Estimation of Shear Strength Parameters from Easily-Collected Soil Physical Properties Using Bagging Learning Technique

Chau-Hoang-Quyen NGUYEN[1,2], The-Khang NGUYEN[1,2], Thi-Hoai-Thuong VO[1,2], Ba-Quang-Vinh NGUYEN*[1,2]

[1]*School of Civil Engineering and Management, International University, Quarter 6, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University, Ho Chi Minh city, Vietnam; E-mail: nbqvinh@hcmiu.edu.vn*

Abstracts: Shear strength parameters, including cohesion and friction angle, are among the most crucial factors in soil mechanics, playing a pivotal role in the design and construction of engineering projects. This paper aims to estimate these essential soil shear strength parameters using an ensemble learning model. To achieve this, the current study employs the Random Forest (RF) model incorporating various physical parameters of soil, such as density (□), saturation degree (S_r), liquid limit (*LL*), silt content (*SC*), clay content (*CC*) to predict cohesion (*c*), and friction angle (□). In order to assess the predictive performance of the used model, this research used various metrics, including the mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (R^2), to evaluate the model's accuracy. The results reveal that RF performs superior predictive capabilities. Furthermore, the proposed model prediction ability was compared to the previous empirical equations. The comparison results indicated that the prediction capability of RF outperforms the previously developed equations.

Keywords: Cohesion, Ensemble learning, Friction angle, Random Forest

1. INTRODUCTION

The Shear strength (SS) of soil refers to the inherent capacity of soil to resist shear forces without undergoing failure. It is a critical parameter used to assess soil mechanical behavior, which, in turn, informs the design and construction of geotechnical structures (Chitra & Gupta, 2014). In accordance with the Mohr-Coulomb Failure Criterion, shear strength is determined by two key parameters: cohesion (*c*), which represents the inherent bonding strength between soil particles, and the internal friction angle (*φ*), signifying the maximum angle between the horizontal plane and an inclined plane where soil particles maintain their equilibrium. The SS parameters can be received through methods conducted either in the field or in the laboratory. In the laboratory, two commonly employed tests are the Direct Shear Test, and Triaxial Compression Test. The in-situ tests include Standard Penetration Test (SPT), Cone Penetration Test (CPT), Pressuremeter Test, Vane Shear Test. However, the process of measuring the SS parameters, whether in the field or laboratory, is inherently costly, time-consuming, and labor-intensive (Khanlari et al., 2012; Mohammadi et al., 2022). Furthermore, obtaining precise undisturbed soil samples from the field poses significant challenges, owing to issues such as the handling of samples, transportation, release of overburden pressure, and maintaining ideal laboratory conditions. This necessitates vigilant supervision and care throughout the entire process (Yoseph, 2022). Hence, certain researchers have put forth various models for the estimation of the SS parameters based on multiple physical properties, including soil type, grain size distribution, density, water content, Atterberg limits, void ratio, saturation, etc…Many empirical equations have been developed for estimating the SS parameters using multiple linear and non-linear regression methods (Adunoye, 2014a; Adunoye, 2014b; Ersoy et al., 2013; Roy et al., 2019). However, determining the appropriate adjustment coefficients for these formulas can be a challenging task (Salari et al., 2015; Zhu et al., 2022). This challenge significantly diminishes their predictive accuracy (Stefanow & Dudziński, 2021; Zhu et al., 2022). With the advent and widespread application of machine learning (ML) in addressing engineering challenges, it has become a valuable tool for handling big data and complex conditions. Numerous studies have emerged utilizing ML techniques to predict shear strength of soil, employing a range of algorithms, most notably Artificial Neural Networks (ANN) (Zhu et al., 2022; Chao et al., 2021; Pham et al., 2018; Iyeke et al., 2016; Khanlari et al., 2012; Mohammadi et al., 2022; Zakharov et al., 2022). Besides the ANN model, Support Vector Machine (SVM) is also a popularly used technique for estimating the shear strength of soil (Zhu et al., 2022; Chao et al., 2021; Pham et al., 2018). Random Forest (RF), a bagging ensemble learning model, stands out as a robust choice and finds

