Jordan Journal of Civil Engineering, Volume **17**, No. **3**, **2023**
DOI: https://doi.org/10.14525/JJCE.v17i3.14

Regression Analysis for Predicting Soil Strength in Bangladesh

Shadman Rahman Sabab[1], Hossain Md. Shahin[2], Md. Muftashin Muhim Bondhon [3] and Md. Ehsan Kabir [4]

[1] Undergraduate Student, Dept. of Civil Engineering, Islamic University of Technology, Dhaka K B Bazar-1704, Bangladesh. * Corresponding Author. E-Mail: shadmansabab@iut-dhaka.edu
[2] Professor, Dept. of Civil Engineering, Islamic University of Technology, Dhaka K B Bazar-1704, Bangladesh. E-Mail: shahin@iut-dhaka.edu
[3] Undergraduate Student, Dept. of Civil Engineering, Islamic University of Technology, Dhaka K B Bazar-1704, Bangladesh. E-Mail: muftashinmuhim@iut-dhaka.edu
[4] Undergraduate Student, Dept. of Civil Engineering, Islamic University of Technology, Dhaka K B Bazar-1704, Bangladesh. E-Mail: ehsankabir@iut-dhaka.edu.

ABSTRACT

This study focuses on establishing a robust relationship between Standard Penetration Test-N values (SPT-N), geotechnical parameters and unconfined compressive strength (q_u) using regression analysis. The proposed relationship offers a reliable method for estimating q_u based on SPT-N values. A comprehensive dataset comprising approximately 200 soil samples collected from various boreholes across Dhaka city was utilized. Multiple Linear Regression (MLR), Rando-forest Regression (RFR) and AdaBoost Regression techniques were employed to develop a unified correlation model. Evaluation metrics including R-squared (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), along with Trend-behavior Analysis were employed to assess and compare the performances of the models. Additionally, sensitivity analysis was carried out on the selected model in order to assess the importance of each parameter used to predict q_u. Finally, the selected model was compared against the existing empirical models that were published in previous studies. In terms of evaluation metrics and Trend-behavior Analysis, the results showed that the RFR model performed better than the others. Additionally, the selected model outperformed the others, demonstrating the highest R^2 score, the smallest RMSE and MAE values and lower residuals compared to the previous models. Hence, the proposed model provides accurate predictions of q_u for clayey soil in Bangladesh. Its implementation could ensure more efficient geotechnical designs, specifically adjusted to the geological conditions of the Dhaka region. While previous studies have established regional equations for various parts of the world, our model uniquely has incorporated the Plasticity Index (PI) as a predictor for q_u and is specifically calibrated for the geological characteristics of Dhaka city. The findings of this study highlight the effectiveness and applicability of regression analysis in predicting q_u for Dhaka's soil properties, thus introducing a valuable tool for enhancing the accuracy and effectiveness of geotechnical assessments and design in the region.

KEYWORDS: Unconfined compressive strength, Standard penetration test-N values, Plasticity index, Multiple linear regression, Random-forest regression, AdaBoost regression, Evaluation metrics, Trend-behavior analysis, Sensitivity analysis.

INTRODUCTION

Unlocking the secrets hidden beneath the earth's surface lies the key to successful structural design. Soil properties play a vital role in shaping the stability and reliability of any construction project. In the field of engineering, soil is characterized as an unbound collection of mineral grains and decomposed organic materials, surrounding voids filled with liquid and gas. As an adaptable construction material, soil serves as a fundamental component in numerous civil-engineering

Received on 21/3/2023.
Accepted for Publication on 8/6/2023.

works, providing the essential groundwork for structural foundations. Consequently, a comprehensive understanding of soil properties becomes vital for civil engineers. These properties cover a wide array of factors, including the soil's origin, distribution of grain sizes, drainage capabilities, compressibility, shear strength and capacity to withstand imposed loads. Probing into the details of soil properties empowers engineers to make informed decisions, ensuring the integrity and stability of their construction projects (Das, 2010). However, traditional methods of gathering this crucial information through labor-intensive and time-consuming *in-situ* testing have often proven to be deterrent for engineers and budget-conscious individuals. To get information on the geological properties of soil, the Standard Penetration Test (SPT) is conducted.

SPT is a widely utilized method worldwide for estimating the *in-situ* properties of granular soil. It involves conducting tests to visually examine disturbed samples and conducting laboratory testing to assess the ground's consistency. The N value, which represents the number of blows required for 12-inch (300mm) penetration after a 150-mm seating drive, is a key parameter obtained from the SPT. The SPT apparatus consists of a free-fall hammer, along with a hollow cylindrical mass that slides over a steel rod, which is lifted to a height of 760mm using a wire and then automatically released to drive a split-spoon sampler into the ground. The disturbed samples collected by the split-spoon sampler are visually inspected before being sent for laboratory testing (Skempton, 1986).

Clayey soil strata are collected in undisturbed samples for laboratory testing to evaluate their engineering properties. The term "undisturbed" refers to the degree of disruption to the *in-situ* qualities of the soil. These samples are taken utilizing specialized tools to minimize disturbance to the soil's *in-situ* structure and moisture content. In Bangladesh, a thin-walled Shelby tube is utilized to extract the undisturbed samples of cohesive soils for strength (Mayne et al., 2001).

In the soil of Dhaka city, two common types of cohesive soils are encountered: fat clay and lean clay. Fat clay is categorized by its high plasticity and compressibility. It has a greasy texture due to its elevated mineral content. When damp, it can become challenging to work with, but it exhibits a considerable amount of strength when dry. The liquid limit of fat clay exceeds 50 and its plasticity index (PI) ranges from 30 to 50. On the other hand, lean clay contains a higher concentration of silt or sand. Its PI ranges from low to medium and its liquid limit is less than 50. The PI of lean clay typically falls between 10 and 30 (ASTM D 2487-11, 2011).

To describe the consistency of fine-grained clay and silt soils with different moisture contents, Atterberg has defined the transitions into the following terms. The shrinkage limit is defined as the moisture content (in percent) at which the change from solid to semi-solid occurs. The moisture content at the point of transition from semi-solid to plastic is the plastic limit, while the moisture content at the point of transition from plastic to liquid is the liquid limit. These are also known as the Atterberg limits (Das, 2010).

For this study, unconfined compressive strength (q_u) is a crucial parameter under consideration. It represents the maximum axial compressive stress that a cohesive soil can bear under zero confining stress. Measuring q_u is not only cost-effective, but also provides a quick assessment of the shear strength of cohesive soil. The unconfined compression test (ASTM D2166) is commonly used to determine this parameter. However, this test may not always be readily available due to challenges in obtaining soil specimens, limited laboratory facilities or a shortage of skilled technicians proficient in operating the required instruments. In such cases, the utilization of empirical models emerges as a viable alternative that offers effectiveness and efficiency in predicting the outcome by establishing correlations with other readily available parameters. In order to get beyond the drawbacks of conventional testing techniques, this study investigates the possibility of regression analysis as an innovative way of estimating soil strength in Bangladesh.

Regression analysis is one of the leading strategies that provide an effective and efficient alternative to estimating soil strength. New prospects for geotechnical engineers may be unlocked by utilizing data and statistical modeling, which might change their knowledge of and ability to utilize soil in a variety of building projects. Regression analysis would not only provide an accurate prediction of soil-strength parameters, but also offer significant benefits to the field of geotechnical engineering. Its cost-effectiveness eliminates the need for extensive laboratory testing, saving both time and resources without compromising accuracy. The time efficiency of regression analysis would allow for rapid assessment and evaluation of soil strength, facilitating timely decision-making in project planning. Moreover, the insights gained from regression analysis could enable engineers to improve design and construction processes, optimizing the structural integrity and stability of geotechnical structures. By incorporating data-driven predictions, regression analysis enhances risk management by identifying potential challenges and guiding appropriate mitigation strategies. Moreover, regression analysis can raise innovation in geotechnical engineering, since it progressively improves the understanding of soil behavior and paves the path for more efficient and sustainable construction approaches. With these benefits, regression analysis could emerge as a powerful tool in the pursuit of safer, more reliable and cost-effective geotechnical-engineering solutions.

