Journal of GeoEngineering, Vol. 19, No. 2, pp. 083-094, June 2024
http://dx.doi.org/10.6310/jog.202406_19(2).4

USE OF CPT AND OTHER PARAMETERS FOR ESTIMATING SPT-N VALUE USING OPTIMISED MACHINE LEARNING MODELS

Mehedi A. Ansary [1*] and Mushfika Ansary [2]

ABSTRACT

In this study, the SPT-N value of soil is modeled using five cutting-edge machine learning procedures, comprising multilayer perceptron artificial neural networks, random forests, ridge regression, support vector regressors, and extremely gradient boosting. The hyper-parameters of these algorithms are optimized utilizing the randomized search cross-validation (RSCV) algorithm. The mean average error, root mean square error, R-squared, and variance accounted for values are applied as evaluation indicators to assess the efficiency of optimized machine learning procedures on a dataset with 1113 data. The comparison shows that the RSCV approach is effective in the hyper-parameter tuning and that the optimized machine learning procedures have tremendous prospects to evaluate the SPT-N value of soils. Among the five Optimized Machine Learning Models (OMLs) used for the testing dataset, Random Forrest (RF) and Support Vector Regression (SVR) display excellent performances ($R^2 = 0.9205$ and 0.8956, respectively). The depth and CPT cone resistance are the variables that have the greatest influence on determining the SPT-N value of soil with a variable importance score of 24.06% and 23.61%, respectively. The performance of RF and SVR is compared with the existing models. It is found that the OML models such as RF and SVR outperform the existing models.

Key words: Cone penetration test (CPT), SPT-N value of soils, optimized machine learning methods, randomized search cross-validation (RSCV) algorithm.

1. INTRODUCTION

Due to their lower costs, in-situ investigative techniques like the standard penetration test (SPT) and cone penetration test (CPT) are widely used in projects of all sizes. The fundamental drawback of these approaches is that they rely on correlations for engineering design rather than explicitly measuring any soil parameters. To estimate soil parameters utilizing in situ techniques using information gathered in various locations, many correlations have been devised. However, because there aren't any other affordable alternatives, their use has spread rapidly.

Numerous scholars have looked into the relationships between SPT-N and CPT parameters under various geotechnical conditions. For instance, based on a sizable dataset, Robertson and Campanella (1983) established a link between SPT-N values and cone tip resistance (q_c) for sands. They developed an empirical equation that is frequently used to calculate tip resistance from SPT-N data. Correlations between SPT-N and CPT parameters are more complicated in cohesive soils because of the impact of variables including soil type, plasticity, and sensitivity. Mayne and Kulhawy (1982) concentrated on correlations between SPT-N and other factors, such as friction angle (ϕ) for cohesive soils. To address the impact of regional soil conditions, regional correlations between SPT-N and CPT parameters have been constructed for particular places. For instance, researchers have created correlations particular to regional soil types and geotechnical features in studies carried out in various parts of the world. According to Suzuki *et al.* (1998), cone penetration resistance is correlated with soil physical characteristics, SPT-N value, and shear wave velocity. Robertson's proposed soil behavior type index I_c, fines content, and mean grain size are used as indexes to categorize soil types. They concluded that the fines concentration and mean grain size are well correlated with the soil behavior type index I_c and the CPT-SPT resistance ratio (q_t/N) varies not only with soil type but also with SPT-N value or CPT q_t-value.

In Adapazari, Turkey, Kara and Gündüz (2010) investigated the relationship between CPT and SPT. They examined data from 65 SPT boreholes and 47 CPT locations while taking into account the varied soil composition. The correlation coefficients were slightly lower than those reported in the literature after the N values were adjusted for energy efficiency. When analyzing filtered data, the coefficients became more accurate. For the sands of Oil sands, researchers Elbanna *et al.* (2011) examined the relationship between SPT and CPT. Reviewing and evaluating the precision of these correlations, particularly for Muskeg River Mine (MRM) tailings sand, was their main objective. They also contrasted the relationships with measurements from other oil sands tailings sites and the average particle size (D_{50}). The investigation was conducted at the MRM oil sand mine near Fort McMurray in northern Alberta. In a silty sand deposit in Egypt, Shahien and Albatal (2014) also looked into the relationship between SPT and CPT to determine the correlation between the two tests, considering grain sizes, fines content, and soil behavior type index. The study highlighted the cost-effectiveness and ease of use of CPT while noting the consistency and standardization of SPT. The N value in SPT was discovered to be influenced by elements such as borehole diameter, water level, hammer type, and lifting procedures. When establishing the soil behavior type index for CPT, the normalised tip resistance and friction ratio was the main considerations. Stratification, soil type, grain characteristics, soil density, and the

Manuscript received November 21, 2023; revised April 10, 2024; accepted May 13, 2024.

[1*] Professor (corresponding author), Department of Civil Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh (e-mail: ansary@ce.buet.ac.bd).

[2] Graduate Student, Department of Architecture, University of Asia-Pacific, Dhaka, Bangladesh.

presence of gravel had an impact on the association between SPT and CPT. In Goluck, Turkey, Asci *et al.* (2014) looked into the relationship between SPT and CPT data. They looked at several soil types and saw that the correlation coefficients varied. It has been observed that the correlation coefficient for sandy silt in their particular research region was highest. Based on a database of in-situ tests in China, Zhao and Cai (2015) used statistical and regression methods to evaluate the SPT-CPT association. There are two suggested correlation equations for the SPT-N value and CPT cone tip resistance. For the evaluation of the liquefaction potential, the equations are used. SPT-CPT correlation is examined for three effect parameters; including soil type, mean particle size, and fine content. Similarly, Demir and Sahin (2022) investigates and compares the performance of three tree-based Machine Learning (ML) methods, Canonical Correlation Forest (CCF), Rotation Forest (RotFor), and Random Forest (RF), for predicting the liquefaction potential of soils based on the cone penetration test (CPT) case history datasets collected from previous studies in Turkey. Demir and Sahin (2023) also presents the prediction of soil liquefaction from the SPT dataset by using relatively new and robust tree based ensemble algorithms, namely Adaptive Boosting, Gradient Boosting Machine, and eXtreme Gradient Boosting (XGBoost). A total of 620 SPT records with 12 parameters collected from two major Turkish earthquakes in 1999 are considered for this study.

Jarushi *et al.* (2015) examined the relationship between SPT and CPT in various sandy soils in Florida, USA. The study established empirical relationships for evaluating soil performance by using data from initiatives funded by the Florida Department of Transportation (FDOT). The results showed that in sandy soils, SPT-N value, CPT tip resistance q_c and sleeve resistance f_s had a positive linear association. The alteration of the soil's permeability, compressibility, fines concentration, and q_c/N ratio all had an impact. Zhao *et al.* (2021) employed a database of about 900 data pairs from 230 sites in the South Island of New Zealand that were co-located cone penetration test (CPT) soundings and boreholes with standard penetration tests (SPT). This study evaluates the applicability of several SPT-CPT correlations that are already in existence. Correlations based on the soil behavior type index (I_c) using CPT data as well as simple linear SPT-CPT correlations for various soil types were evaluated. For an alluvial soil deposit in Dhaka, Arifuzzaman and Anisuzzaman (2022) sought to present connections between the SPT-N value, cone tip resistance (q_c), sleeve friction resistance (f_s), soil behavior index (I_c), and mean particle size (D_{50}). It is discovered that the coarser soil layers exhibit a coefficient of correlation (R^2) of 0.7106, which imply a reliable association for the relationship between equivalent SPT-N_{60}-value and SPT-N_{60}-value. Additionally, there is a high correlation between cone tip resistance (q_c) and SPT-N_{60}-value that is extremely close to the Meyerhof correlation that was proposed. Hasan (2023) aimed to establish empirical correlations for the SPT-N value and soil unit weight with the CPT parameters for the Dhaka Metropolitan Development Plan (DMDP) area in Bangladesh. The collected data includes SPT-N values, CPT parameters (cone tip resistance, sleeve friction), unit weight of soil, soil type, moisture content, *etc.* Considering depth, cone tip resistance and side friction as independent variables and SPT-N value as the dependent variable, a multiple linear regression equation has been developed, where obtained R^2 is only 0.6399. Table 1 presents summary of all those above findings. In addition works of several other researchers' such as Akca (2003), Lingwanda *et al.* (2015), Dos Santos and Bicalho (2017), Khodaparast *et al.* (2020) and Khan *et al.* (2022) have been presented in Table 1.