extensive utility in addressing geotechnical engineering issues (Nguyen & Kim, 2021a; Cheng et al., 2021; Nguyen et al., 2022). Nonetheless, Random Forest has not seen widespread utilization in the prediction of shear strength of soil. Furthermore, the development of a digital data system of engineering geological in the context of Vietnam has not garnered significant attention. From the above analysis, this research focuses on the development of a framework of Random Forest aimed at predicting shear strength parameters (cohesion, and internal friction angle) based on soil physical properties. The primary objective is to provide a simple and accurate means of estimating these parameters for a variety of soil types in Dong Nai province, Vietnam. The performance of the employed model was assessed using a range of metrics, encompassing the mean absolute error (MAE), root mean square error (RMSE), and the correlation coefficient (R^2). Furthermore, the impact of various physical properties on the predictive capabilities of the used model is assessed using Decision Trees Feature Importance (DTFI).

2. STUDY AREA AND DATA COLLECTION

Dong Nai is located in the Southeast region of Vietnam, next to Ho Chi Minh City. The database is established upon the compilation of borehole data gathered for geotechnical investigations conducted in Dong Nai province. Within the survey area, a representative soil sample is obtained from each layer in the borehole. These prototypes serve as the basis for conducting experiments to determine the physical, and mechanical properties of soil. The soil samples undergo testing utilizing tools and methods in accordance with ASTM standards. For each soil sample, each property is tested twice in parallel, ensuring that the results do not deviate beyond the permissible margin of error.

A total of 332 samples collected from 40 boreholes in Dong Nai province were used to determine the properties of soil in the laboratory. The observed data including cohesion, and friction angle of soil were obtained from the direct shear test following ASTM. The physical properties, including water content (w), density (□), void ratio (e), saturation degree (S_r), plastic limit (PL), liquid limit (LL), silt content (SC), and clay content (CC) have been considered as features data for training the RF model.

Table 1 presents the initial statistical analysis of the dataset, detailing the units for each variable. This table provides comprehensive statistical information, such as the mean, standard deviation, and quantiles for all the variables.

Table 1. Statistical descriptions of the input features

Statistical descriptions	□(g/cm^3)	S_r	LL	SC	CC	□□□degree)	c (kPa)
Mean	2.717	0.877	0.421	0.225	0.292	14.986	24.982
Standard Deviation	0.062	0.083	0.129	0.115	0.136	4.863	8.660
Sample Variance	0.004	0.007	0.017	0.013	0.019	23.653	74.988
Minimum	2.590	0.195	0.045	0.030	0.027	1.833	2.157
Maximum	3.020	1.000	0.867	0.552	1.870	26.217	41.678
Sum	902.192	291.233	139.757	74.773	96.858	4975.480	8294.007

□: density, S_r: degree of saturation, LL: liquid limit, SC: silt content, CC: clay content, □: friction angle, c: cohesion.

3. METHODS

The proposed methodology involved five distinct steps, as depicted in Fig. 1. These steps are shown below:

(i) Collecting, and analyzing databases.
(ii) Assessing correlations among the features through Pearson and multicollinearity tests.
(iii) Creating training (80% of the database), and testing (20% of the database) datasets.
(iv) Employing the Random Forest model, which predicted the SS parameters.
(v) Assessment and comparative analysis of model performance through the numerous metrics: MAE, RMSE, R^2.