In the past, several countries including Japan, Iran, Malaysia, Canada, Turkey, India and even Bangladesh have developed empirical correlations specific to their respective regions. However, these existing correlations exhibit certain limitations. It has been observed that the values obtained from laboratory tests significantly differ from those predicted using the empirical formulae derived from previous research studies. These discrepancies can be attributed to the variations in study areas and the distinct soil properties and behaviors exhibited. As Bangladesh undergoes rapid development, numerous upcoming projects necessitate the prediction of q_u in situations where laboratory testing may not be feasible. Hence, the primary objective of this study is to establish an empirical equation that precisely and accurately predicts q_u for the Dhaka region of Bangladesh, addressing the shortcomings of previous correlations and providing a reliable tool for engineering practitioners.

The first ever study to determine the relationship between C and SPT-N was conducted by Terzaghi & Peck (1967). They proposed Equation (1) for fine-grained soil:

$$C = 6.25\ N \qquad (1)$$

where, C = cohesion in kPa, N=field SPT-N (value for limited range).

Schmertmann (1979) and Sowers (1979) proposed that C increases with an increase in the plasticity index.

For highly plastic soil: $C = 12.5\ N$ (2)

For medium plastic soil: $C = 7.5\ N$ (3)

For kow plastic soil: $C = 3.75\ N$ (4)

In 1996, Serajuddin & Alim Chowdhury studied the correlation between SPT-N and unconfined compressive strength (q_u) of Bangladesh soil deposits. 420 soil samples were collected from different locations in Bangladesh, especially from Dhaka metropolitan, Tangail and Modhupur. Atterberg-limit tests, sieve analysis and unconfined compression tests were conducted. The liquid limit and plasticity index for all the soil samples have been plotted in the Casagrande plasticity chart to define the classification of soil samples by USCS. Soil samples of different ranges of liquid limit were plotted to identify the K values for a range of liquids. From the graphs, Equation (5) was found.

$$q_u = KN \qquad (5)$$

where, K=14.3 for liquid limit <= 35%, K=16.9 for liquid limit =36% to 50% and K=17.8 for liquid limit >=51%, N = SPT-N.

Hettiarachchi & Brown (2009) developed a model using a correlation between corrected SPT-N and undrained shear strength for clayey soil using data obtained from Commonwealth Associates, Inc. Equation (6) which they have proposed suggested that N_{60} is directly proportional to s_u of clay.

$$\frac{s_u}{P_a} = \alpha' N_{60} \qquad (6)$$

where, s_u = undrained shear strength, N_{60} = corrected SPT-N, P_a = atmospheric pressure (100kPa) and α' = 0.041.

Kumar et al. (2016) proposed Equation (7) using a random number-generation method between field SPT-N and cohesion (kPa). The range of SPT-N values was between 2 and 30. The ranges were compiled using random number-generation technology and information from the literature.

$$c = - 2.2049 + 6.484\ N. \quad (7)$$

Undrained shear strength (s_u) can be used to express a sub-soil's bearing capacity. Unconfined compressive strength (q_u) and the value of undrained shear strength (s_u) are equivalents. Theoretically, this number is twice as large as cohesion (c); thus, (q_u) is also twice as large as (c) (Widodo et al., 2012).

In a recent study conducted by Larbi et al. (2019), the authors examined the influence of database size on the use of Artificial Neural Network (ANN) for predicting the compressive strength of concretes containing reclaimed asphalt pavement. They emphasized that ANN, as a machine-learning algorithm, demonstrates effective predictive capabilities for the considered materials. In their research, they utilized feed-forward networks and back-propagation algorithms to construct an ANN model that more accurately estimates the compressive strength of the concretes. Their findings revealed a significant correlation between database size and prediction accuracy, indicating that larger databases yield improved results.

Hossain et al. (2021) conducted a study in which they employed multiple-regression analysis techniques, including Multiple Linear Regression (MLR), Artificial Neural Network (ANN) and Support Vector Machine (SVM), to estimate the internal frictional angle (Φ) of soil. SVM, like ANN, is a machine-learning algorithm capable of effectively predicting a target variable. In their study, similar to the present study, they utilized evaluation metrics to identify the best-performing model for predicting Φ and compared it with existing models. They concluded that SVM performs better than both MLR and ANN in predicting Φ.

In a study focused on enhancing the unconfined compressive strength (q_u) of soils through the use of additives and predicting the strength behavior of stabilized soils, Tabarsa et al. (2021) employed Artificial Neural Network (ANN) and Support Vector Machine (SVM) models. Their objective was to develop a model that could estimate the required number of additives for predicting q_u. Two different architectures of ANN were utilized: one with a single hidden layer consisting of seven neurons and another with two hidden layers, each containing four neurons. The authors also used the SVM model, utilizing Gaussian radial basis functions (SVM-RBF) and polynomial functions (SVM-poly). Metrics like the average absolute percentage error (AAPE) and correlation coefficient (R) were used to compare the ANN and SVM models. The results showed that SVM was more accurate in predicting q_u. The study also showed that both models performed better than multiple-regression analysis. Sensitivity analysis was also carried out and the results showed that cement and lime concentrations had a stronger impact on q_u.

Saadat & Bayat (2022) used the Adaptive Neuro Fuzzy Inference System (ANFIS) and Non-linear Regression (NLR) models for predicting unconfined compressive strength (q_u) in a separate study that aimed to examine the effects of stabilizer content, curing time and moisture content on q_u using 150 samples of stabilized soil. In order to capture the underlying physical correlations between input and output variables, ANFIS is a hybrid model that combines ANN with fuzzy logic. NLR was chosen due to the fact that there is an inverse correlation between q_u and moisture content. Metrics such as Root Mean Square Error (RMSE), bias and correlation coefficient (R) were used to compare the two models' performances. According to the findings, ANFIS performed better in predicting q_u than NLR. Additionally, sensitivity testing revealed that the cement content had the greatest impact on the q_u value.

The models developed in Larbi et al. (2019), Hossain et al. (2021), Tabarsa et al. (2021) and Saadat & Bayat (2022) studies were not used in comparison with the current study. This is because Larbi et al.'s (2019) study was to estimate the compressive strength of concretes, whereas Hossain et al. (2021) built a model to predict the internal frictional angle (Φ) of soil. Although Tabarsa et al. (2021) and Saadat & Bayat (2022) aimed to predict q_u, their models used different input variables compared to the present study. Consequently, ANN and ANFIS were not considered as regression models in the present study due to their reliance on substantial amounts of data, pre-processing steps like feature scaling and the need to address outliers. In contrast, tree-based regression models, such as RFR and AdaBoost, which were employed in the current study, do not have the same extensive data requirements and pre-processing steps as ANN and ANFIS. Although SVM performs well with smaller datasets, it shares similar limitations with ANN in terms of predictive performance, which is why it was not considered as a regression model for the present study.

METHODOLOGY

Study Area

The samples were collected from Dhaka. 205 samples were collected from Dhaka mass rapid transit development project (MRT) line-6, which covers the area of Uttara, Pallabi, Mirpur11, Mirpur10, Kazipara, Shewrapara, Agargaon, Bijoy Sarani, Farmgate, Kawranbazar, Shahbag, Dhaka University, Bangladesh Secretariat & Motijheel, Feasibility study and preliminary design of Dhaka subway covering area of Gabtoli, Mohakhali, Sayedabad bus terminals, Sadarghat river port, Kamalapur and Tongi Junction railway stations. A standard penetration test was conducted at 1.5m intervals to investigate the soil properties. The collected samples of the Dhaka zone mainly consist of lean clay with some fat clay and a small proportion of other clay types. A small quantity of silty sand was also found; however, it was not considered when developing the ML model. Figure 1 shows the location of boreholes in Dhaka and Gazipur.

