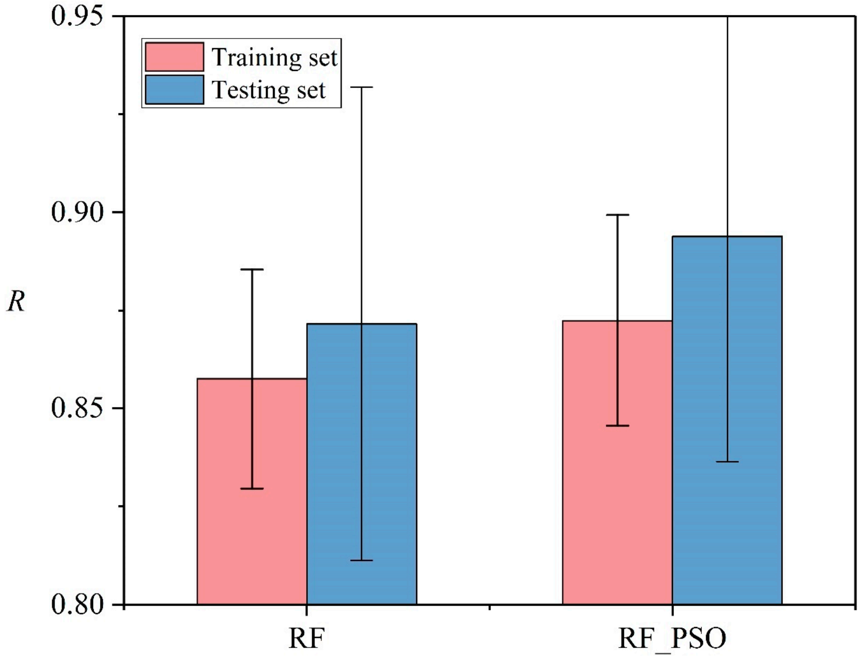

Figure 5. Predictive capability of the models.

4.4. Sensitivity Analysis of Input Parameters

A sensitivity analysis was carried out to evaluate the importance of input parameters for the modeling using partial dependence plots, which is an efficient way to investigate the relationship between inputs and output. Details are available in the published work [70]. Figure 6 shows the relative importance of the inputs to the output. It can be seen that the moisture content was the most significant variable for the shear strength of soil, which achieved an average importance score of 0.337. The void ratio ranked the second with an average importance score of 0.271, followed by liquid limit (0.166), clay (0.109), and plastic limit (0.093). The specific gravity achieved the smallest average importance score (0.025), indicating that it had the lowest influence on the undrained total normal shear strength of soil.