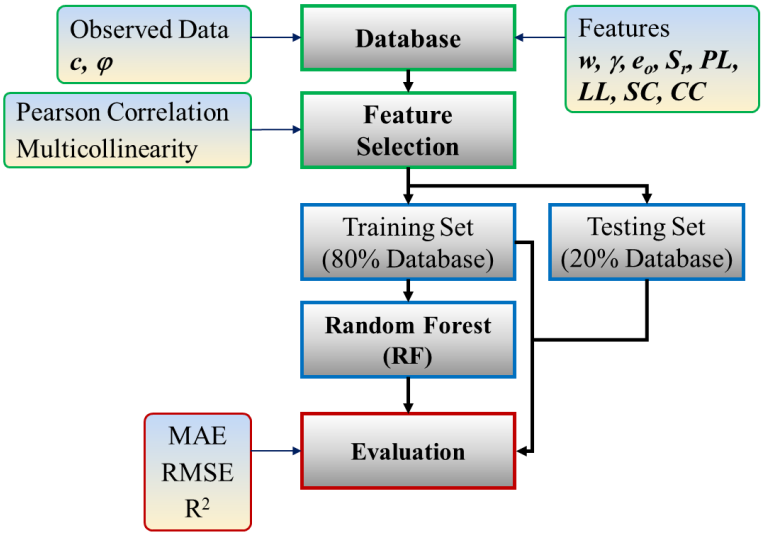

Figure 1. The flow chart of the proposed method

Pearson correlation

The Pearson correlation coefficient is the prevalent method for quantifying the linear correlation of the features in the ML training process (Amin Benbouras & Petrisor, 2021; Mohammadi et al., 2022). This coefficient, ranging from -1 to 1, assesses both the intensity and direction of the association between two variables. This coefficient near 1 indicates a robust positive correlation, while a coefficient near -1 signifies a strong negative correlation. Conversely, coefficients near zero suggest the absence of a linear correlation.

Multicolinearity tests

The interconnections among independent variables hold significant importance in data analysis, profoundly impacting the accuracy of the models (Arabameri et al., 2020; Chen & Chen, 2021). Hence, a multicollinearity test was employed to assess the feature correlations in this study. This test can uncover multicollinearity issues that might lead to unaccuracy results. The evaluation of multicollinearity was performed using the tolerance (TOL) and variance inflation factor (VIF) (Nguyen & Kim, 2021a; Bui et al., 2011). If VIF <10 or TOL >0.1, these values were suitable to consider as input data (Chen et al., 2018).

Random forest model

The Random Forest algorithm is an ensemble learning model comprising multiple individual decision trees. It is employed for the creation of both regression and classification models (Cheng et al., 2021). The outcome of this algorithm is influenced by two key variables: the number of trees and the maximum depth of each tree (Nguyen & Kim, 2021). As a result, in this study, a trial-and-error procedure was implemented to determine these variables. This method assisted in preventing overfitting and minimizing the occurrence of errors in the results produced by the Random Forest algorithm.

Model performance assessment

To assess the predictive capabilities of the models, three widely accepted validation criteria were chosen and employed: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) (Chao et al., 2021; Pham et al., 2018; Khanlari et al., 2012). R^2 serves to signify the statistical relationship between the observed values and the predicted values from the models (Mohammadi et al., 2022). It ranges from 0 to 1, where 0 indicates an incorrect model and 1 signifies a precise model. Higher R^2 values reflect superior model performance. On the other hand, RMSE represents the average squared difference between observed and predicted values (Khanlari et al., 2012), while MAE computes the average absolute difference between predicted and observed values (Khanlari et al., 2012). Both RMSE and MAE offer insights into the model's error assessment,

with lower values indicating enhanced model performance. The calculation of these values (MAE, RMSE, R^2) can be executed using the following equations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - X_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - X_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where, Y_i is measured value, X_i is predicted value, $\bar{Y}$ is the average of measured value, and n is data numbers.

4. RESULTS AND DISCUSSIONS

Feature selection results

Pearson correlation

Fig. 2 provides the heat map of Pearson's correlation coefficients, showcasing the relationships among soil physical properties. The correlation coefficients among *w*, *e*, *PL*, and *LL* show a high value from 0.8 to 1.0 and achieved a statistically significant positive association. Highly correlated input features have the potential to diminish the performance of machine learning algorithms. In order to ensure optimal model performance, these input features will be evaluated for potential removal from the database due to their significant similarity to each other (Mohammadi et al., 2022). Therefore, *w*, *e*, *PL* will be eliminated from the input features of the RF model.

Multicollinearity test

Following the removal of three input features, a multicollinearity test was carried out on the remaining features to evaluate the correlations between each feature and the others. The results of this assessment showed that all features have no multicollinearity, according to the VIF (1.099–1.363) and TOL values (0.734–0.910).