Tarawneh (2014) carried out a study on silty sand to sandy silt soils in the UAE. The study concentrated on creating models to forecast N-values from CPT data. The study used SR and MLR methodologies and comprised 66 CPT-SPT couples. The rod energy ratio was used to normalize the SPT-N values because the CPT data fell inside defined soil behavior areas. In comparison to MLR, symbolic regression models also demonstrated improvement. Fernando *et al.* (2021) examined using an ANN to predict SPT values based on CPT data and soil physical parameters on cohesive soils with and without data normalization. The tip resistance, sleeve resistance, effective soil overburden pressure, liquid limit, plastic limit, and percentage of sand, silt, and clay are the input data used in this study. The outcomes demonstrated that the ANN could successfully predict events on networks with and without data standardization. The ANN without data normalization displayed a lower error value in this investigation than the ANN with data normalization, it was discovered.

It is observed from the above studies that very limited researches have been undertaken using machine learning models to estimate the SPT-N value of soil from CPT parameters. Only Artificial Neural Networks is mainly used for particular type of soils. Other machine learning models like Random Forest (RF), Support Vector Machine (SVM), Ridge Regression, Extreme Gradient Boosting (XGB) *etc.* have not been used to predict SPT-N value from CPT parameters. However, there are still a number of issues that require correct resolution before applying the ML techniques for estimating SPT-N value from CPT parameters, such as: (1) The viability of other methods has not been extensively investigated, and only a small number of sophisticated ML algorithms have been used in SPT-N value estimate; (2) before using ML algorithms on datasets, it is essential to correctly adjust their hyper-parameters; and (3) there is still a need for an organized, thorough comparison of the existing ML techniques. The efficacy of modern ML algorithms when used to estimate SPT-N value may differ significantly, necessitating further research.

This paper employs five optimized machine learning (OML) techniques in a comparative manner, utilizing the programming language Python, to fill a gap in the existing papers concerning the estimation of SPT-N value. The extreme gradient boosting (XGB), multilayer perceptron artificial neural network (MLPANN), RIDGE regression (RIDGE), random forest (RF), and support vector machine (SVM) are the five ML algorithms are used for this purpose. In this study, a recently collected 54 CPT-SPT collocated points having 1113 dataset have been used. The depth of the soil sample collected (D), moisture content of the soil sample (MC), fine content of the soil sample (FC), cone tip resistance (q_c) and cone local resistance (f_s) are the input features of the algorithms and will be discussed in the data processing section. Mean absolute error (MAE), root mean square error (RMSE), R-squared value (R^2) and variance accounted for (VAF) are used as performance indicators for evaluating the efficiency of the five ML methods. Investigations on the comparative significance of important input variables for SPT-N value are also conducted. The limitations of many existing methods are resolved by the current study, which can more effectively estimate the SPT-N value of soil than is currently done.

Table 1 Correlations developed between SPT-N value of soil and CPT parameters in the past years

Author	Correlation equations
Schmertmann (1970)	q_c/N = 3.5 (R^2 = 0.24) for clean sand where q_c in kg/cm^2
Robertson *et al.* (1986)	$(q_c/P_a)/N_{60} = 5$ (R^2 = 0.38) for clean sand
Suzuki *et al.* (1998)	$(q_c/P_a)/N_{60} = 0.0026FC^2 - 0.263FC + 12.34$; $0 \le N < 10$ $(q_c/P_a)/N_{60} = 0.00085FC^2 - 0.120FC + 8.733$; $10 \le N < 30$ $(q_c/P_a)/N_{60} = 0.001FC^2 - 0.059FC + 5.59$; $30 \le N$, $FC \le 20$
Akca (2003)	$q_c/N_{60} = 0.47$ (R^2 = 0.31) for clean sand $q_c/N_{60} = 0.55$ (R^2 = 0.34) for silty sand $q_c/N_{60} = 0.32$ (R^2 = 0.48) for sandy silt where q_c in MPa (unfiltered data)
Asci *et al.* (2014)	$q_c = 7.187 \exp(-0.4827 N_{60}) + 1.938 \exp(0.00989 N_{60})$ (R^2 = 0.8005) for sandy silts
Shahien and Albatal (2014)	$(q_c/P_a)/N_{60} = 17.13 \times (D_{50})^{0.26}/[(N_{60})^{0.49} \times (FC)^{0.27}]$ (R^2 = 0.52) for silty sands
Jarushi *et al.* (2015)	q_c= 0.291N + 2.430 (R^2 = 0.60) for fine sand (SP) q_c = 0.121N + 5.086 (R^2 = 0.35) for silty fine sand (SM) q_c = 0.155N + 7.260 (R^2 = 0.11) for fine sand with silt (SP-SM)
Lingwanda *et al.* (2015)	$(q_c + f_s) = 0.161 N_{60} + 7.87$ (R^2 = 0.604) filtered data for sandy soil
Zhao and Cai (2015)	$N = [0.02\rho_c - 3.48D_{50} - 0.1\alpha_p + 2{:}53] \times q_c$ where ρ_c = 3%, if $\rho_c \le$ 3%; and ρ_c 15, if $\rho_c \ge$ 15%; D_{50} = 0.03, if $D_{50} \le$ 0:03 and D_{50} = 0.1, if $D_{50} \ge$ 0.10; α_p = 0 for silt; and α_p = 1 for silty sand (R^2 = 0.58 to 0.78)
Dos Santos and Bicalho (2017)	$q_c/N_{60} = 0.44$ (R^2 = 0.86) for Vittoria sand where q_c in MPa
Khodaparast *et al.* (2020)	$q_c = 0.245N_{60} + 5.861$ (R^2 = 0.24) for clean sand $q_c = -4.609 + 5.823\ln(N_{60})$ (R^2 = 0.36) for silty sand $q_c = 1.678 + 4.716/(N_{60})$ (R^2 = 0.67) for sandy silt
Arifuzzaman and Anisuzzaman (2022)	$q_c = 4N_{60} * 0.098$ (MPa) (R^2 = 0.6758) $f_s = 4.66N_{60} + 70.2$ (R^2 =0.6408) for all soils
Khan *et al.* (2022)	$q_c = 0.52N_{60} + 9.36$ (R^2 = 0.45) for gravelly sand $f_s = 3.13N_{60} + 116.13$ (R^2 = 0.35) for gravelly sand
Hasan (2023)	$N = 0.877D + 0.706q_c + 22.835f_s + 0.834$ (R^2 = 0.6399) where D is the depth in m, for all soils

2. METHODOLOGY

The SPT-N value of soil and its affecting variables are investigated in this study using five ML algorithms. The hyper-parameters of these five algorithms are optimized utilizing the randomized search cross-validation (RSCV) algorithm. The five ML algorithms and RSCV are briefly described in this section.