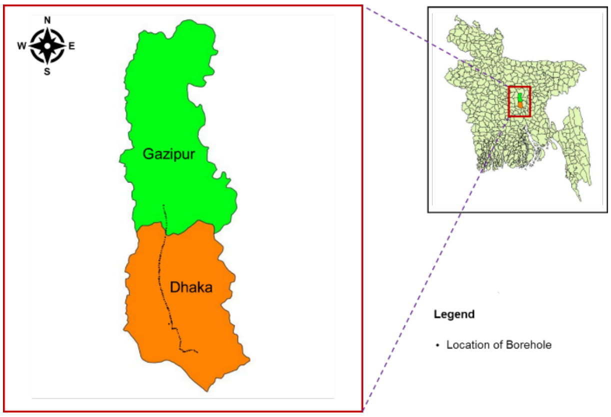

Figure (1): Borehole locations of the current study

Collection and Preparation of Data

197 datasets were incorporated in building the regression models. Each observation comprises SPT-N value, depth of collection, Atterberg limits, water content, specific gravity and q_u test results. The physical properties are shown in Table 1. The SPT-N collected from the field needs to be corrected, as several factors can cause its variation. The type of SPT hammer used, borehole diameter, sampling method and rod length can cause such changes. Therefore, SPT-N is corrected using Equation (8),

$$N_{60} = \frac{N\eta_H\eta_B\eta_S\eta_R}{60} \tag{8}$$

where, N_{60} = corrected SPT-N, according to field conditions.

N = SPT-N measured in the field.

η_H = efficiency of hammer.

η_B = borehole-diameter correction coefficient.

η_S = correction coefficient of sampler.

η_R = rod-length correction coefficient.

The hammer efficiency was taken as η_H = 0.60, as the doughnut type hand-dropped was used to carry out the field test (Jay et al., 2014).

Table 1. Physical properties of soil

Soil properties	Unit	Range
Moisture Content	%	17.39-97.10
Specific Gravity	N/A	2.67-2.77
Unconfined Compressive Strength	kPa	26.2-220.1
Plastic Limit	%	14.3-38.3
Liquid Limit	%	29.5-81.6
Plasticity Index	%	10.01-59.78

Regression Models

Multiple Linear Regression (MLR)

Multiple linear regression is the extension of simple linear regression for data with various predictor variables with one outcome. Equation (9) is the general expression of the MLR model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad \varepsilon_i \overset{indep}{\sim} N(0,\sigma^2) \tag{9}$$

where β_0 is the intercept, β_k is the slope or partial regression coefficients with respect to predictor variables x_k (k=1, 2, …, p-1) and y is the outcome (Eberly, 2007).

Random-forest Regression (RFR)

Random forest is a regression technique that mixes the performance of various decision-tree algorithms to predict a variable (Breiman, 2001; Gou et al., 2011; Rodriguez-Galiano et al., 2012b, as cited in Rodriguez-

Galiano, 2015). So, when random forest receives an (x) input vector, consisting of values of different features analyzed for a given training area, random forest builds a K number of regression trees and the results are averaged. After K number of trees, $\{T(x)\}_1^K$ are made. The random-forest regression predictor is shown in Equation (10),

$$f_{rf}^{K}(x) = \frac{1}{K}\sum_{K=1}^{K} T(x). \qquad (10)$$

To avoid the correlation of different trees, random forest increases the trees' diversity by growing them through various training data sub-sets through a process called bagging. Bagging resamples the original dataset randomly with replacement, without deletion of data selected from the input sample for generating the next sub-set {h(x, Θ_k), k = 1,…,K}, where {Θ_k} are independent random vectors with the same distribution, thus creating new training data. As a result, some data is used more than other data, while some data is never used during training. This helps obtain greater stability, as it becomes more robust whenever it finds any slight variation in input data and increases the accuracy of prediction (Rodriguez-Galiano, 2015).

AdaBoost

AdaBoost, a boosting algorithm, uses a lot of weak learners to create classifiers that constantly perform better than random guessing. AdaBoosting performs by frequently running the weak learners on different distributions of the training datasets and combining the outputs. On the machine's performance in the previous iteration, the training-data distribution in the next iteration is affected (Solomatine & Shrestha, 2004).

Initially, it was a boosting approach using large training data for classification problems (Schapire, 1990, as cited in Solomatine & Shrestha, 2004). It was then improved into another algorithm called AdaBoost (Freund & Schapire, 1996, as cited in Solomatine & Shrestha, 2004). Mainly used for classification problems, Freund and Schapire (1997, as mentioned in Solomatine & Shrestha, 2004) extended it to solve regression problems and named it AdaBoost R. Later, Drucker (1997, as cited in Solomatine & Shrestha, 2004) modified the algorithm and called it AdaBoost R2.

Evaluation Metrics

Coefficient of determination (R^2 score), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used in evaluating the performance of the three machine-learning algorithms. In previous studies, it can be observed that a larger value of R^2 score Eq. (11) and smaller values of RMSE (Eq. 12) and MAE (Eq. 13) show that the model is more accurate.

Coefficient of determination:

$$R^2 score = 1 - \frac{\sum_{i=1}^{n}(y_{actual}-y_{predict})^2}{\sum_{i=1}^{n}(y_{actual}-\bar{y}_{average})^2} \qquad (11)$$

Root Mean Squared Error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{actual}-y_{predict})^2}{n}} \qquad (12)$$

Mean Absolute Error:

$$MAE = \frac{\sum_{i=1}^{n}|(y_{actual}-y_{predict})|}{n} \qquad (13)$$

Here, y_{actual} = actual observation, $y_{predict}$ = predicted observation, $\bar{y}_{average}$= average of observations, n = total number of datasets.

The whole process of building the model is as follows:

1. Data is collected and SPT-N is corrected.
2. The dataset is split into training and testing machine-learning models.
3. Each regression model is run and the best model is decided through evaluation metrics and trend-behavior analysis.
4. Sensitivity analysis is carried out in order to determine the importance of each parameter to predict q_u in the chosen model.
5. The chosen model is then compared with other widely used existing models.

RESULTS AND DISCUSSION

Statistical Inspection of the Dataset

Table 2 shows the distribution of all the variables used in the regression. Here, all the variables are positively skewed; only PI (Plasticity Index) is relatively symmetrical, while N_{60}, depth(m) and q_u (unconfined compressive strength) are moderately skewed. Figure 2 shows the correlation between the variables. It can be observed that there is a very strong correlation between q_u and N_{60}, while both depth(m) and PI have a weak correlation with q_u.

Table 2. Statistical details of the dataset

	N_{60}	Depth(m)	PI	q_u
Count	197	197	197	197
Mean	8.9389	10.9832	30.0511	95.4000
Standard deviation	4.7248	5.5622	10.0623	44.9926
Min.	1.9000	1.9500	10.0100	26.2300
25%	5.0000	6.5000	22.6900	61.0900
50%	8.0000	9.5000	29.1400	87.3400
75%	12.3500	14.0000	37.2400	122.7100
Max.	22.0000	29.0000	59.7800	220.1100
Skewness	0.6483	0.8099	0.3221	0.6007
Kurtosis	-0.3801	0.1811	-0.3147	-0.3801

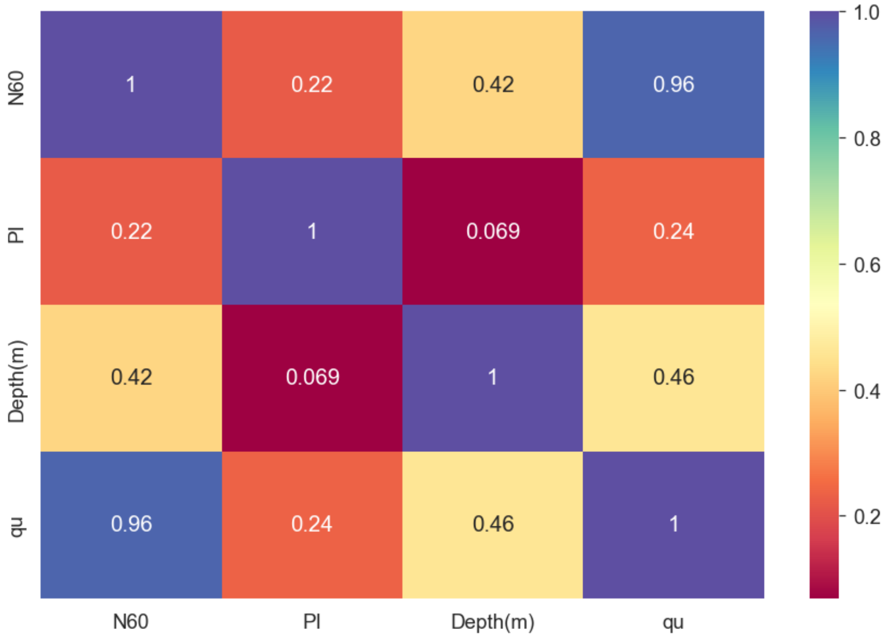

Figure (2): Correlation heatmap of the dataset

Output of Regression Models

The three regression models were operated using scikit-learn (Pedregosa et al., 2011). The correlation heatmap was generated using seaborn (Michael Waskom, 2021), while the residual plots were generated using YellowBrick (Bengfort et al., 2018).