Predicted shear strength parameters

Random forest was utilized to estimate *c* and *φ* in this study. The RF model was designed with input layers comprising soil physical properties (□, S_r, *LL*, *SC*, *CC*), while the output layer consisted of *c* and *φ*. For developing RF model to predict the SS parameters, the available data was divided into two sets: a training set, which encompassed 80% of the data, and a test set, comprising the remaining 20% of the data. The training set served the purpose of fitting and training the RF model, while the test set was employed to assess the model's performance on previously unseen data. The outcomes of this evaluation are presented in Fig. 3. The RF model exhibited a high performance, with an R^2 of 0.9385 for cohesion prediction during the training stage (Fig. 3a), and an R^2 of 0.9764 for friction angle prediction (Fig. 3b). These superior results were consistently observed when the models were tested using the validation dataset, achieving an R^2 of 0.9233 for cohesion prediction (Fig. 3a) and an even more impressive R^2 of 0.9773 for friction angle (Fig. 3b). Table 2 shows the MAE, RMSE values from the training, and testing datasets. For the cohesion prediction, the RF model achieved the MAE, and RMSE values of 1.7162, 1.9630, and 2.2620, 2.4138 for the training, and testing datasets, respectively. When predicting friction angle, the RF model performs the MAE, and RMSE values of 0.5248, 0.5376, and 0.7887, 0.7210 for the training, and testing datasets, respectively. The evaluation results validate the success of the RF model in accurately predicting cohesion and friction angle using readily available physical properties.

Table 2. Evaluation results of the RF model

Metrics	*c*		□	
	Train	Test	Train	Test
MAE	1.7162	1.9630	0.5248	0.5376
RMSE	2.2620	2.4138	0.7887	0.7210
R^2	0.9385	0.9233	0.9764	0.9773

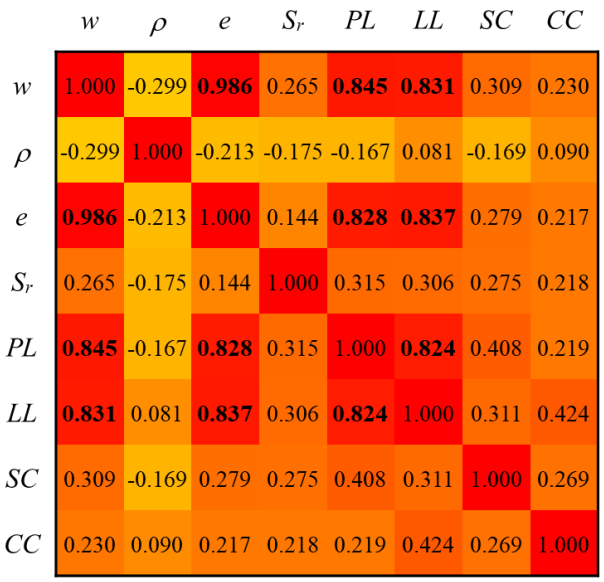

Figure 2. Heat map of Pearson's correlation

Numerous empirical formulas for predicting soil shear strength parameters are available in the literature. For the purpose of evaluating the comparative performance of the RF model, four specific studies were selected. These equations, proposed by Ersoy et al., (2013) (study 1); Roy & Dass, (2014) (study 2); Adunoye, (2014a, 2014b) (study 3) were chosen because the database employed in this study encompasses the majority of the parameters needed for calculating shear strength parameters via these methods.

The performance of these empirical methods and the RF model on the training and validation sets is presented in Figs. 4 and 5, specifically for cohesion and angle of friction, respectively. Regarding the prediction of cohesion (Fig. 4), the used model demonstrates the highest R^2 values compared to the previous empirical equations for the training, and testing, respectively (Fig. 4a). In addition, the RF model exhibits lower MAE (Fig. 4b), and RMSE (Fig. 4c) values in comparison to the previous model. These results are similarly observed in predicting friction angle (Fig. 5). These outcomes underscore the enhanced predictive capacity of the Random Forest (RF) model when compared to the earlier models. The main reason for this finding comes from the RF model, being a bagging ensemble learning technique grounded in decision tree algorithms, gains performance improvements by aggregating the decisions of individual models (Nguyen & Kim, 2021). Consequently, the predicting results of the RF exhibits higher accuracy compared to the empirical equations, which were obtained from linear regression method. This method relies on the assumption of linearity between the features, which can potentially yield less accurate results (Iyeke et al., 2016). Furthermore, this approach can lead to excessive similarities between an analysis and a dataset, potentially resulting in the failure to generate reliable predictions and accurately forecast future observations.