2.1 Extreme Gradient Boosting (XGB)

XGB is a standard machine learning technique known for its efficiency and efficiency in both regression and classification tasks. It uses a more sophisticated version of the gradient boosting architecture that utilizes an optimized algorithm and various regularization techniques to improve model accuracy and generalization. XGB model works as follows: (a) Gradient Boosting Framework: The foundation of XGB is the gradient boosting framework, which creates a powerful ensemble model by integrating a number of weak predictive models (usually decision trees). The goal of gradient boosting is to continuously train models that minimize the errors produced by the past models; (b) Optimization Algorithm: XGB employs a highly optimized algorithm to efficiently build the ensemble of weak models. The algorithm leverages parallel processing and tree pruning techniques to decrease memory utilization and gear up the training process; (c) Regularization Techniques: XGB incorporates several regularization procedures to avoid overfitting and enhance the quality of generalization. Regularization methods include shrinkage (learning rate), this regulates how much each tree contributes to the total forecast, which add penalties to the model's complexity; (d) Tree Construction: XGB uses decision trees as base learners. It constructs trees in a greedy manner by iteratively splitting the data based on specific criteria, such as reducing the loss or maximizing the information gain. The tree construction process is guided by optimization objectives and constraints to find the best splits and create trees that capture important patterns in the data; (e) Feature Importance: XGB provides a measure of feature importance, which indicates the relative importance of each input feature in the prediction process. Based on how often a feature is utilized to divide the data among all the ensemble trees, feature significance scores are determined; and (f) Hyperparameter Tuning: A variety of hyperparameters are available in XGB that can be adjusted to enhance the efficiency of the model. XGB offers a comprehensive hyperparameters that can be tweaked to optimize the model's performance. Hyperparameters govern various aspects of the algorithm, such as the learning rate, regularization factors, tree depth, subsampling ratio, *etc.*

XGB has gained popularity for its exceptional performance in machine learning competitions and real-world applications. It is capable of handling large datasets, capturing complex relationships, and providing accurate predictions. However, proper hyperparameter tuning and careful validation are important to achieve optimal results and prevent overfitting.

2.2 Multilayer Perceptron (MLP) Artificial Neural Network

MLP artificial neural networks are a common tool for classification and regression among other machine learning problems. It draws inspiration from the design and operation of the human brain. The MLP is made up of numerous layers of neurons, which are interconnected nodes. Usually, there are three different types of layers: (a) A set of features or attributes may be received by the input layer as input data; (b) Layers between the input and output layers are considered hidden layers. Each neuron in a hidden layer takes information from the layer below and processes it before sending the results to the layer above it. The network can learn intricate representations and patterns in the input thanks to the hidden layers; and (c) The network's ultimate output is produced by the output layer. The type of task determines how many neurons are present in the output layer. For example, in a regression task, there is typically a single neuron for predicting a continuous value, while in a classification task; there is one neuron per class for predicting class probabilities.

The neurons in an MLP are connected by weighted connections, which determine the strength and importance of the information flowing between neurons. Each neuron generates an output by applying an activation function to the weighted sum of its inputs. In order to reduce the discrepancy between the projected outputs and the actual targets, the MLP modifies the weights of the connections during training. The weights are iteratively updated based on the computed error in order to do this using an optimization approach like gradient descent. Finding the ideal collection of weights is the goal in order to reduce prediction errors and increase the network's capacity to generalize to new inputs.

A non-linear function, such as the hyperbolic tangent (tanh) function, rectified linear unit (ReLU) function, sigmoid function, may be employed as the activation function in an MLP. The network can learn intricate connections between the inputs and outputs thanks to non-linear activation functions. MLPs are known for their ability to approximate complex functions and learn non-linear forms in the data. However, they can be susceptible to over-fitting if not suitably regularized or if the network architecture is not appropriately designed. MLPs have become popular in various domains due to their flexibility, scalability, and effectiveness in handling complex datasets. They have been effectively used in a variety of machine learning applications, including time series analysis, natural language processing, image identification, and many more.

2.3 Random Forest (RF)

An ensemble learning technique called RF combines the predictions of multiple decision trees to get predictions that are more accurate. Both classification and regression tasks can be accomplished with this flexible and effective technique. RF method works as follows: (a) Ensemble Learning: Random Forest belongs to the family of ensemble learning methods, which combine multiple individual models to make collective predictions. In the case of Random Forest, the individual models are decision trees; (b) Decision Trees: Decision trees are predictive models that learn a series of hierarchical if-else rules based on the features of the data. Each decision tree makes predictions by following a track from the root node to a leaf node, where the final prediction is made; (c) Randomness and Diversity: Random Forest introduces randomness and diversity into the modeling process. Randomness is introduced by randomly selecting subsets of the original data for training each decision tree (bootstrap aggregating or bagging). Diversity is achieved by arbitrarily picking a subset of features for each split in the decision tree; (d) Voting and Aggregation: When making predictions, every decision tree in the Random Forest independently forecasts the target variable. For classification tasks, the class with the majority of votes among the trees is selected as the final prediction. For regression tasks, the average or median of the predicted values from all the trees is taken as the final prediction; (e) Feature Importance: Random Forest provides a measure of feature importance, indicating the comparative significance of each input feature in making predictions. The importance is calculated based on how much each feature contributes to the reduction of impurity or variance across all the decision trees; and (f) Hyper-parameter Tuning: There are several hyper-parameters in Random Forest that can be tweaked to enhance efficiency. The maximum depth of each tree, the quantity of trees in the forest, the amount of features taken into account for each split, *etc.* are some significant hyper-parameters.

Random Forest is known for its robustness, scalability, and capacity for handling high-dimensional data. It is less prone to over-fitting compared to individual decision trees and often yields better performance in terms of accuracy. However, like any algorithm, proper hyper-parameter tuning and careful validation are crucial to achieve optimal results.

2.4 Ridge Regression (RIDGE)

The Ridge method, also known as Ridge regression or Tikhonov regularization, is a regularization technique used in linear regression models. It helps to address the issue of multi-collinearity (high correlation between features) and reduce the impact of less important features on the model. Ridge method works as follows: (a) Objective function: The objective function for linear regression includes a penalty term thanks to the Ridge technique. The objective function attempts to reduce the quantity of squared residuals, which gauges the difference between expected and observed values. The penalty term is the L2 norm (squared values) of the coefficients multiplied by a regularization parameter (lambda or alpha); (b) L2 regularization: The L2 norm penalty in the Ridge method is the sum of the squared values of the coefficients. This penalty term discourages extreme values of the coefficients and encourages smaller, more spread-out values. It aids in managing the model's complexity and lessens the effects of multi-collinearity; (c) Shrinkage of coefficients: The Ridge method shrinks the coefficients towards zero, reducing their magnitudes. The amount of shrinkage is controlled by the regularization parameter. A larger regularization parameter results in more aggressive shrinkage and smaller coefficients; (d) Multi-collinearity handling: Ridge regression is particularly useful when dealing with datasets that have multi-collinearity, where features are highly correlated. By shrinking the coefficients, Ridge regression reduces the impact of highly correlated features and prevents them from dominating the model; (e) Bias-variance tradeoff: The Ridge method helps in striking a balance between bias and variance in the model. Increasing the regularization parameter increases the bias of the model but reduces its variance, while decreasing the regularization parameter has the opposite effect. Proper tuning of the regularization parameter is important to find the right balance for the given dataset; and (f) Hyper-parameter tuning: Ridge regression involves tuning the regularization parameter (lambda or alpha) to optimize the model's efficiency. Cross-validation techniques can be utilized to evaluate different values of the regularization parameter and select the optimal one.

The Ridge method is widely used in various domains to handle multi-collinearity and improve the stability and generalization of linear regression models. By shrinking the coefficients, in the presence of strongly linked predictors, it aids in determining and ranking the most important features.

2.5 Support Vector Regression (SVM)

Notable supervised machine learning methods for classification and regression include Support Vector Machine (SVM). It is particularly effective in handling complex datasets with clear margin or separation between different classes. A SVM model works as follows: (a) Basic concept: In a high-dimensional feature space, SVM seeks to identify the best hyperplane for classifying the data points. In binary classification, SVM seeks to find a hyperplane that maximizes the margin between each class's nearest data points. Support vectors are utilized to express these nearby data points; (b) Feature space and hyperplane: The feature space refers to the transformed space where the input data points are mapped using a kernel function. A hyperplane which most effectively divides the data points has been identified by SVM. In two dimensions, the hyperplane is a line, while in higher dimensions, it becomes a hyperplane; (c) Margin and support vectors: The margin is the gap between the nearest data points for each class and the hyperplane. SVM seeks to increase this margin, as a larger margin usually implies better generalization and robustness to new data. The data points represent the support vectors that lie on the margin or are misclassified. These points influence the position and orientation of the hyperplane; (d) Linear and non-linear separation: SVM can handle both linearly separable and non-linearly separable data. For linear separation, a linear kernel (*e.g.,* the linear function) is used to create a linear decision boundary. For non-linear separation, SVM utilizes kernel functions (for instance, a polynomial or a radial basis function) to translate the data into a space with more dimensions, where a linear separation is possible; (e) Training process: Given a labeled training dataset, SVM determines the optimal hyperplane by solving an optimization problem. The optimization problem involves identifying the hyperplane that increases the margin while minimizing the classification errors. The solution is obtained by solving a quadratic programming problem or through convex optimization techniques; and (f) Prediction: Once the optimal hyperplane is determined, SVM can predict the class label of new, unseen data points by evaluating which side of the hyperplane they fall on.