Multiple Linear Regression (MLR)

The regression equation obtained from training the model (Eq. 14) is as follows:

$$q_u = 9.045N_{60} + 0.586depth(m) + 0.0672PI + 6.062. \qquad (14)$$

It can be noted that N_{60} has the largest coefficient among the variables, indicating its significant influence in predicting q_u. The dataset was divided into training and testing sets using a 70:30 split ratio and a random state of 98. The performance of the model was evaluated using R^2, RMSE and MAE scores. As shown in Table 3, the R^2 score obtained from the training dataset is 0.932, which indicates that approximately 93.2% of the variance in q_u can be explained by the model. The

RMSE and MAE values are 11.876 and 9.208, respectively, suggesting that the model can effectively capture the underlying patterns in the training data and produce accurate predictions. Similarly, the R^2 score obtained from the testing dataset is 0.911. However, the slightly higher RMSE and MAE values for the testing dataset of 12.603 and 10.323, respectively, indicate that the model has a slightly higher level of prediction errors when the model is applied to unseen data. Figures 3 and 4 depict the relationships between the actual and predicted values for the training and testing datasets, respectively. Both figures demonstrate a strong linear relationship, which indicates the model's ability to accurately predict q_u. Additionally, Figure 5 illustrates the residual plot for both the training and testing datasets. The model captures the underlying patterns in the data, because the majority of the residuals are randomly distributed across the horizontal axis. The residuals' histogram shows that the errors are typically between 20 kPa and -20 kPa, indicating that the model's predictions are generally close to the actual values.

Table 3. The results obtained from each regression model

	Model	MLR	RFR	AdaBoost
Training	R^2 score	0.932	0.986	0.944
	RMSE	11.876	5.448	10.782
	MAE	9.208	4.397	9.373
Testing	R^2 score	0.911	0.914	0.904
	RMSE	12.603	12.422	13.115
	MAE	10.323	9.544	9.613

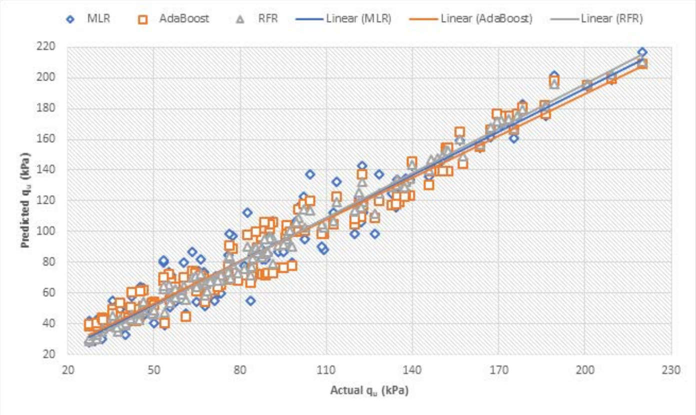

Figure (3): Actual q_u (kPa) *vs.* predicted q_u (kPa) for training dataset

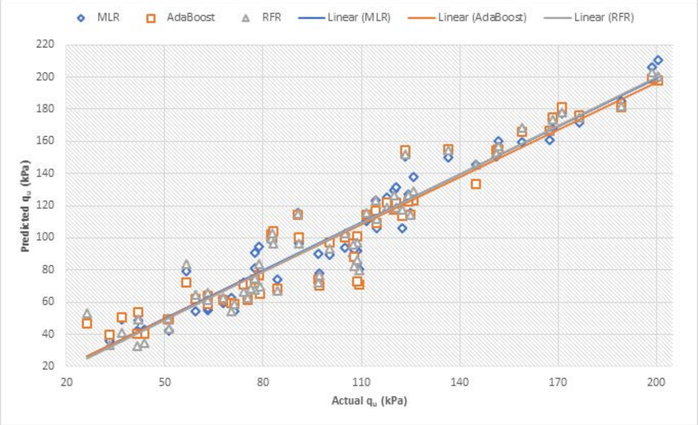

Figure (4): Actual q_u (kPa) *vs.* predicted q_u (kPa) for testing dataset

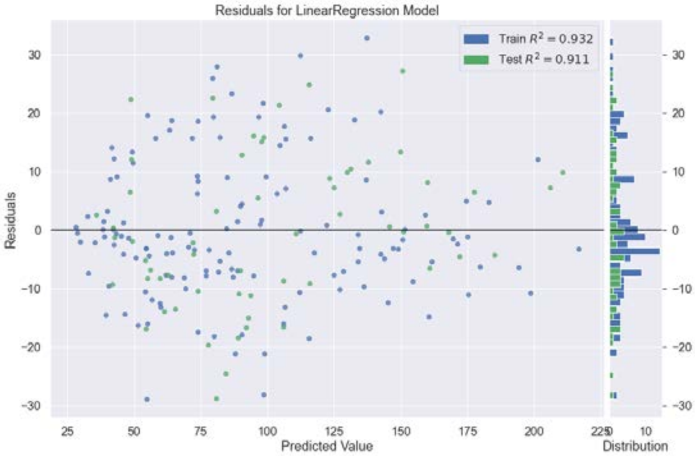

Figure (5): Residual q_u (kPa) *vs.* predicted value for MLR

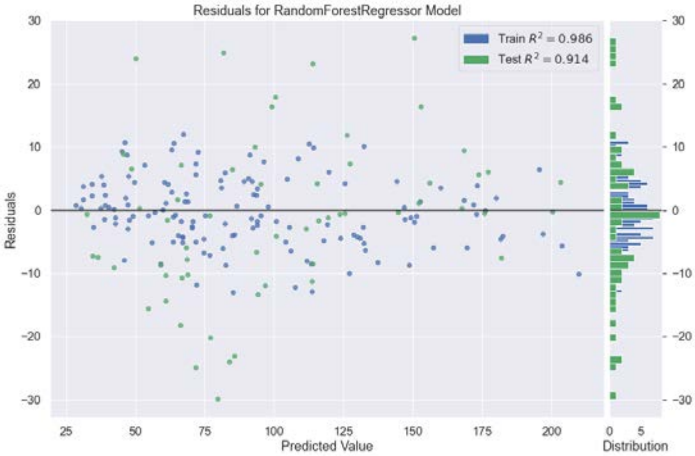

Figure (6): Residual q_u (kPa) *vs.* predicted value for RFR

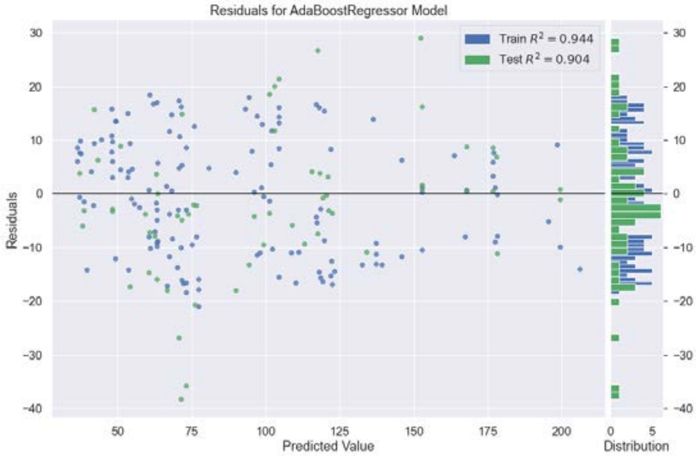

Figure (7): Residual q_u (kPa) *vs.* predicted value for AdaBoost

AdaBoost

The AdaBoost regression model was applied to the dataset, which was split into training and testing sets using a 70:30 ratio and a random state of 98. Just like MLR, the performance of this model was also evaluated using R^2, RMSE and MAE scores. From Table 3, the R^2 score obtained from the training dataset is 0.944, indicating that approximately 94.4% of the variance in

q_u can be explained by the model. The RMSE and MAE values are 10.782 and 9.373, respectively, indicating that the model can effectively capture the underlying patterns in the training data and produce accurate predictions. Similarly, the R^2 score obtained from the testing dataset is 0.904, demonstrating the model's ability to generalize well to unseen data. However, the slightly higher RMSE and MAE values for the testing dataset of 13.115 and 9.613, respectively, indicate that the model has a slightly higher level of prediction errors when the model is applied to unseen data. Figures 3 and 4 show the relationship between the training and testing datasets' actual and predicted values. These graphs show how well the model's predictions match the actual values. Additionally, Figure 7 shows the residual plots for both the training and testing datasets which represent the differences between the actual and predicted values. In the training set, the majority of the residuals fall within the range of 24 kPa to 76 kPa. Similarly, in the testing set, most of the residuals are concentrated within the range of 94 kPa to 124 kPa. Moreover, the histogram of the residuals for both the training and testing sets provides further insights into the errors. It indicates that most of the errors in both datasets are concentrated within the residual range of 22 kPa to -20 kPa.