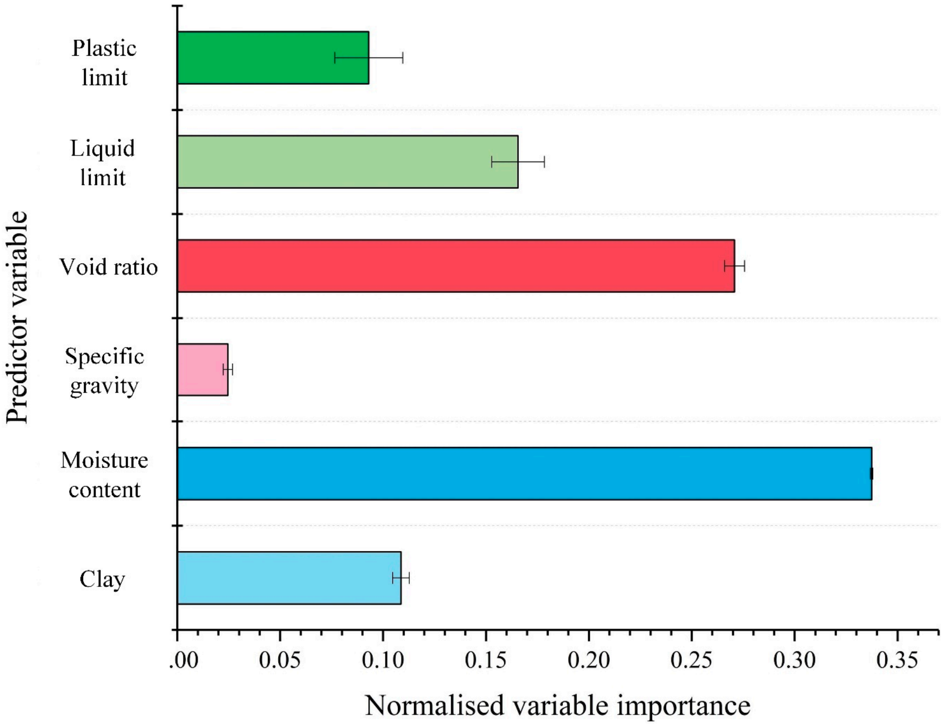

Figure 6. Variable importance analysis using the model.

Figure 7 illustrates the partial dependence of the output to the inputs. As shown, the shear strength had an overall negative correlation with clay, moisture content, specific gravity, and void ratio. Moreover, the variation of shear strength was less significant with the variation of specific gravity compared with the variation of clay, moisture content, and void ratio. This result indicates the specific gravity has a relatively lower influence on the undrained shear strength, which agrees well with the important score results (Figure 6). The undrained shear strength decreased first and then increased with the increasing liquid limit. Finally, the undrained shear strength increased with the increase of plastic limit.

In general, out of the input factors, moisture content is considered as the most important factor affecting the undrained shear strength of soil. This is reasonable as the water reduces the friction and cohesion between the soil particles; thus, increase of moisture content leads to decrease of shear strength of soil [31,71].