Key characteristics and considerations of the SVM models are: (a) Versatility: SVM can handle both linear and non-linear classification tasks; (b) Robustness: SVM is less prone to overfitting due to the margin maximization objective; (c) Kernel functions: The choice of kernel function can significantly impact SVM's performance and ability to handle complex datasets; and (d) Model complexity: The complexity of the SVM model depends on the number of support vectors, which affects training and prediction time.

SVM is frequently utilized in many fields, including image classification, text classification, and bioinformatics. Effectively separate classes and handling of high-dimensional data makes it a valuable tool in machine learning.

2.6 Randomized Search Cross-Validation (RSCV)

The term "Randomized Search CV" refers to cross-validation. It is a method for selecting models and tweaking hyperparameters in machine learning. Hyperparameters are settings made by the user prior to training a machine learning model rather than ones that are learned from the data. The rate of learning, the quantity of hidden layers in a neural network, or the regularization strength is a few examples of hyperparameters.

To determine the ideal set of hyperparameters for a particular model, Randomized Search CV combines cross-validation and random sampling of hyperparameters. It operates by selecting a subset of hyperparameter combinations at random from a predetermined search space and assessing their effectiveness using cross-validation. Here is a detailed explanation of how Randomized Search CV operates: (a) Establish a search area: Indicate the range of values or distributions from which to sample the hyperparameters; (b) Randomly sample hyperparameter combinations: Pick a selection of hyperparameter combinations at random from the search space; (c) Evaluate each combination: Utilizing each combination of hyperparameters, develop and test the model. Typically, k-fold cross-validation is used for this, where the data is divided into k subsets (folds), the model is trained and assessed k times, and each time, a different fold is used as the validation set; (d) Choose the optimal combination: The performance metric (such as accuracy, precision, or recall) achieved during the cross-validation procedure should be used to determine the optimum hyperparameter combination; and (e) Train the model again: On the complete training dataset, train the model using the optimal combination of hyperparameters.

Randomized Search CV rapidly explores a wide variety of hyperparameter combinations without analyzing all potential possibilities by using random sampling as opposed to an exhaustive grid search. This makes it appropriate in situations where the hyperparameter search space is huge or when there are not enough processing resources. In general, Randomized Search CV aids in automating the hyperparameter tuning process, enabling the choice of ideal hyperparameters for a machine learning model.

3. PREPARATIONS OF DATA AND INTERPRETATION

In order to evaluate SPT-N value of soil through a comparison analysis, five ML procedures are utilized to the dataset of 1113 instances of Cone Penetration Testing (CPT) data. These are collected from 54 collocated CPT and SPT within the DMDP area of Bangladesh as shown in Fig. 1. It should be noted that the dataset is fairly thorough and contains a wide range of metrics that are important for figuring out the SPT-N values. For determining SPT-N values, it is critical to choose the factors that will have the biggest impact. As input characteristics, five variables are chosen, including depth (D), moisture content (MC), fine content (FC), CPT tip resistance (q_c) and CPT local friction (f_s) are nominated as an input parameters (see Table 2).

Figure 2 shows the correlation matrix for the affecting variables as a heat map and Fig. 3 shows the affecting variables as a pair panel. The correlation matrix displays the correlation coefficient between the variables, while the pair panel displays the histogram of individual variable and scatter plots between two variables.

3.1 Data Splitting and Cross-Validation

To preserve the model's capacity to simplify while addressing the overfitting issue, in this work, around 80% of the instances are considered in the training set and 20% of specimens are allotted to the testing set utilizing arbitrary selection. It should be mentioned that before performing any modeling, we have normalized the dataset. The objective of normalization is to convert the dataset's values to a mutual scale without affecting variations in the value ranges.

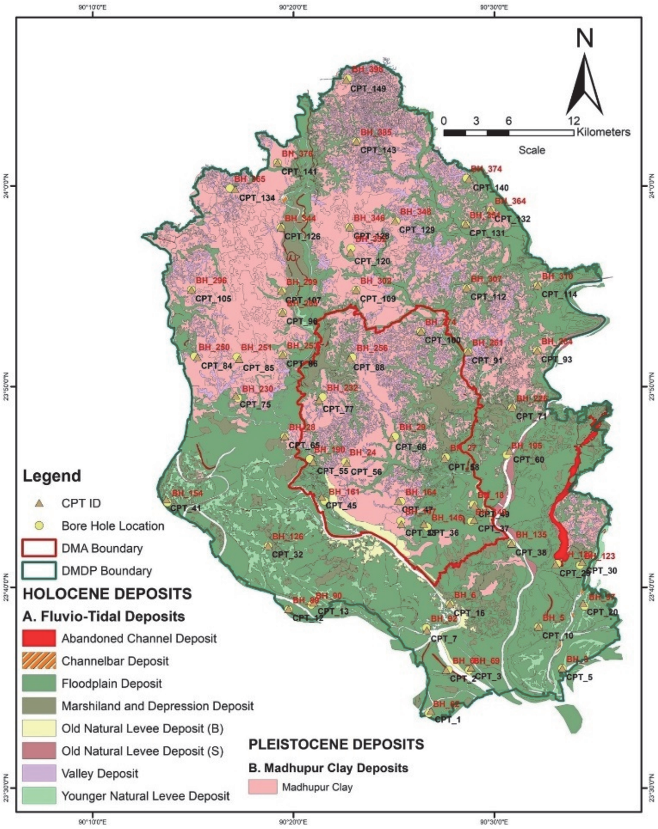


Fig. 1 **Locations of CPT/SPT tests performed in the DMDP area along with the geology**

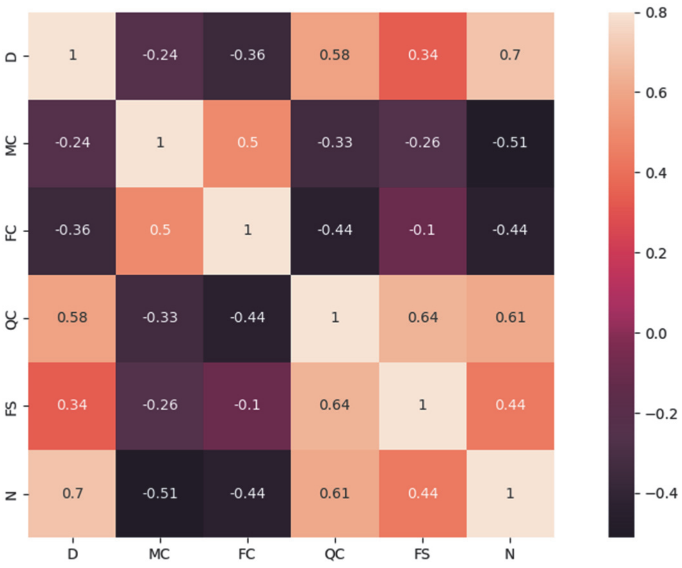


Fig. 2 **Correlation matrix of one output variable and five input variables**

Table 2 Summary of the dataset

	D (m)	MC (%)	FC (%)	q_c (MPa)	f_s (MPa)	SPT-N
Mean	17.0	24.9	58.9	7.12	0.1285	22.6
STD	10.1	8.9	31.8	5.70	0.1068	14.8
Min	1.5	1.4	5.0	0.01	0.0010	1
Max	49.5	89.1	100.0	34.49	0.7140	50

Dataset: 1113

The predictive power of five OML algorithms is assessed in this study using K-fold cross-validation on the same data. The data can be subjected to cross-validation techniques to reduce the likelihood of overfitting and bias during selection in the ML approaches. The data is split into K equal-sized subsets for the K-fold CV. The single surviving subset of the K subsets is employed as the testing data, while the K-1 subsets are utilized as training data. Then, this procedure is carried out K times using various subsets as the testing subset. In order to evaluate OML algorithms on a small sample of data, CV is a resampling approach. The 5-fold CV is the most popular CV, which has been utilized in this study.