Random Forest Regression (RFR)

The RFR model was applied to the dataset, which was divided into training and testing sets using a 70:30 split ratio and a random state of 98. Like previous models, the performance of the model was assessed using R^2, RMSE and MAE. The results in Table 3 show that the model achieved a high R^2 score and low RMSE and MAE scores for both the training and testing datasets compared to MLR and AdaBoost. The R^2 score derived from the training dataset is 0.986, indicating that the model can explain 98.6% of the variance in q_u, while the RMSE and MAE values are 5.448 and 4.397, respectively, indicating that the model can effectively capture the underlying patterns in the training data and produce accurate predictions. Similarly, the R^2 score from the testing dataset is 0.914, demonstrating the model's ability to generalize effectively to unknown data. However, the slightly higher RMSE and MAE values for the testing dataset of 12.422 and 9.544, respectively, indicate that the model has a slightly higher level of prediction errors when the model is applied to unseen data. Still, these values are lower than those produced in MLR and AdaBoost models, indicating that this model's ability to predict unknown data is much better than the others'. Figures 3 and 4 show the relationships between the training and testing datasets' actual and predicted values. These graphs show how well the model's predictions match the actual values. Furthermore, Figure 6 presents the residual plots for both the training and testing datasets. These plots illustrate the distribution of the residuals, which stands for the differences between the actual and predicted values. Most of the residuals in both training and testing sets are within the 27 kPa to 130 kPa range and are randomly distributed along the horizontal axis. In the histogram of the training set residuals, it can be seen that most of the errors are within the range of 10 kPa to -10 kPa. In the testing set, it can be noticed that there is a slightly larger number of residuals, within the 12 kPa to -20 kPa region.

Comparing the evaluation metrics shown in Table 3, it is evident that the RFR model outperforms both the MLR and AdaBoost models in terms of performance for both the training and testing datasets, as it yields a higher R^2 score, as well as lower RMSE and MAE values compared to the other two models. These results indicate that the RFR model effectively predicts q_u based on the input variables of N_{60}, depth(m) and PI.

To further compare the models, trend-behavior analysis has been performed as shown in Figures (9), (10) and (11). In Figure (9), the relationship between N_{60} and q_u is displayed while taking into account the effects of depth(m) and PI. N_{60} is represented on the x-axis, while q_u is displayed on the y-axis. The blue data points represent the "actual" marker. These points reflect the true values of the unconfined compressive strength corresponding to the respective N_{60} values. What can be observed from the plot is that as N_{60} increases, q_u increases, which indicates a positive trend and all three models have perfectly captured the trends. The models did not consistently overestimate or underestimate the unconfined compressive strength for the N_{60} values. However, from observing the data points, it is apparent that RFR overall fits in trend and makes better predictions. Figures (10) and (11) display the relationships between depth(m) and PI with q_u. However, no trend can be observed; an increase or decrease in the value of independent variable has no effect on q_u. R^2 and RMSE have to be observed to determine any relationship. Table 4 shows the R^2 and RMSE of the trends, where RFR model has the highest values of R^2 and RMSE for all three trends, except for the PI, where the RMSE is a bit higher compared to the RMSE values of the other models.

Table 4. R^2 and RMSE obtained from trend-behavior analysis

		Linear Regression	**AdaBoost**	**RFR**
N_{60}	R^2	0.906	0.902	0.912
	RMSE	13.001	13.248	12.526
Depth (m)	R^2	0.0787	0.0991	0.125
	RMSE	40.620	40.169	39.591
Plasticity Index	R^2	-0.122	0.00162	-0.419
	RMSE	44.835	42.286	50.419

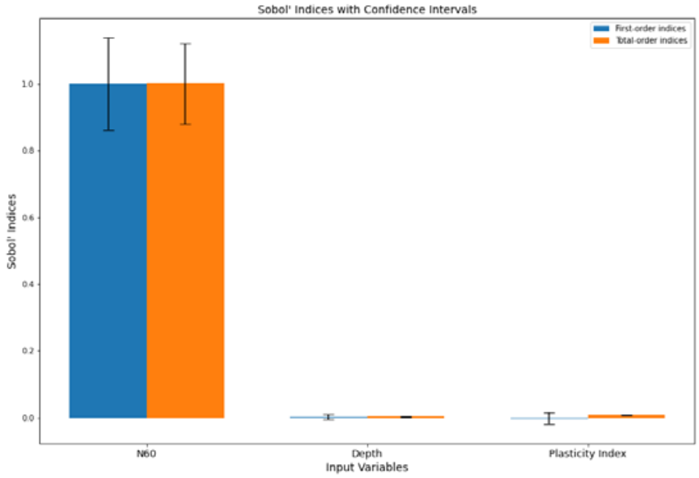

Figure (8): Bar chart indicating Sobol's indices of each independent variable

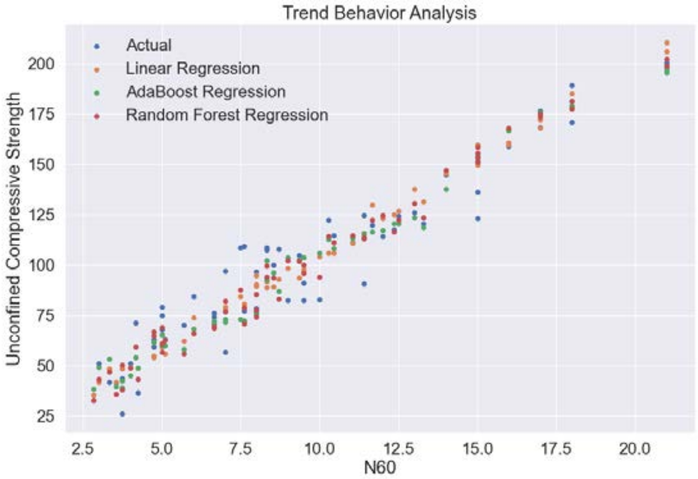

Figure (9): Trend-behavior analysis of q_u *vs.* N_{60}

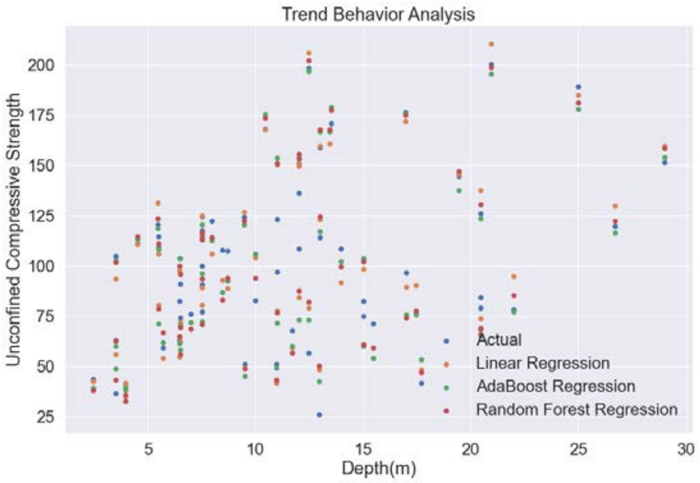

Figure (10): Trend-behavior analysis of q_u *vs.* depth(m)

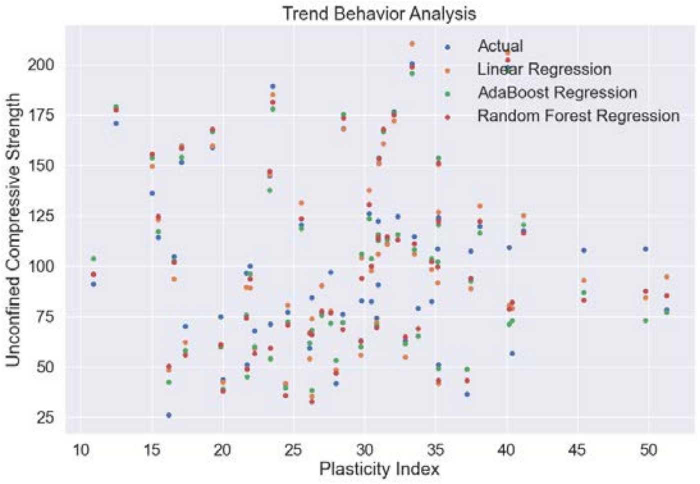

Figure (11): Trend-behavior analysis of q_u *vs.* plasticity index

Thus, based on the evaluation metrics, the findings from the random distribution of residuals and from the trend-behavior analysis, it can be concluded that the RFR model demonstrates its accuracy, precision and superior performance in predicting qu. Its ability to capture the underlying patterns and generalize well to new observations makes it a favorable choice for predicting q_u based on the given input variables.