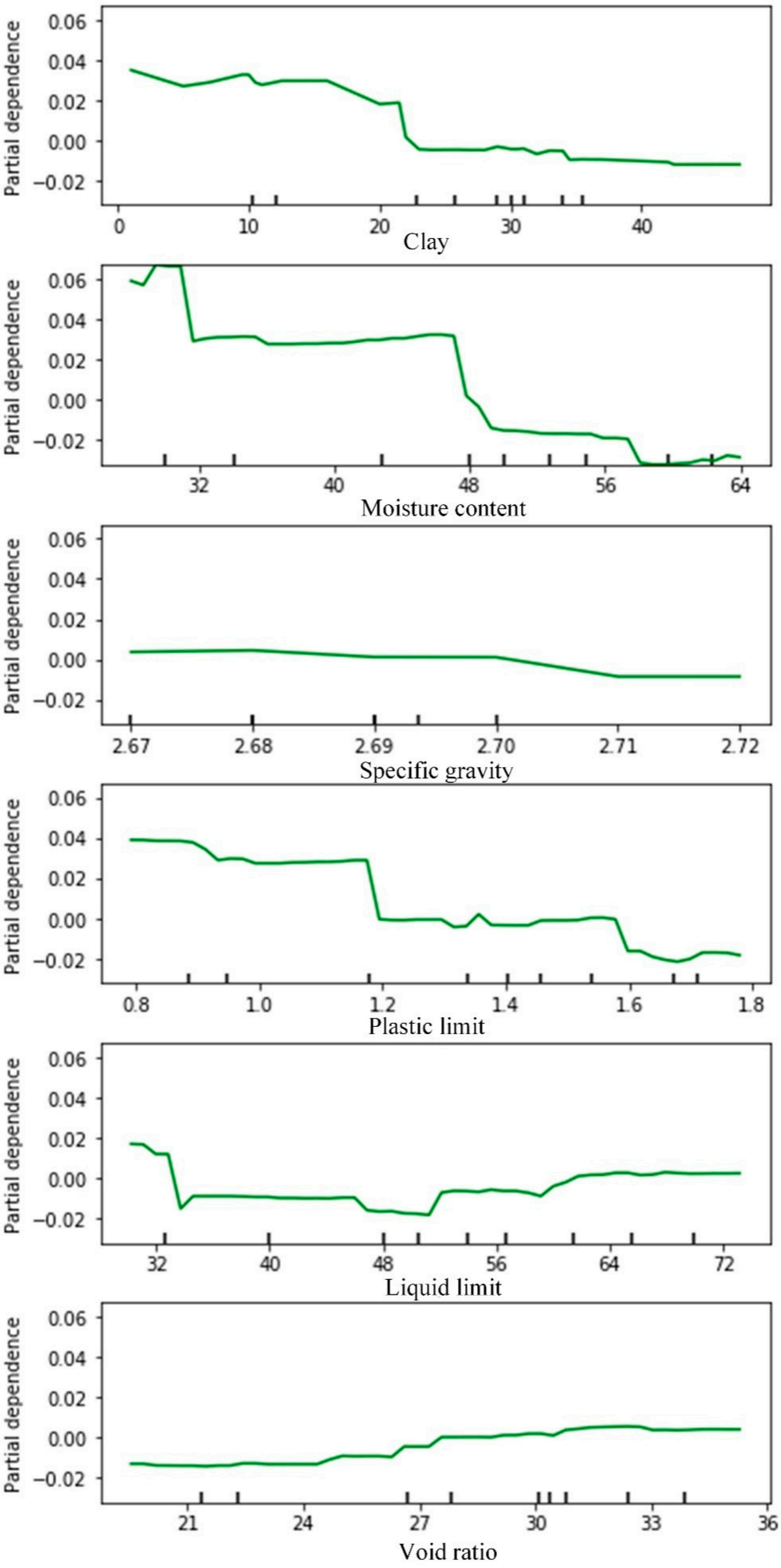

Figure 7. Importance of the variables used for the model.

5. Conclusions

In this study, a novel hybrid machine learning namely RF-PSO model, which is a combination of RF and PSO models, was proposed and applied to estimate the undrained shear strength of soil for the design purpose of the construction projects. In total, the experimental results of 127 samples were used to create datasets for validating and training models. A statistical measure such as R was used to validate and compare the models. The results show that performance of the models improved and stabilized from 0.79 to 0.84 with the increase of training dataset size from 30% to 80%. Performance of the RF-PSO hybrid model is best with R = 0.89 and RMSE = 0.453, followed by RF with R = 0.87 and RMSE = 0.48 in the estimation of soil shear strength.

In addition, the sensitivity analysis using partial dependence plots was carried out to evaluate the importance of input parameters in the model study. Results show that moisture content is considered as the most important parameter for modeling of prediction of the undrained shear strength though other parameters considered are also important.

In this study it is seen that that the proposed hybrid model RF-PSO is capable of predicting shear strength of undrained soil quickly with basic soil properties in a better way for use in the design of civil engineering structures. A limitation of this study is the number of samples tested from one of the highway projects of Vietnam. It would be better to use large/big data from other projects to confirm its wider applicability.

Author Contributions: Conceptualization, B.T.P., T.N.-T., N.A.-A., M.D.N., and I.P.; Data curation, M.D.N., H.D.N. and H.V.L.; Formal analysis, L.S.H., T.N.-T., N.A.-A., M.D.N., H.D.N., H.-B.L. and H.V.L.; Methodology, B.T.P., C.Q., N.A.-A. and H.-B.L.; Project administration, B.T.P. and T.N.-T.; Supervision, T.N.-T., N.A.-A. and I.P.; Validation, T.N.-T., N.A.-A., C.Q., and I.P.; Visualization, L.S.H. and C.Q.; Writing—original draft, all authors; Writing—review & editing, B.T.P., T.N.-T., N.A.-A., and I.P.; Funding: T.N.-T. and N.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Poulos, H.G. Design of reinforcing piles to increase slope stability. *Can. Geotech. J.* **1995**, *32*, 808–818. [CrossRef]