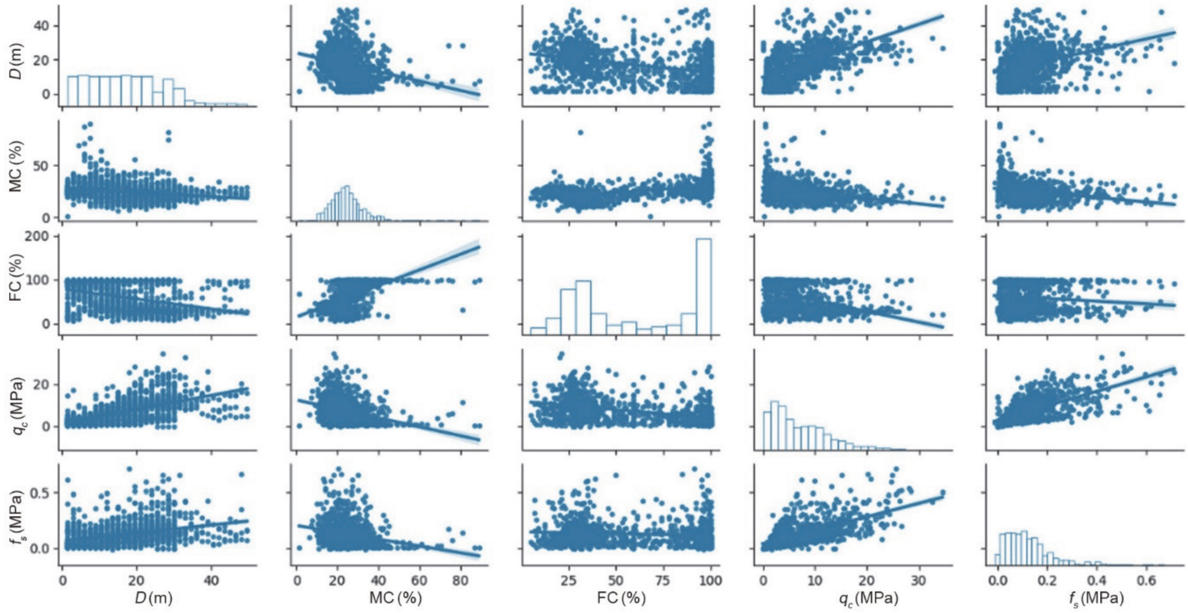


Fig. 3 **Pair panel of input variables**

3.2 Measures of Performance

The SPT-N value (N) is investigated between actual and estimated values using mean absolute error (MAE), root mean square error (RMSE), variance accounted for (VAF) and R-squared value (R^2) to show the accuracy of five OML algorithms' estimation. The mean absolute error is the mean absolute error between actual and predicted values. The most often used metric for assessing models is the mean squared error. Here, the difference between actual values and anticipated values is squared, and the average of those values is computed for each data point. The MSE can be a helpful statistic to employ when the dataset contains unforeseen values, either very high or low values. However, the MSE can either overstate or underestimate how awful the prediction is when dealing with noisy data, *i.e.,* when the data are not completely dependable. The RMSE is described as a root of MSE. Another statistic for regression issues that calculates the variance between the actual values and anticipated values is variance accounted for. A statistical parameter R-squared may be used to assess the accuracy of the fit which represents how narrowly an algorithm resembles the real data points.

4. ANALYSIS FINDINGS

This section discusses hyper-parameter tuning, a comparison of five ML methods for estimating the SPT-N value of soil, and the significance of influencing variables. Figure 4 depicts the process for SPT-N value evaluation of soil using OML approaches. Training and testing datasets are created from the initial dataset. Five cutting-edge ML algorithms are optimized after being trained on the training dataset. Then, the OML models are utilized to the test dataset in order to compare their results.

4.1 Hyper-Parameter Tuning Results

The hyper-parameters of every ML method which have been obtained through the randomized search cross-validation (RSCV) algorithm are shown in Table 3, along with their tuned values. Figure 5 displays the evolution of the root mean squared error value over the training dataset's iterations. Figure 5 shows that hyper-parameter adjustment, especially for MLP, SVR, and XGB, has a significant impact on how well ML algorithms perform.

Table 3 Hyper-parameter tuning

ML algorithms	Optimum value
XGB	'subsample': 1.0, 'reg_lambda': 0.5, 'reg_alpha': 0, 'n_estimators': 300, 'max_depth': 3, 'learning_rate': 0.01, 'colsample_bytree': 0.8
MLP	'activation': 'tanh', 'alpha': 0.0024062425041415756, 'hidden_layer_sizes': 91, 'learning_rate': 'adaptive', 'max_iter': 610, 'solver': 'adam'
RF	max_depth = 9, max_features = 'sqrt', max_leaf_nodes = 9, n_estimators = 150
RIDGE	'solver': 'saga', 'alpha': 0.24770763559917114
SVR	'kernel': 'rbf', 'gamma': 0.0001, 'C': 1000

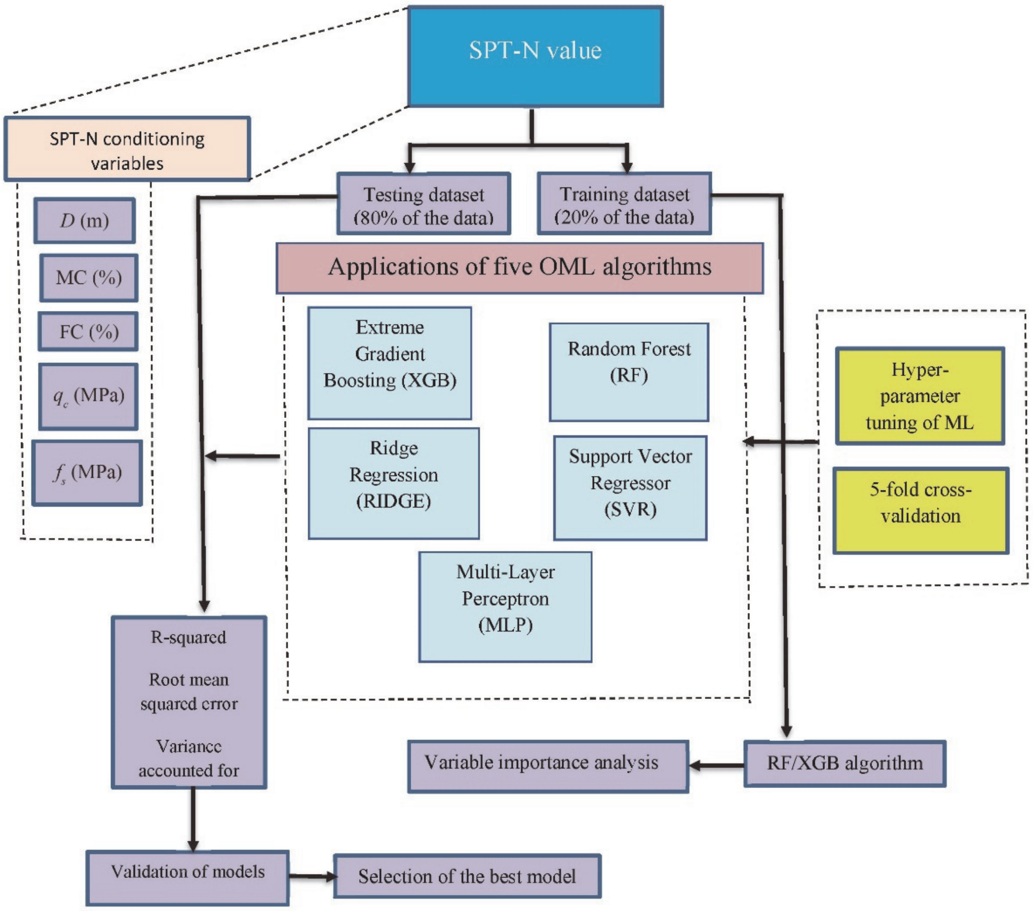

SPT-N value
SPT-N conditioning variables
D (m)
MC (%)
FC (%)
q_c (MPa)
f_s (MPa)
Testing dataset (80% of the data)
Training dataset (20% of the data)
Applications of five OML algorithms
Extreme Gradient Boosting (XGB)
Random Forest (RF)
Ridge Regression (RIDGE)
Support Vector Regressor (SVR)
Multi-Layer Perceptron (MLP)
Hyper-parameter tuning of ML
5-fold cross-validation
R-squared
Root mean squared error
Variance accounted for
Variable importance analysis
RF/XGB algorithm
Validation of models
Selection of the best model