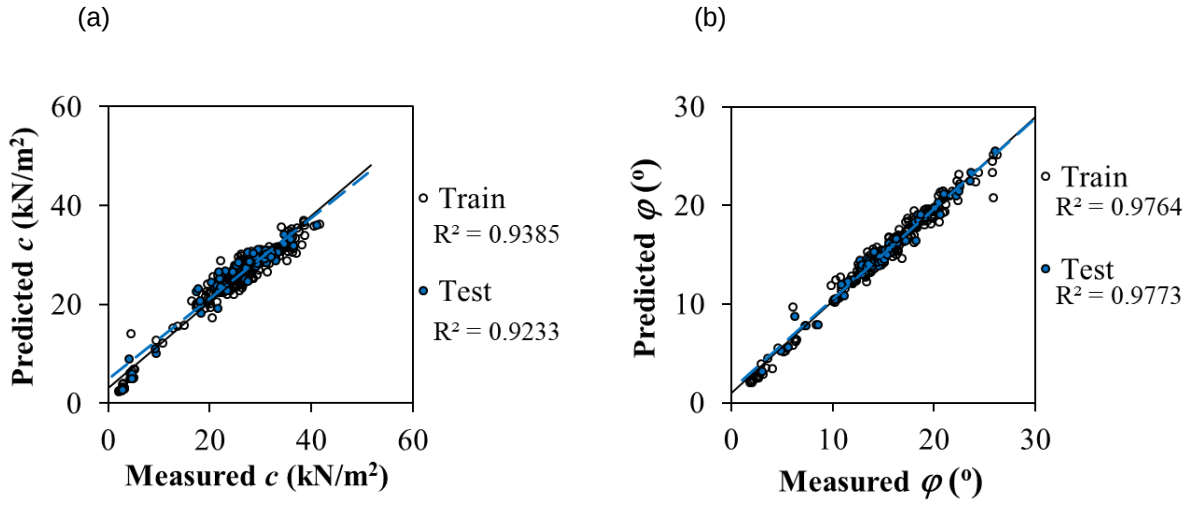

Figure 3. Correlation between the measured and predicted values: (a) cohesion, (b) friction angle

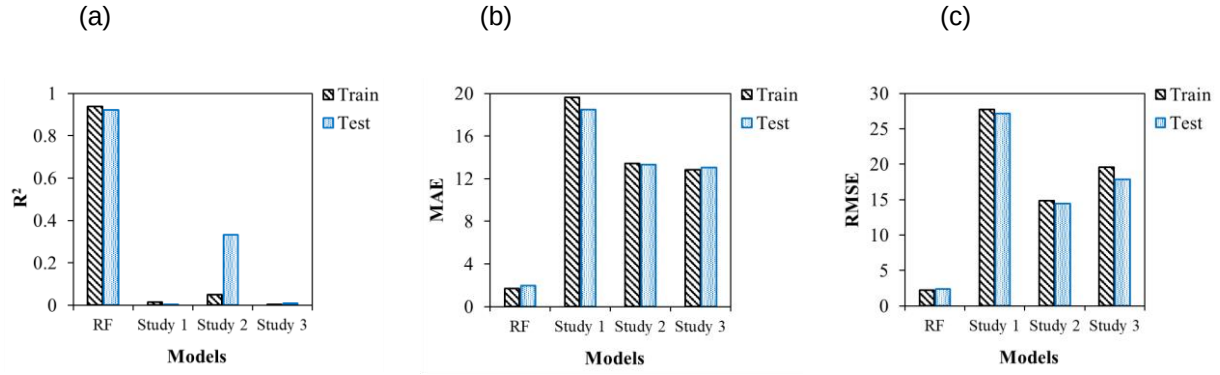

Figure 4. Comparison between this study and the previous studies for cohesion prediction: (a) R^2, (b) MAE, (c) RMSE. (study 1: Ersoy et al., 2013; study 2: Roy & Dass, 2014; study 3: Adunoye, 2014b)