2. Liu, Y.-J.; Wang, T.-W.; Cai, C.-F.; Li, Z.-X.; Cheng, D.-B. Effects of vegetation on runoff generation, sediment yield and soil shear strength on road-side slopes under a simulation rainfall test in the Three Gorges Reservoir Area, China. *Sci. Total Environ.* **2014**, *485*, 93–102. [CrossRef] [PubMed]
3. Hettiarachchi, H.; Brown, T. Use of SPT blow counts to estimate shear strength properties of soils: Energy balance approach. *J. Geotech. Geoenviron. Eng.* **2009**, *135*, 830–834. [CrossRef]
4. Motaghedi, H.; Eslami, A. Analytical approach for determination of soil shear strength parameters from CPT and CPTu data. *Arab. J. Sci. Eng.* **2014**, *39*, 4363–4376. [CrossRef]
5. Cha, M.; Cho, G.-C. Shear strength estimation of sandy soils using shear wave velocity. *Geotech. Test. J.* **2007**, *30*, 484–495.
6. Garven, E.; Vanapalli, S. Evaluation of empirical procedures for predicting the shear strength of unsaturated soils. In *Unsaturated Soils 2006, Fourth International Conference on Unsaturated Soils, Carefree, AZ, USA, 2–6 April 2006*; American Society of Civil Engineers: Reston, VA, USA, 2006; pp. 2570–2592.
7. Kim, B.-S.; Shibuya, S.; Park, S.-W.; Kato, S. Application of suction stress for estimating unsaturated shear strength of soils using direct shear testing under low confining pressure. *Can. Geotech. J.* **2010**, *47*, 955–970. [CrossRef]
8. Ohu, J.O.; Raghavan, G.; McKyes, E.; Mehuys, G. Shear strength prediction of compacted soils with varying added organic matter contents. *Trans. ASAE* **1986**, *29*, 351–355. [CrossRef]
9. Tiwari, B.; Marui, H. A new method for the correlation of residual shear strength of the soil with mineralogical composition. *J. Geotech. Geoenviron. Eng.* **2005**, *131*, 1139–1150. [CrossRef]
10. Vilar, O.M. A simplified procedure to estimate the shear strength envelope of unsaturated soils. *Can. Geotech. J.* **2006**, *43*, 1088–1095. [CrossRef]