In order to determine the importance of each parameter in the RFR model, sensitivity analysis has been carried out.

Sensitivity Analysis

Sensitivity analysis (SA) is the study of how multiple sources of uncertainty in a model's input parameters may be connected to the uncertainty in its output (Saltelli et al., 2010). SA is essential for interpreting model behavior and understanding how input parameters affect the model's output. SA helps in determining the elements that have a major impact on the model's predictions by examining the impacts of input-variable uncertainty. In this work, SA is used to evaluate the significance of N_{60}, depth(m) and PI as input parameters in the random-forest regression (RFR) model for predicting q_u.

Local SA (LSA) and global SA (GSA) methods are the two main categories into which SA models can be generally divided. LSA techniques concentrate on how particular variables affect the model's output at particular locations in the parameter space. They normally use partial derivatives and provide information on how sensitive the model is to changes in particular variables. However, LSA approaches frequently overlook the interactions between variables in non-linear models and have difficulties in capturing the overall dependence of the model outcome on the input variables. GSA techniques, on the other hand, provide a thorough analysis of the impacts and interactions of variables throughout the whole parameter space. These methods may use screening or variance decomposition techniques to determine the relative relevance of input variables. GSA techniques can reveal non-linear relationships, measure the significance of individual variables and show cooperative or synergistic effects between variables by taking into account the combined contribution of variables and their interactions (Tosin et al., 2020).

According to Glen and Isaacs (2012), Sobol's method is a popular GSA technique that determines the variance of a model's output parameter by calculating the contribution of each input variable or combination of variables.

Sobol's indices, like the first-order and total-order indices, are used to measure how much the input variables affect the variance of the output. The total-order Sobol's index considers the contributions of the variable itself as well as its interactions with other factors, in contrast to the first-order Sobol's index, which only assesses the contribution of a single input variable. Following are the first-order (S_u) and total-order (ST_u).

$$S_u = \frac{V[E(y|u)]}{V[y]} \qquad (15)$$

$$ST_u = \frac{E[V(y|v)]}{V[y]} \qquad (16)$$

where, V[.] = the unconditional variance operator.

V[.l.] = conditional variance.

E[.] = the mathematical expectation.

E[.l.] = the conditional expectation.

"u" = group of input variables.

"y" = the output parameter of the model.

"V" = variance.

$ST_u \geq S_u$ and the Sobol's indices have a range of 0 to 1, with a higher value indicating a greater contribution of the corresponding variable to the output variance. The first-order Sobol's index of a set of inputs, u, is given by Equation (15), while the total-order Sobol's index of u is given by Equation (16). ST_u gives a thorough measurement of the variable's entire effect, taking into account both its direct impact and its interactions with other factors, as opposed to S_u, which just considers the direct impact of the variable. The sum of the total-order and first-order Sobol's indices always adds up to 1 (Azzini et al., 2021).

Sobol's indices were generated in this work using the open-source SALib software (Iwanaga et al., 2022) and (Herman and Usher, 2017) and the significance of the independent factors in the RFR model's prediction of q_u was established. Figure (8) shows how N_{60}, depth(m) and PI computed Sobol's indices related to the output variability of the RFR model. It was found that N_{60} exhibited the highest Sobol's index, indicating its significant influence on output variability. Specifically, N_{60} had the highest first-order and total-order indices (1.00036511 and 1.0014458, respectively), suggesting its strong influence both individually and in interaction with other variables. On the other hand, depth(m) and PI showed relatively lower Sobol's index values, indicating lesser impacts on the output variance. Their first-order and total-order indices were 0.00265658, 0.00348034 and -0.00204188, 0.00746667, respectively. Therefore, depth(m) and PI have the least relative importance and combined effects on the prediction of q_u.

RFR is a popular and reliable algorithm that has shown its effectiveness across various real-world scenarios. Nevertheless, similar to other machine-learning algorithms used for regression analysis, RFR also has some limitations. These limitations are the difficulty of understanding the exact relationships between features and the target variable, the risk of overfitting, the computational complexity and time constraints and the possible inaccuracy in predicting outcomes outside the range of the training data.

Comparison

In Bangladesh, Equation (1) proposed by Terzaghi & Peck (1967) is mostly used to predict the cohesion of clayey soil. Similarly, Schmertmann (1979), Sowers (1979), Serajuddin & Chowdhury (1996) and Hettiarachchi & Brown (2009) have proposed equations that are also used. Recently, Kumar et al. (2016) proposed a model using random-number generation with high predictability in determining cohesion. All these models are compared with the present study of the testing data and the results are displayed in Table 5.

Table 5. Comparison of the present study with previous models on testing data

Model Name	R^2	RMSE	MAE
Present Study	0.914	12.422	9.544
Terzaghi & Peck (1967)	0.318	47.911	32.221
Schmertmann (1975) and Sowers (1979)	0.282	129.804	104.442
Serajuddin & Chowdhury (1996)	0.311	47.308	61.900
Hettiarachchi & Brown (2009)	0.905	25.999	23.176
Kumar et al (2016)	0.899	34.612	29.023

The actual *vs.* predicted graph in Figure 12 compares the results of the present study with those obtained by using the equations of Terzaghi & Peck (1967), Schmertmann (1979) and Sowers (1979). For the proposed model, a well-fit plot can be observed, where most of the data points are closely aligned with the trend line, which indicates a strong correlation between the predicted and actual values. Conversely, the other equations displayed more scattered data points and lack a well-defined fit. To further evaluate the models, the residual plot in Figure 13 is presented, which shows the residuals for the present study, Terzaghi & Peck (1967), Schmertmann (1979) and Sowers (1979) equations. The residual plot of the present study exhibits a random distribution of data points along the x-axis, which indicates unbiased predictions. However, the majority of the residuals in the Terzaghi & Peck (1967) equation are near and below the x-axis, which shows that there is a tendency for over-prediction. Notably, there is a group of outliers that move away from the trend line and show greater magnitudes within the range of 160 kPa to 178 kPa. The residuals for the Schmertmann (1979) and Sowers (1979) equations mostly demonstrate over-prediction as magnitudes of actual q_u values rise. This suggests that these equations consistently over-estimate q_u. Similar to this, the actual *vs.* predicted graph in Figure 14 shows the findings of the present study with those attained utilizing Serajuddin & Chowdhury (1996) and Kumar et al. (2016) equations. A well-fitted plot for the present study can be seen, that demonstrates a good correlation between the predicted and actual values. Interestingly, a well-fit plot can also be observed for Kumar et al. (2016). However, the plot for Serajuddin & Chowdhury (1996) displays more scattered data points and lacks a well-defined fit. To further evaluate the models, the residual plot in Figure 15 is presented, which shows the residuals for the present study, Serajuddin & Chowdhury (1996) and Kumar et al. (2016) equations. The residual plot of the present study exhibits a random distribution of data points along the x-axis, suggesting unbiased predictions. In contrast, the residuals for Serajuddin & Chowdhury (1996) equation predominantly show over-prediction, with magnitudes increasing as the actual q_u values increase. Notably, there is a cluster of outliers within the range of 160 kPa to 178 kPa, which deviate from the trend line and exhibit larger magnitudes. Regarding Kumar et al. (2016), the residuals also display over-prediction and most data points are clustered within the range of 110 kPa to 125 kPa.