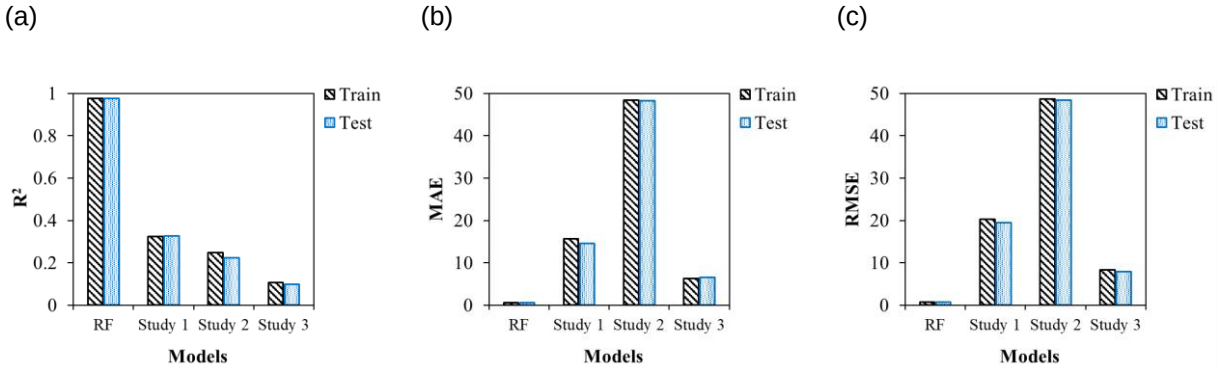

Figure 5. Comparison between this study and the previous studies for friction angle prediction: (a) R2, (b) MAE, (c) RMSE. (study 1: Ersoy et al., 2013; study 2: Roy & Dass, 2014; study 3: Adunoye, 2014a)

Sensitivity analysis was undertaken to assess the importance of the input features concerning the cohesion, and friction angle prediction. In this study, assessing the contributions and levels of importance of the input features through Decision Tree Feature Importance (DTFI) technique. The DTFI yields importance scores for the features computed using Gini impurity. The significance of the input parameters is visualized in Fig. 6. From Fig. 6, *SC*, *LL*, □, and *CC* have significantly affected both cohesion (Fig. 6a), and friction angle (Fig. 6b) of soil. These findings

align with prior researches that investigated the correlation between physical properties and shear strength parameters (Murthy, 2002; Kayadelen et al., 2009; Mousavi et al., 2011; Dadkhah et al., 2010; Tafari et al., 2021; Jiang et al., 2021; Ersoy et al., 2013; Roy & Dass, 2014; Adunoye, 2014a).