11. Huang, B.; Qiu, M.; Lin, J.; Chen, J.; Jiang, F.; Wang, M.-K.; Ge, H.; Huang, Y. Correlation between shear strength and soil physicochemical properties of different weathering profiles of the non-eroded and collapsing gully soils in southern China. *J. Soils Sediments* **2019**, *19*, 3832–3846. [CrossRef]
12. Zhai, Q.; Rahardjo, H.; Satyanaga, A.; Dai, G. Estimation of unsaturated shear strength from soil–water characteristic curve. *Acta Geotech.* **2019**, *14*, 1977–1990. [CrossRef]
13. Leong, E.-C. Soil-water characteristic curves-Determination, estimation and application. *Jpn. Geotech. Soc. Spec. Publ.* **2019**, *7*, 21–30. [CrossRef]
14. Bui, D.T.; Nhu, V.-H.; Hoang, N.-D. Prediction of soil compression coefficient for urban housing project using novel integration machine learning approach of swarm intelligence and multi-layer perceptron neural network. *Adv. Eng. Inform.* **2018**, *38*, 593–604.
15. Chen, W.; Wang, Y.; Cao, G.; Chen, G.; Gu, Q. A random forest model based classification scheme for neonatal amplitude-integrated EEG. *Biomed. Eng. Online* **2014**, *13*, S4. [CrossRef]
16. Chou, J.-S.; Pham, A.-D. Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. *Constr. Build. Mater.* **2013**, *49*, 554–563. [CrossRef]
17. Koopialipoor, M.; Fallah, A.; Armaghani, D.J.; Azizi, A.; Mohamad, E.T. Three hybrid intelligent models in estimating flyrock distance resulting from blasting. *Eng. Comput.* **2019**, *35*, 243–256. [CrossRef]
18. Koopialipoor, M.; Ghaleini, E.N.; Tootoonchi, H.; Armaghani, D.J.; Haghighi, M.; Hedayat, A. Developing a new intelligent technique to predict overbreak in tunnels using an artificial bee colony-based ANN. *Environ. Earth Sci.* **2019**, *78*, 165. [CrossRef]
19. Pham, B.T.; Bui, D.T.; Prakash, I.; Dholakia, M. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *Catena* **2017**, *149*, 52–63. [CrossRef]
20. Samui, P. Prediction of friction capacity of driven piles in clay using the support vector machine. *Can. Geotech. J.* **2008**, *45*, 288–295. [CrossRef]
21. Shahin, M.A.; Jaksa, M.B.; Maier, H.R. Recent advances and future challenges for artificial neural systems in geotechnical engineering applications. *Adv. Artif. Neural Syst.* **2009**, *2009*, 5. [CrossRef]
22. Dao, D.V.; Adeli, H.; Ly, H.-B.; Le, L.M.; Le, V.M.; Le, T.-T.; Pham, B.T. A Sensitivity and Robustness Analysis of GPR and ANN for High-Performance Concrete Compressive Strength Prediction Using a Monte Carlo Simulation. *Sustainability* **2020**, *12*, 830. [CrossRef]
23. Pham, B.T.; Avand, M.; Janizadeh, S.; Phong, T.V.; Al-Ansari, N.; Ho, L.S.; Das, S.; Le, H.V.; Amini, A.; Bozchaloei, S.K. GIS Based Hybrid Computational Approaches for Flash Flood Susceptibility Assessment. *Water* **2020**, *12*, 683. [CrossRef]
24. Pham, B.T.; Prakash, I.; Dou, J.; Singh, S.K.; Trinh, P.T.; Tran, H.T.; Le, T.M.; Van Phong, T.; Khoi, D.K.; Shirzadi, A. A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. *Geocarto Int.* **2019**, 1–25. [CrossRef]
25. Pham, B.T.; Bui, D.T.; Prakash, I.; Nguyen, L.H.; Dholakia, M. A comparative study of sequential minimal optimization-based support vector machines, vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS. *Environ. Earth Sci.* **2017**, *76*, 371. [CrossRef]
26. Pham, B.T.; Bui, D.T.; Pham, H.V.; Le, H.Q.; Prakash, I.; Dholakia, M. Landslide hazard assessment using random subspace fuzzy rules based classifier ensemble and probability analysis of rainfall data: A case study at Mu Cang Chai District, Yen Bai Province (Viet Nam). *J. Indian Soc. Remote Sens.* **2017**, *45*, 673–683. [CrossRef]
27. Sharma, L.; Singh, R.; Umrao, R.; Sharma, K.; Singh, T. Evaluating the modulus of elasticity of soil using soft computing system. *Eng. Comput.* **2017**, *33*, 497–507. [CrossRef]
28. Kalkan, E.; Akbulut, S.; Tortum, A.; Celik, S. Prediction of the unconfined compressive strength of compacted granular soils by using inference systems. *Environ. Geol.* **2009**, *58*, 1429–1440. [CrossRef]
29. Nhu, V.H.; Hoang, N.D.; Duong, V.B.; Vu, H.D.; Bui, D.T. 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). *Eng. Comput.* **2019**, 1–14. [CrossRef]
30. Moavenian, M.; Nazem, M.; Carter, J.; Randolph, M. Numerical analysis of penetrometers free-falling into soil with shear strength increasing linearly with depth. *Comput. Geotech.* **2016**, *72*, 57–66. [CrossRef]
31. Pham, B.T.; Hoang, T.-A.; Nguyen, D.-M.; Bui, D.T. Prediction of shear strength of soft soil using machine learning methods. *Catena* **2018**, *166*, 181–191. [CrossRef]

32. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
33. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
34. Jahed Armaghani, D.; Hajihassani, M.; Yazdani Bejarbaneh, B.; Marto, A.; Tonnizam Mohamad, E. Indirect measure of shale shear strength parameters by means of rock index tests through an optimized artificial neural network. *Measurement* **2014**, *55*, 487–498. [CrossRef]
35. Hajihassani, M.; Armaghani, D.J.; Kalatehjari, R. Applications of particle swarm optimization in geotechnical engineering: a comprehensive review. *Geotech. Geol. Eng.* **2018**, *36*, 705–722. [CrossRef]
36. Hasanipanah, M.; Noorian-Bidgoli, M.; Armaghani, D.J.; Khamesi, H. Feasibility of PSO-ANN model for predicting surface settlement caused by tunneling. *Eng. Comput.* **2016**, *32*, 705–715. [CrossRef]
37. Kalatehjari, R.; Ali, N.; Kholghifard, M.; Hajihassani, M. The effects of method of generating circular slip surfaces on determining the critical slip surface by particle swarm optimization. *Arab. J. Geosci.* **2014**, *7*, 1529–1539. [CrossRef]
38. Das, B.M.; Sobhan, K. *Principles of Geotechnical Engineering*; Cengage Learning: Stamford, CT, USA, 2013.
39. Terzaghi, K.; Peck, R.B.; Mesri, G. *Soil Mechanics*; John Wiley & Sons: New York, NY, USA, 1996.
40. Hong, H.; Pourghasemi, H.R.; Pourtaghi, Z.S. Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology* **2016**, *259*, 105–118. [CrossRef]
41. Stumpf, A.; Kerle, N. Object-oriented mapping of landslides using Random Forests. *Remote Sens. Environ.* **2011**, *115*, 2564–2577. [CrossRef]
42. Archer, K.J.; Kimes, R.V. Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* **2008**, *52*, 2249–2260. [CrossRef]
43. Biau, G.; Devroye, L.; Lugosi, G. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **2008**, *9*, 2015–2033.
44. Trigila, A.; Iadanza, C.; Esposito, C.; Scarascia-Mugnozza, G. Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology* **2015**, *249*, 119–136. [CrossRef]
45. Eberhart, R.; Kennedy, J. A new optimizer using particle swarm theory. In Proceedings of the MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 4–6 October 1995; pp. 39–43.
46. Cheng, Y.; Li, L.; Chi, S.-C.; Wei, W. Particle swarm optimization algorithm for the location of the critical non-circular failure surface in two-dimensional slope stability analysis. *Comput. Geotech.* **2007**, *34*, 92–103. [CrossRef]
47. Awad, Z.K.; Aravinthan, T.; Zhuge, Y.; Gonzalez, F. A review of optimization techniques used in the design of fibre composite structures for civil engineering applications. *Mater. Des.* **2012**, *33*, 534–544. [CrossRef]
48. Chen, W.; Panahi, M.; Pourghasemi, H.R. Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modelling. *Catena* **2017**, *157*, 310–324. [CrossRef]
49. Qi, C.; Fourie, A.; Chen, Q.; Zhang, Q. A strength prediction model using artificial intelligence for recycling waste tailings as cemented paste backfill. *J. Clean. Prod.* **2018**, *183*, 566–578. [CrossRef]
50. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
51. Qi, C.; Chen, Q.; Fourie, A.; Zhang, Q. An intelligent modelling framework for mechanical properties of cemented paste backfill. *Miner. Eng.* **2018**, *123*, 16–27. [CrossRef]
52. Eberhart, R.C.; Shi, Y. Comparing inertia weights and constriction factors in particle swarm optimization. In Proceedings of the 2000 Congress on Evolutionary Computation, CEC00 (Cat. No.00TH8512), La Jolla, CA, USA, 16–19 July 2000; Volume 81, pp. 84–88.
53. Van den Bergh, F.; Engelbrecht, A.P. A study of particle swarm optimization particle trajectories. *Inf. Sci.* **2006**, *176*, 937–971. [CrossRef]
54. Li-ping, Z.; Huan-jun, Y.; Shang-xu, H. Optimal choice of parameters for particle swarm optimization. *J. Zhejiang Univ. Sci. A* **2005**, *6*, 528–534. [CrossRef]
55. Qi, C.; Fourie, A.; Chen, Q.; Tang, X.; Zhang, Q.; Gao, R. Data-driven modelling of the flocculation process on mineral processing tailings treatment. *J. Clean. Prod.* **2018**, *196*, 505–516. [CrossRef]