Fig. 4 Methodological flowchart of this study

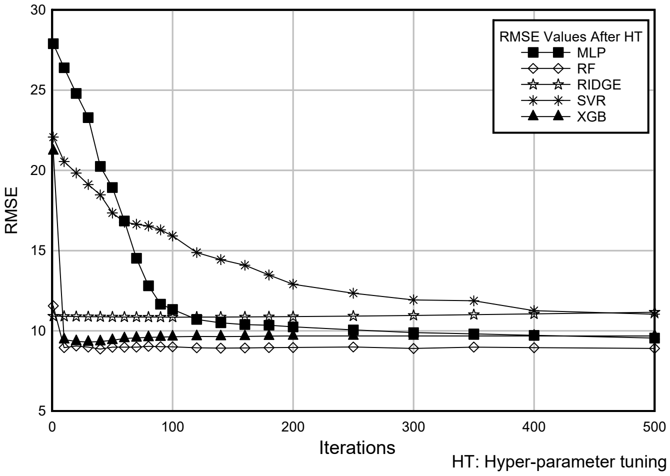


Fig. 5 RMSE values along with iterations on the training dataset

4.2 Review and Comparison of Five Machine Learning Models

4.2.1 Training Dataset Results

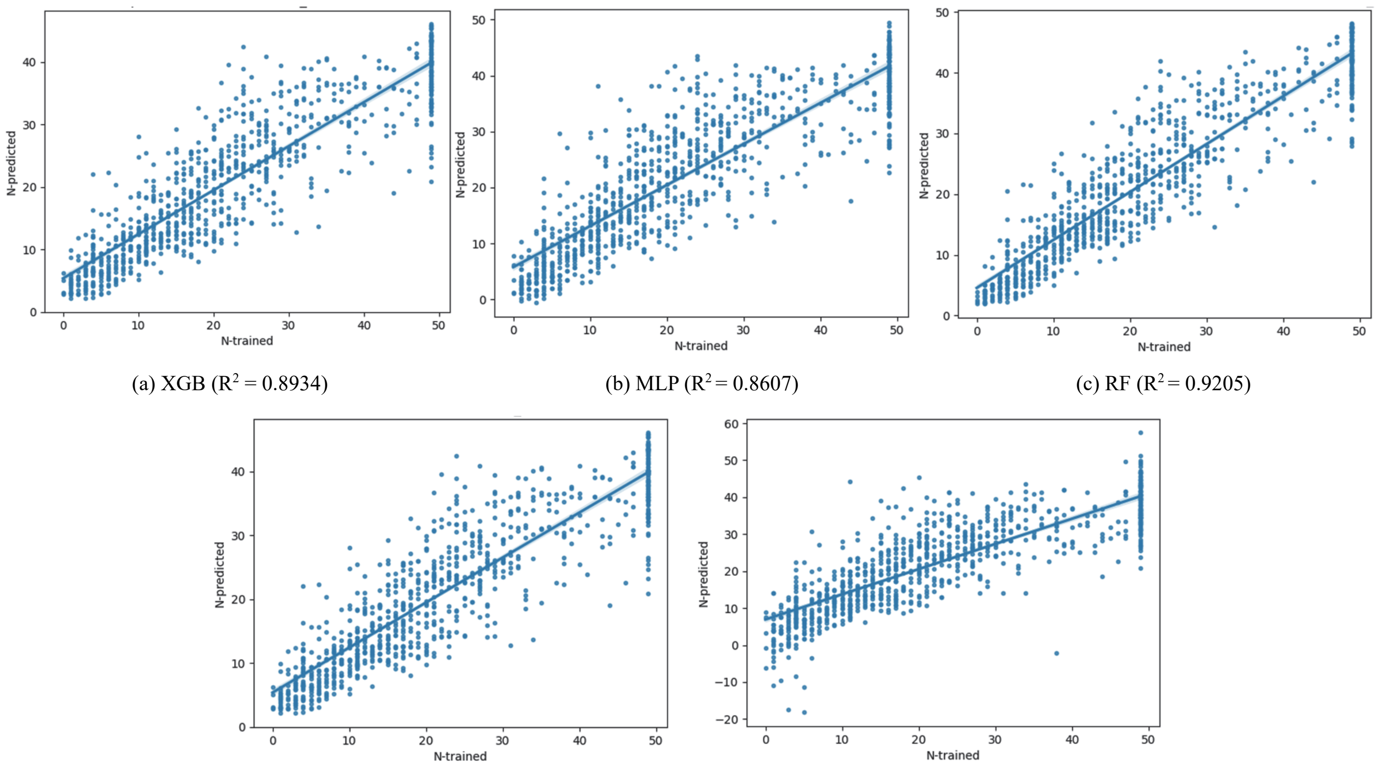
On the training data of 890 SPT-N value and corresponding CPT and other soil parameters, five OML algorithms are used. The regression graphs for each of these techniques are displayed in Fig. 6. Among the five OMLs, RF and SVR exhibit the best performances (R^2 = 0.9205 and 0.8956, respectively), which is exceptional performance. XGB, which has an R-squared value of 0.8934, comes next. The performances of MLP (R^2 = 0.8607) and RIDGE (R^2 = 0.8252) are adequate.

4.2.2 Testing Dataset Results

The SPT-N values of the soils in the testing dataset are now estimated using five OML models that were trained in the former section. Five OML approaches' performance on 223 samples from the testing data—where no training procedure was applied—is assessed. Figure 7 shows the regression graphs for the test set of five OML models. The ranking of OML technique performance for the testing dataset is different from that for the training dataset, as shown by a comparison of Figs. 6 and 7, and R-squared values also vary between the training and testing datasets. On the testing dataset, SVR and MLP display excellent performances (R^2 = 0.87677 and 0.8681, respectively). Both XGB (R^2 = 0.8647) and RF (R^2 = 0.8579) exhibit good performances. This is followed by RIDGE (R^2 = 0.8110), which shows acceptable performance.

4.2.3 Results Comparison

Each OML algorithm's mean absolute error, root mean squared error, variance accounted for, and R-squared values for training and testing datasets are presented in Table 4. Each technique's performance ranking is also displayed. According to Table 4, RF and RIDGE, which are ranked first and last among OML models for training data, respectively obtain the uppermost and lowermost R-squared values. Similarly for the testing datasets, SVR and RIDGE are categorized first and last among OML algorithms for testing datasets, respectively. For testing datasets, MLP is ranked 2nd and XGB is ranked 3rd. The efficiency of OML procedures on the testing data is more significant to be taken into account as a utilization of each OML algorithm meanwhile the testing data may be seen as an illustration of an actual situation.