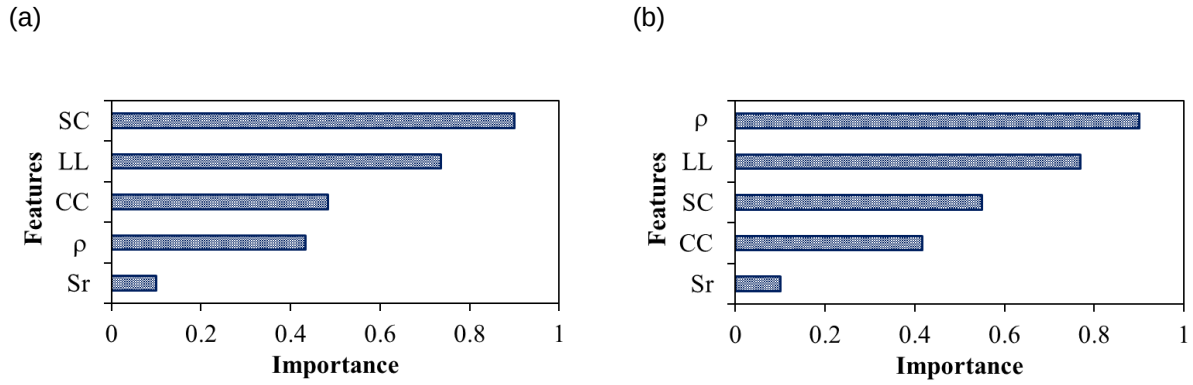

Figure 6. Importance of the input features: (a) cohesion, (b) friction angle

CONCLUSIONS

This study explores the potential of the Random forest model in predicting shear strength parameters of soil. To assess the performance of the used method, the study employs metrics such as R^2 (determination coefficient), RMSE (root mean square error), and MAE (mean absolute error). The values of these metrics proved that the proposed framework is successful in predicting shear strength parameters from the easily-available physical properties.

The results obtained from this study were compared to the results from the previous empirical equations. This comparison demonstrated the outperforming predictive ability of the RF model in shear strength parameters estimation.

The impact of the input features on the prediction outcomes was also assessed through the Decision Tree Feature Importance (DTFI) method. The results from the DTFI indicate that *SC*, *LL*, □, and *CC* wield significant influence on the prediction of cohesion and friction angle.

ACKNOWLEDGEMENT

This research is funded by International University, Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number SV2022-CE-03.

REFERENCES

[1]. Adunoye, G. O. (2014a). Fines content and angle of internal friction of a lateritic soil: an experimental study. *American Journal of Engineering Research*, *3*(3), 16–21.

[2]. Adunoye, G. O. (2014b). Study of relationship between fines content and cohesion of soil. *British Journal of Applied Science & Technology*, *4*(4), 682–692.

[3]. Amin Benbouras, M., & Petrisor, A.-I. (2021). Prediction of swelling index using advanced machine learning techniques for cohesive soils. *Applied Sciences*, *11*(2), 536.

[4]. Arabameri, A., Saha, S., Roy, J., Chen, W., Blaschke, T., & Tien Bui, D. (2020). Landslide susceptibility evaluation and management using different machine learning methods in the Gallicash River Watershed, Iran. *Remote Sensing*, *12*(3), 475.

[5]. Bui, D. T., Lofman, O., Revhaug, I., & Dick, O. (2011). Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural Hazards*, *59*(3), 1413.

[6]. Chao, Z., Fowmes, G., & Dassanayake, S. M. (2021). Comparative study of hybrid artificial intelligence approaches for predicting peak shear strength along soil-geocomposite drainage layer interfaces. *International Journal of Geosynthetics and Ground Engineering*, *7*(3), 60.

[7]. Chen, W., Zhang, S., Li, R., & Shahabi, H. (2018). Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Science of the Total Environment*, *644*, 1006–1018.

[8]. Chen, X., & Chen, W. (n.d.). GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. *Catena*, *196*, 104833.

[9]. Cheng, Y.-S., Yu, T.-T., & Son, N.-T. (2021). Random forests for landslide prediction in tsengwen river watershed, central taiwan. *Remote Sensing*, *13*(2), 199.

[10]. Chitra, R., & Gupta, M. (2014). Neural networks for assessing shear strength of soils. *Int J Recent Dev Eng Technol*, *3*(4), 24–32.

[11]. Dadkhah, R., Ghafoori, M., Ajalloeian, R., & Lashkaripour, G. R. (2010). The effect of Scale Direct Shear Tests on The Strength parameters of Clayey Sand in Isfahan city, Iran. *Journal of Applied Sciences*, *18*.

[12]. Ersoy, H., Karsli, M. B., Cellek, S., Kul, B., Baykan, I., & Parsons, R. L. (2013). Estimation of the soil strength parameters in Tertiary volcanic regolith (NE Turkey) using analytical hierarchy process. *Journal of Earth System Science*, *122*, 1545–1555.

[13]. Iyeke, S. D., Eze, E. O., Ehiorobo, J. O., & Osuji, S. O. (2016). Estimation of shear strength parameters of lateritic soils using artificial neural network. *Nigerian Journal of Technology*, *35*(2), 260–269.