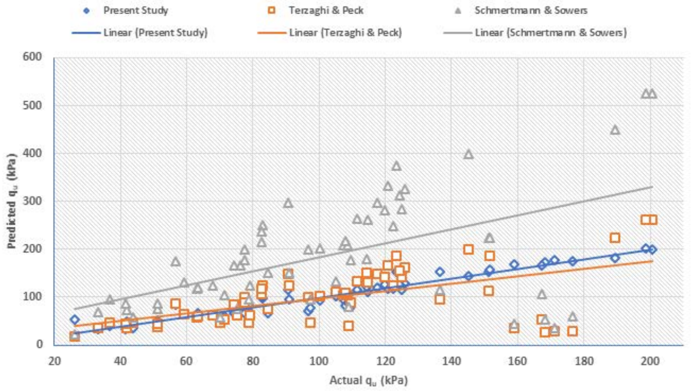

Figure (12): Relationship between the actual *vs.* predicted q_u for testing models devised by Terzaghi & Peck and Schmertmann and Sowers

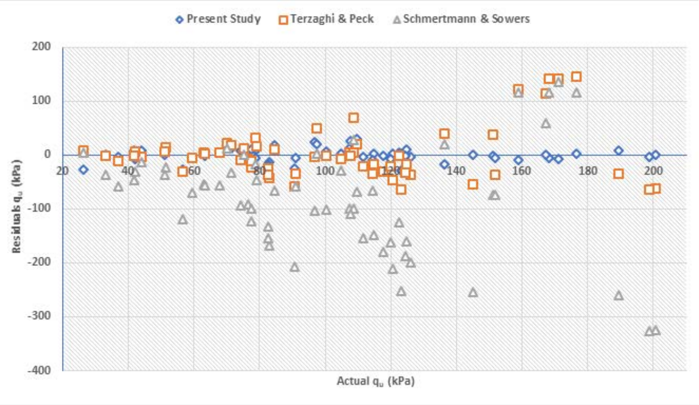

Figure (13): Relationship between residual and actual q_u of Terzaghi & Peck and Schmertmann and Sowers's models

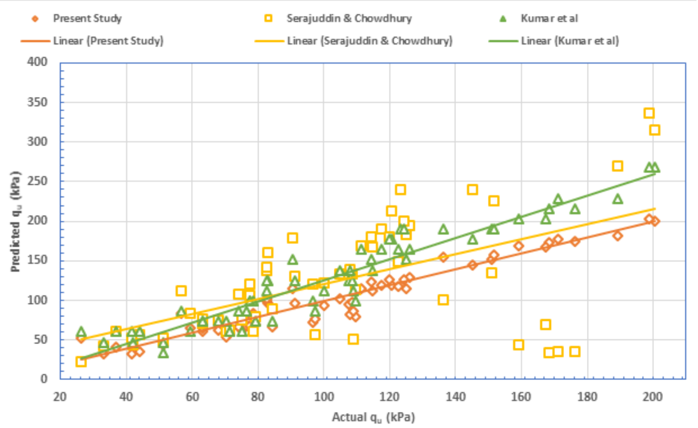

Figure (14): Relationship between the actual *vs.* predicted q_u for testing models devised by Serajuddin & Chowdhury and Kumar et al.

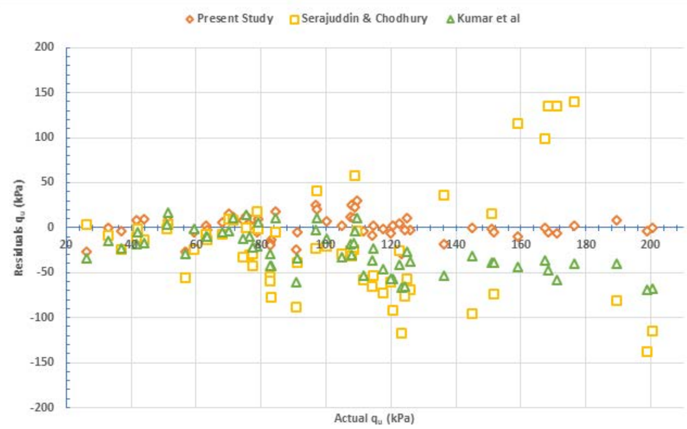

Figure (15): Relationship between residual and actual q_u in Serajuddin & Chowdhury and Kumar et al.'s models

The actual vs. predicted graph in Figure 16 compares the results of the present study with those obtained using the equation of Hettiarachchi & Brown (2009). Compared to the well-fit plot produced in the proposed model, surprisingly, Hettiarachchi & Brown (2009) had also shown a well-fit plot. However, in the residual plot in Figure 17, for Hettiarachchi & Brown (2009), the residuals mainly show under-prediction and only a few data points are over-predicted.

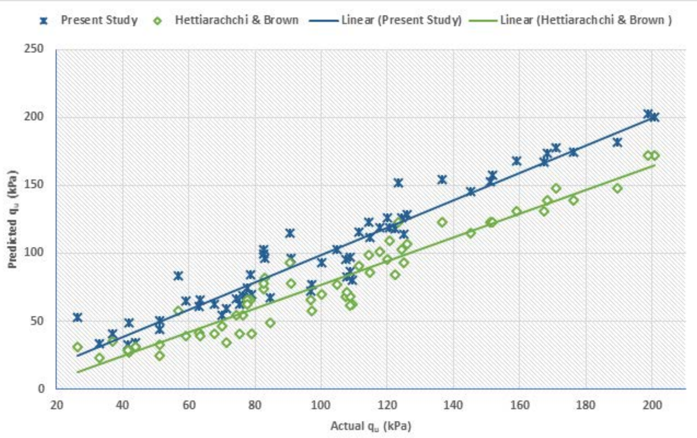

Figure (16): Relationship between residual and actual q_u in Hettiarachchi & Brown's model

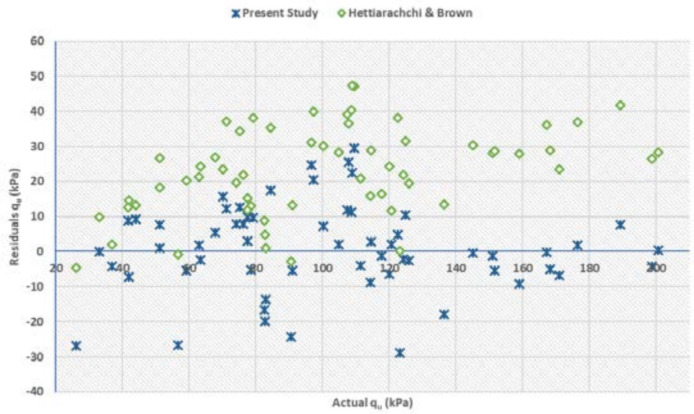

Figure (17): Relationship between residual and actual q_u in Hettiarachchi & Brown's model

In Table 5, the evaluation results of all models are summarized. It is evident that the present study has the highest R^2: 0.914 and the lowest values for both RMSE and MAE: 12.422 and 9.544, respectively, compared to the other models. In contrast, the models from the previous studies show certain limitations. For instance, Terzaghi & Peck (1967) and Hettiarachchi & Brown (2009) models tend to under-predict the unconfined compressive strength, while Schmertmann (1979), Sowers (1979), Serajuddin & Chowdhury (1996) and Kumar et al. (2016) models tend to over-predict it. Over-prediction can be a problem, as it may cause the design of geotechnical structures that are deemed unsafe, such as excavation, tunneling or retaining walls. On the other hand, under-prediction may result in a safer structure, but with higher project costs. Ideally, a reliable model should exhibit residuals that are randomly dispersed around the horizontal axis, which appears to be the case for the model developed in this study.