(a) XGB (R^2 = 0.8934)

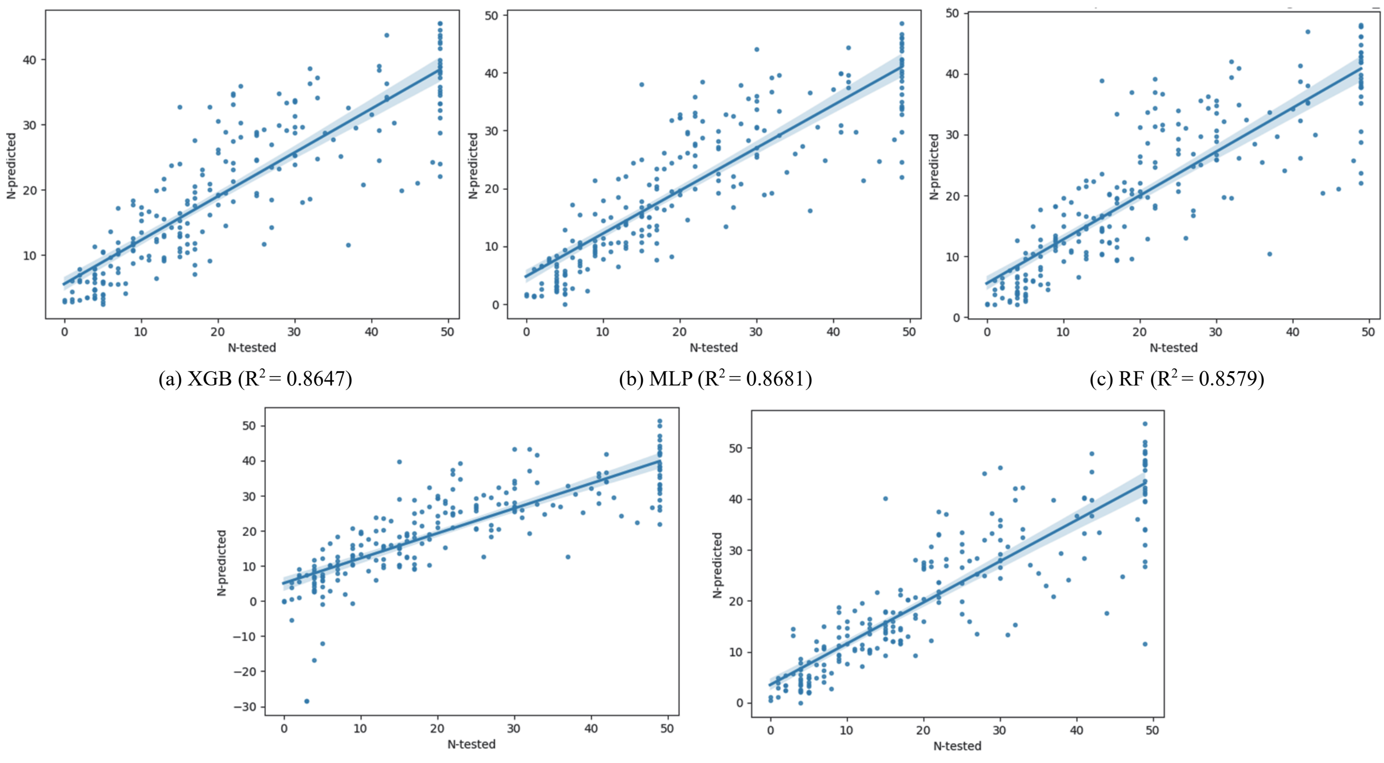
(b) MLP (R^2 = 0.8607)

(c) RF (R^2 = 0.9205)

(d) RIDGE (R^2 = 0.8252)

(e) SVR (R^2 = 0.8956)

Fig. 6 Regression plots for the training set of five OML algorithms



(a) XGB ($R^2 = 0.8647$)

(b) MLP ($R^2 = 0.8681$)

(c) RF ($R^2 = 0.8579$)

(d) RIDGE ($R^2 = 0.8110$)

(e) SVR ($R^2 = 0.8767$)

Fig. 7 Regression plots for the testing set of five OML algorithms

Table 4 Statistical analyses of five ML algorithms

ML algorithms		MAE	RMSE	VAF	R^2	Ranking
XGB	Training	5.1915	6.8599	0.7825	0.8934	3
	Testing	5.6720	7.8080	0.7920	0.8647	3
MLP	Training	5.5984	7.4895	0.7407	0.8607	4
	Testing	5.4257	7.5100	0.7521	0.8681	2
RF	Training	4.3984	5.8213	0.8433	0.9205	1
	Testing	5.5750	7.7442	0.8502	0.8579	4
RIDGE	Training	6.2990	8.3070	0.6810	0.8252	5
	Testing	6.6240	8.9220	0.6950	0.8110	5
SVR	Training	4.2836	6.5594	0.8011	0.8956	2
	Testing	4.9374	7.2897	0.8098	0.8767	1

Considering depth, cone tip resistance and side friction as independent variables and SPT-N value as the dependent variable, a multiple linear regression equation has been developed by Hasan (2023), where obtained R^2 is only 0.6399. In this study, using SVR and MLP ML algorithms for testing dataset, this R^2 value has been increased to 0.8767 and 0.8681, respectively.

4.3 Results of Variable Importance

The RF and SVR exhibit the top relative efficiency in the training and the testing data respectively, according to the comparison. In order to evaluate the significance of influencing parameters for the SPT-N value estimation of soils, RF is used. The normalized values for variable importance are displayed in Fig. 8. In this analysis, it has been assumed that there exists a trained model (in this case RF). For RF model, feature importance represent a score that measures how useful each feature is in the construction of the RF within the ensemble. These importance are typically computed based on how much each feature decreases impurity across all the RF in the ensemble. These importance are computed during the training of the model and reflect how much each feature contributes to the model's predictive performance. Overall, feature importance provide valuable insights into which features are most relevant to the model's predictions, helping users understand the model's behavior and potentially identify important features in the dataset.

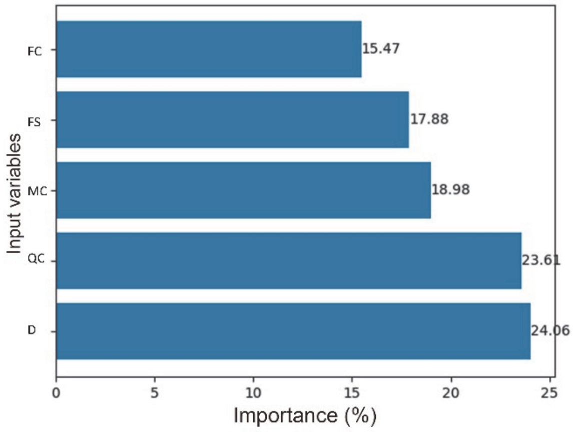


Fig. 8 Relative variable importance for SPT-N value (for the case RF)

According to Fig. 8, depth (D) is the factor that has the biggest impact on estimating the SPT-N value of soils (score = 0.2406). The remaining factors' importance values for estimating the SPT-N value of soils fall in the following order: cone tip resistance (score = 0.2361) > moisture content (score = 0.1898) > cone shaft friction (score = 0.1788) > fine content (score = 0.1547). It is crucial to keep in mind that when alternative datasets and models are used, scores may vary. Additionally, it is essential to mark that these five input features are of non-ignorable relevance because they are used as the fundamental input parameters in the majority of engineering projects.

4.4 Comparison with Existing Models, RF and XGB

In this section, previously established SPT-N value versus CPT correlations developed by Arifuzzaman and Anisuzzaman (2022) and Hasan (2023) for all soil types have been used to estimate SPT-N value of soils and the results are then compared to the RF and SVR algorithms results. Figure 9(a) exhibits that the RF and SVR approach outperformed Arifuzzaman and Anisuzzaman (2022) model in terms of performance. On the other hand, Hasan (2023) model performs relatively better in estimating SPT-N value of soils. The positive and negative residuals (difference between predicted and measured SPT-N values) of Fig. 9(b) show that Arifuzzaman and Anisuzzaman (2022) model under-predicts and Hasan (2023) model slightly over-predicts the majority of the dataset.