[14]. Jiang, Q., Cao, M., Wang, Y., Wang, J., & He, Z. (2021). Estimation of Soil Shear Strength Indicators Using Soil Physical Properties of Paddy Soils in the Plastic State. *Applied Sciences*, *11*(12), 5609.

[15]. Kayadelen, C., Günaydın, O., Fener, M., Demir, A., & Özvan, A. (2009). Modeling of the angle of shearing resistance of soils using soft computing systems. *Expert Systems with Applications*, *36*(9), 11814–11826.

[16]. Khanlari, G. R., Heidari, M., Momeni, A. A., & Abdilor, Y. (2012). Prediction of shear strength parameters of soils using artificial neural networks and multivariate regression methods. *Engineering Geology*, *131*, 11–18.

[17]. Mohammadi, M., Fatemi Aghda, S. M., Talkhablou, M., & Cheshomi, A. (2022). Prediction of the shear strength parameters from easily-available soil properties by means of multivariate regression and artificial neural network methods. *Geomechanics and Geoengineering*, *17*(2), 442–454.

[18]. Mousavi, S. M., Alavi, A. H., Gandomi, A. H., & Mollahasani, A. (2011). Nonlinear genetic-based simulation of soil shear strength parameters. *Journal of Earth System Science*, *120*, 1001–1022.

[19]. Murthy, V. N. S. (2002). *Geotechnical engineering: principles and practices of soil mechanics and foundation engineering*. CRC press.

[20]. Nguyen, B.-Q.-V., & Kim, Y.-T. (2021). Landslide spatial probability prediction: a comparative assessment of naive Bayes, ensemble learning, and deep learning approaches. *Bulletin of Engineering Geology and the Environment*, *80*, 4291–4321.

[21]. Nguyen, B.-Q.-V., Song, C.-H., & Kim, Y.-T. (2022). A Hybrid Physical and Machine Learning Model for Assessing Landslide Spatial Probability Caused by Raising of Ground Water Table and Earthquake in Atsuma, Japan—Case Study. *KSCE Journal of Civil Engineering*, *26*(8), 3416–3429.

[22]. Pham, B. T., Hoang, T.-A., Nguyen, D.-M., & Bui, D. T. (2018). Prediction of shear strength of soft soil using machine learning methods. *Catena*, *166*, 181–191.

[23]. Roy, S., & Dass, G. (2014). Statistical models for the prediction of shear strength parameters at Sirsa, India. *International Journal of Civil and Structural Engineering*, *4*(4), 483.

[24]. Roy, S., Prajapati, A. K., & Maurya, A. K. (2019). Prediction of Shear Strength Parameters Using Multiple Regression Analysis. *International Journal of Landscape Planning and Architecture*, *5*(2), 26–39.

[25]. Salari, P., Lashkaripour, G., & Ghafoori, M. (2015). Presentation of empirical equations for estimating internal friction angle of SP and SC soils in Mashhad, Iran using standard penetration and direct shear tests and comparison with previous equations. *International Journal of Geography and Geology*, *4*(5), 89.

[26]. Stefanow, D., & Dudziński, P. A. (2021). Soil shear strength determination methods–State of the art. *Soil and Tillage Research*, *208*, 104881.

[27]. Tafari, T., Quezon, E. T., & Yasin, M. (2021). Statistical Analysis on Shear Strength Parameter from Index Properties of Fine-grained Soils. *Available at SSRN 3811717*.

[28]. Yoseph, H. (2022). *ESTIMATION OF SOIL SHEAR STRENGTH PARAMETERS FROM INDEX PROPERTIES USING ANN THE CASE OF ADDIS ABABA*.

[29]. Zakharov, A., Shenkman, R., Ofrikhter, I., & Ponomaryov, A. (2022). Estimation of soil properties by an artificial neural network. *Magazine of Civil Engineering*, *110*(2), 11011.

[30]. Zhu, L., Liao, Q., Wang, Z., Chen, J., Chen, Z., Bian, Q., & Zhang, Q. (2022). Prediction of soil shear Strength parameters using combined data and different machine learning models. *Applied Sciences*, *12*(10), 5100.

DOI: https://doi.org/10.15379/ijmst.v10i1.2751