CONCLUSIONS

Empirical models play an important role in predicting soil properties when laboratory testing is not available. These models provide an economical and time-saving alternative by utilizing in-*situ* tests along with soil properties. In this study, regression analysis was employed to predict the unconfined compressive strength (q_u) (kPa) of clayey soil in Bangladesh using corrected SPT-N (N_{60}), depth (m) and Plasticity Index (PI) of collected samples. Among the three models examined, random-forest regression (RFR) outperformed AdaBoost and Multiple Linear Regression (MLR) in terms of the evaluation metrics, random distribution of residuals and trend-behavior analysis. Thus, the RFR model demonstrated its accuracy, precision and superior performance in predicting q_u. The RFR model has also demonstrated the highest accuracy while exhibiting neither significant over-prediction nor under-prediction of the data. Furthermore, sensitivity analysis highlighted the significant influence of N_{60} in the model's predictions, while depth(m) and PI had relatively lower importance. The implementation of RFR could be facilitated through the development of a web-based application, making it easily accessible to users *via* smartphones or computers.

Furthermore, the performance of the RFR model was compared with those of the previously existing models proposed by Terzaghi & Peck (1967), Schmertmann (1979), Sowers (1979), Serajuddin & Chowdhury (1996), Hettiarachchi & Brown (2009) and Kumar et al. (2016). Through comprehensive evaluation metrics, the RFR model consistently outperformed these models, demonstrating its superior predictive capability.

By employing the predictions generated by this RFR model in geotechnical designs, engineers can ensure safe, reliable and cost-effective outcomes. The use of this model can assist in efficient decision-making processes and contribute to the advancement of geotechnical engineering practices.

Acknowledgments

Our special gratitude is due to the Prosoil Foundation Consultants for their continuous support in conducting this research. Also, cordial thanks are extended to Tahsina Alam, Lecturer, CEE, IUT, for her most valuable advice throughout this research.

Availability of Data

All the data used and analyzed in this study can be obtained from the corresponding author upon request.

REFERENCES

ASTM D 2166. “Standard test method for unconfined compressive strength of cohesive soil”. West Conshohocken, PA.

ASTM D 2487-11. (2011). “Standard practice for classification of soils for engineering purposes (unified soil classification system)”. Annual Book of ASTM Standards, ASTM International, West Conshohocken, PA.

Azzini, I., Mara, T.A., and Rosati, R. (2021). “Comparison of two sets of Monte Carlo estimators of Sobol’s indices”. Environmental Modeling & Software, 144, 105167. https://doi.org/10.1016/J.ENVSOFT.2021.105167

Bengfort, B., Danielsen, N., Bilbro, R., Gray, L., McIntyre, K., Richardson, G., …, and Keung, J. (2018). “Yellowbrick V0.6”. doi:10.5281/zenodo.1206264

Breiman, L. (2001). “Random forests”. Mach. Learn, 45, 5-32.

Das, B.M. (2010). “Principles of geotechnical engineering”. Stamford, Conn.: Cengage Learning.

Drucker, H. (1997). “Improving regressor using boosting”. Douglas H. Fisher, Jr. (Ed.). Proc. of the 14th Int. Conf. on Machine Learning (107-115), Morgan Kaufmann.

Eberly, L.E. (2007). “Multiple linear regression”. Topics in Biostatistics, 165-187.

Freund, Y., and Schapire, R. (1996). “Experiment with a new boosting algorithm”. Proc. of the 13th Int. Conf. on Machine Learning (148-156), Bari, Italy.

Glen, G., and Isaacs, K. (2012). “Estimating Sobol-sensitivity indices using correlations”. Environmental Modeling & Software, 37, 157-166. https://doi.org/10.1016/J.ENVSOFT.2012.03.014

Guo, L., Chehata, N., Mallet, C., and Boukir, S. (2011). “Relevance of airborne lidar and multi-spectral image data for urban scene classification using random forests”. ISPRS J. Photogramm. Remote Sens. 66, 56-66.

Herman, J., and Usher, W. (2017). “SALib: An open-source Python library for sensitivity analysis”. Journal of Open Source Software, 2 (9). doi:10.21105/joss.00097

Hettiarachchi, H., and Brown, T. (2009). “Use of SPT blow counts to estimate shear-strength properties of soils: Energy-balance approach”. Journal of Geotechnical and Geoenvironmental Engineering, 135 (6), 830-834. https://doi.org/10.1061/(asce)gt.1943-5606.0000016

Hossain, A., Alam, T., Barua, S., and Rahman, M.R. (2021). “Estimation of shear-strength parameter of silty sand from SPT-N60 using machine-learning models”. Geomechanics and Geoengineering, 17:6, 1812-1827, DOI: 10.1080/17486025.2021.1975048

Iwanaga, T., Usher, W., and Herman, J. (2022). “Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses”. Socio-environmental Systems Modeling, 4, 18155. doi:10.18174/sesmo.18155

Jay, A., Sivakugan, N., and Das, B.M. (2014). “Correlations of soil and rock properties in geotechnical engineering”. Development in Geotechnical Engineering, New Delhi, India: Springer.

Kumar, R., Bhargava, K., and Choudhury, D. (2016). “Estimation of engineering properties of soils from field SPT using random-number generation”. INAE Letters, 1 (3-4), 77-84. https://doi.org/10.1007/s41403-016-0012-6

Larbi, R., Benyoussef, H., Morsli, M., Bensaibi, M., and Bali, A. (2019). “Influence of database size on artificial neural network results for the prediction of compressive strength of concretes containing reclaimed asphalt pavement”. Jordan Journal of Civil Engineering, 13 (4).

Mayne, P.W., Christopher, B.R., and DeJong, J. (2001). “Manual on sub-surface investigations”. Nat. Highway Inst. Sp. Pub. FHWA NHI-01-031. Fed. Highway Administ., Washington, DC.

Michael L. Waskom. (2021). “Seaborn: Statistical data visualization”. Journal of Open Source Software, 6 (60), 3021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). “Scikit-learn: Machine learning in Python”. Journal of Machine Learning Research, 12, 2825-2830.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. (2015). “Machine-learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines”. Ore Geology Reviews, 71, 804-818. https://doi.org/10.1016/j.oregeorev.2015.01.001.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sánchez, J.P. (2012b). “An assessment of the effectiveness of a random-forest classifier for land-cover classification”. ISPRS J. Photogramm. Remote Sens., 67, 93-104.

Saadat, M., and Bayat, M. (2022). “Prediction of the unconfined compressive strength of stabilized soil by adaptive neuro fuzzy inference system (ANFIS) and non-linear regression (NLR)”. Geomechanics and Geoengineering, 17 (1), 80-91. https://doi.org/10.1080/17486025.2019.1699668

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). “Variance-based sensitivity analysis of model output: Design and estimator for the total sensitivity index”. Computer Physics Communications, 181 (2), 259-270. https://doi.org/10.1016/J.CPC.2009.09.018

Schapire, R. (1990). “The strength of weak learnability”. Machine Learning, 5 (2), 197-227.

Schmertmann, J.H., and Palacios, A. (1979). “Energy dynamics of the SPT”. Journal of Geotechnical Engineering Division, 105, 8, 909-926.

Serajuddin, M., and Chowdhury, M.A. (1996). “Correlation between standard penetration resistance and unconfined compression strength of Bangladesh cohesive deposits”. Journal of Civil Engineering.

Skempton, A.W. (1986). “Standard penetration test procedures and the effects in sands of overburden pressure, relative density, particle size, ageing and over-consolidation of sands”. Geotechnique, 36 (3), 425-447.

Solomatine, D.P., and Shrestha, D.L. (2004). “AdaBoost.RT: A boosting algorithm for regression problems”. 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2, 1163-1168.

Sowers, George F. (1979). “Introductory soil mechanics and foundations: Geotechnical engineering”. New York: Macmillan.

Tabarsa, A., Latifi, N., Osouli, A., and Bagheri, Y. (2021). “Unconfined compressive-strength prediction of soils stabilized using artificial neural networks and support vector machines”. Frontiers of Structural and Civil Engineering, 15 (2), 520-536. https://doi.org/10.1007/s11709-021-0689-9

Terzaghi, K., and Peck, R.B. (1967). “Soil mechanics in engineering practice”. 2nd Edn., Wiley, New York.

Tosin, M., Côrtes, A.M., and Cunha, A. (2020). “A tutorial on Sobol’s global-sensitivity analysis applied to biological models”. Networks in Systems Biology, 93-118. https://doi.org/10.1007/978-3-030-51862-2_6

Widodo, S., Ibrahim, A., and Hong, S. (2012). “Analysis of different equations of undrained shear-strength estimations using Atterberg limits on Pontianak soft clay”. Challenges of Modern Technology, 3 (3).