56. Qi, C.; Ly, H.-B.; Chen, Q.; Le, T.-T.; Le, V.M.; Pham, B.T. Flocculation-dewatering prediction of fine mineral tailings using a hybrid machine learning approach. *Chemosphere* **2020**, *244*, 125450. [CrossRef] [PubMed]
57. Pham, B.T.; Le, L.M.; Le, T.-T.; Bui, K.-T.T.; Le, V.M.; Ly, H.-B.; Prakash, I. Development of advanced artificial intelligence models for daily rainfall prediction. *Atmos. Res.* **2020**, *237*, 104845. [CrossRef]
58. Dao, D.V.; Ly, H.-B.; Vu, H.-L.T.; Le, T.-T.; Pham, B.T. Investigation and Optimization of the C-ANN Structure in Predicting the Compressive Strength of Foamed Concrete. *Materials* **2020**, *13*, 1072. [CrossRef] [PubMed]
59. Van Dao, D.; Jaafari, A.; Bayat, M.; Mafi-Gholami, D.; Qi, C.; Moayedi, H.; Van Phong, T.; Ly, H.-B.; Le, T.-T.; Trinh, P.T. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *Catena* **2020**, *188*, 104451.
60. Pham, B.T.; Phong, T.V.; Nguyen, H.D.; Qi, C.; Al-Ansari, N.; Amini, A.; Ho, L.S.; Tuyen, T.T.; Yen, H.P.H.; Ly, H.-B. A Comparative Study of Kernel Logistic Regression, Radial Basis Function Classifier, Multinomial Naïve Bayes, and Logistic Model Tree for Flash Flood Susceptibility Mapping. *Water* **2020**, *12*, 239. [CrossRef]
61. Nguyen, V.V.; Pham, B.T.; Vu, B.T.; Prakash, I.; Jha, S.; Shahabi, H.; Shirzadi, A.; Ba, D.N.; Kumar, R.; Chatterjee, J.M. Hybrid machine learning approaches for landslide susceptibility modeling. *Forests* **2019**, *10*, 157. [CrossRef]
62. Nguyen, M.D.; Pham, B.T.; Tuyen, T.T.; Yen, H.; Phan, H.; Prakash, I.; Vu, T.T.; Chapi, K.; Shirzadi, A.; Shahabi, H. Development of an Artificial Intelligence Approach for Prediction of Consolidation Coefficient of Soft Soil: A Sensitivity Analysis. *Open Constr. Build. Technol. J.* **2019**, *13*, 178–188. [CrossRef]
63. Dao, D.V.; Ly, H.-B.; Trinh, S.H.; Le, T.-T.; Pham, B.T. Artificial intelligence approaches for prediction of compressive strength of geopolymer concrete. *Materials* **2019**, *12*, 983. [CrossRef]
64. Dao, D.V.; Trinh, S.H.; Ly, H.-B.; Pham, B.T. Prediction of compressive strength of geopolymer concrete using entirely steel slag aggregates: Novel hybrid artificial intelligence approaches. *Appl. Sci.* **2019**, *9*, 1113. [CrossRef]
65. Pham, B.T.; Nguyen, M.D.; Van Dao, D.; Prakash, I.; Ly, H.-B.; Le, T.-T.; Ho, L.S.; Nguyen, K.T.; Ngo, T.Q.; Hoang, V. Development of artificial intelligence models for the prediction of Compression Coefficient of soil: An application of Monte Carlo sensitivity analysis. *Sci. Total Environ.* **2019**, *679*, 172–184. [CrossRef]
66. Nguyen, H.-L.; Pham, B.T.; Son, L.H.; Thang, N.T.; Ly, H.-B.; Le, T.-T.; Ho, L.S.; Le, T.-H.; Tien Bui, D. Adaptive network based fuzzy inference system with meta-heuristic optimizations for international roughness index prediction. *Appl. Sci.* **2019**, *9*, 4715. [CrossRef]
67. Janizadeh, S.; Avand, M.; Jaafari, A.; Phong, T.V.; Bayat, M.; Ahmadisharaf, E.; Prakash, I.; Pham, B.T.; Lee, S. Prediction Success of Machine Learning Methods for Flash Flood Susceptibility Mapping in the Tafresh Watershed, Iran. *Sustainability* **2019**, *11*, 5426. [CrossRef]
68. Kohestani, V.; Hassanlourad, M.; Ardakani, A. Evaluation of liquefaction potential based on CPT data using random forest. *Nat. Hazards* **2015**, *79*, 1079–1089. [CrossRef]
69. Wan, S. Entropy-based particle swarm optimization with clustering analysis on landslide susceptibility mapping. *Environ. Earth Sci.* **2012**, *68*. [CrossRef]
70. Qi, C.; Fourie, A.; Chen, Q. Neural network and particle swarm optimization for predicting the unconfined compressive strength of cemented paste backfill. *Constr. Build. Mater.* **2018**, *159*, 473–478. [CrossRef]
71. Pham, B.T.; Nguyen, M.D.; Bui, K.-T.T.; Prakash, I.; Chapi, K.; Bui, D.T. A novel artificial intelligence approach based on Multi-layer Perceptron Neural Network and Biogeography-based Optimization for predicting coefficient of consolidation of soil. *Catena* **2019**, *173*, 302–311. [CrossRef]