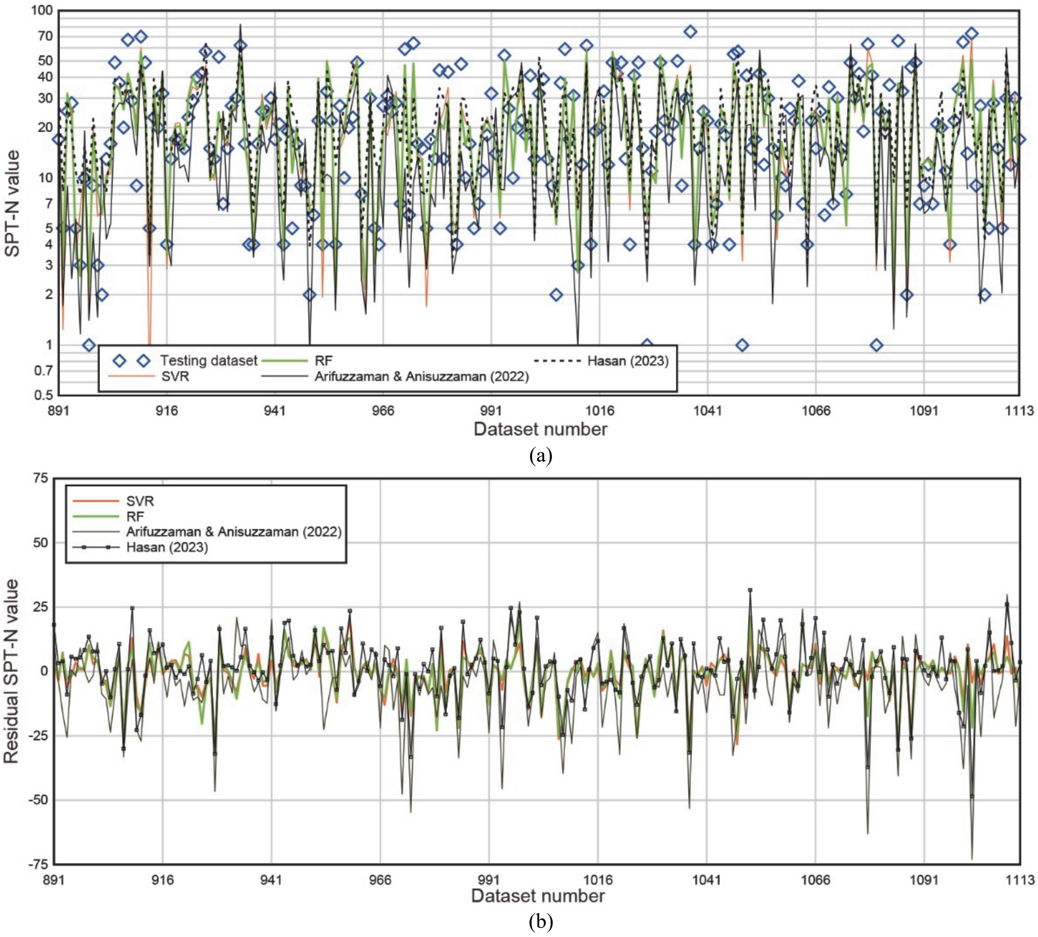
4.5 Discussion

This study's main advantage is its comparison and proposal of five improved machine learning (ML) techniques for estimating the SPT-N value of soils. The following components of this study add to our understanding of the estimation of SPT-N value of soils and other geotechnical engineering fields: For regression problems in geotechnical engineering: (a) the optimized ML approaches are extremely promising; (b) the stability and resilience of regression procedures may be effectively explored using MAE, RMSE, VAF, and R-squared values; (c) the approach presented in this study has significant promise for a wider use in other geotechnical engineering areas where regression complications are frequently faced; and (d) a few guidelines have been given for forecasting the SPT-N value of soils by applying ML algorithms. If more data can be collected, the efficiency of the suggested optimized ML models can be enhanced.

5. CONCLUSIONS AND RECOMMENDATIONS

When laboratory testing cannot be done, empirical equations are employed to determine the engineering properties of soil. In-situ test findings and soil index features are frequently combined to create empirical relationships that offer cost-effective and non-destructive alternatives. The training dataset that is utilized to construct the algorithm affects how effective it is. In this study, the optimum model for estimating the SPT-N value of soils has been identified by a thorough evaluation of five OML models, comprising MLP, RF, RIDGE, SVR, and XGB. 1113 pieces of data make up the dataset used by the OML algorithms. As performance measure, fivefold cross-validation is employed, along with mean average error, root mean square error, R-squared, and variance accounted for values. Following are some significant deductions:

Randomized search cross-validation (RSCV) algorithm is a useful procedure for tweaking the hyper-parameters of ML models, in line with the optimal scores attained by ML algorithms across iterations.



(a)

(b)

Fig. 9 SPT-N values at testing stages: (a) predicted; (b) residual

On the training dataset of 890 SPT-N value and corresponding CPT and other soil parameters, five OML algorithms are used. Among the five OMLs, RF and SVR exhibit the best performances (R^2 = 0.9205 and 0.8956, respectively), which is exceptional performance. XGB, which has an R-squared value of 0.8934, comes next. The performances of MLP (R^2 = 0.8607) and RIDGE (R^2 = 0.8252) are adequate. The SPT-N values of the soils in the testing dataset are also estimated using five OML algorithms that were trained. Five OML approaches' performance on 223 samples from the testing data—where no training procedure was applied—is assessed. The ranking of OML technique performance for the testing dataset is different from that for the training dataset, and that R-squared values also differ between the training and testing data. On the testing dataset, SVR and MLP display excellent performances (R^2 = 0.8767 and 0.8681, respectively). Both XGB (R^2 = 0.8647) and RF (R^2 = 0.8579) exhibit good performances. This is followed by RIDGE (R^2 = 0.8110), which shows acceptable performance.

The CPT cone resistance is the variable that has the second greatest influence on determining the SPT-N value of soil, according to the results of a variable importance study (score = 0.2361).

The performance of RF and SVR is compared with the models of Arifuzzaman and Anisuzzaman (2022) and Hasan (2023) on the testing dataset of the present study. The OML models such as RF and SVR presented in this paper outperform Arifuzzaman and Anisuzzaman (2022) model in terms of fit to the testing dataset. On the other hand, Hasan (2023) model performs relatively better in estimating SPT-N value of soils.

Future research may incorporate additional soil factors that may be important in determining in addition to CPT cone resistance and side friction, depth, and grain size analysis parameters *etc.* For more accuracy, the authors advise using hybrid machine learning models.

FUNDING

The authors received no funding for this work.

DATA AVAILABILITY

The data and/or computer codes used/generated in this study are available from the corresponding author on reasonable request.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

REFERENCES

Akca, N. (2003). "Correlation of SPT–CPT data from the United Arab Emirates." *Engineering Geology*, **67**(3-4), 219-231. https://doi.org/10.1016/S0013-7952(02)00181-3

Arifuzzaman, Anisuzzaman, M. (2022). "An initiative to correlate the SPT and CPT data for an alluvial deposit of Dhaka city." *Geo-Engineering*, **13**(5) 1-13. https://doi.org/10.1186/s40703-021-00170-3

Asci, M., Kurtulus, C., Kaplanvural, I., and Mataracioglu, M.O. (2014). "Correlation of SPT-CPT Data from the Subsidence Area in Golcuk, Turkey." *Soil Mechanics and Foundation Engineering*, **51**(6), 268-272.

Demir, S. and Sahin, E.K. (2022). "Comparison of tree-based machine learning algorithms for predicting liquefaction potential using canonical correlation forest, rotation forest, and random forest based on CPT data." *Soil Dynamics and Earthquake Engineering*, **154**, 107130. https://doi.org/10.1016/j.soildyn.2021.107130

Demir, S. and Sahin, E.K. (2023). "An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost." *Neural Computing and Applications*, **35**, 3173-3190. https://doi.org/10.1007/s00521-022-07856-4

Dos Santos, M.D. and Bicalho, K.V. (2017). "Proposals of SPT-CPT and DPL-CPT correlations for sandy soils in Brazil." *Journal of Rock Mechanics and Geotechnical Engineering*, **9**(6), 1152-1158. https://doi.org/10.1016/j.jrmge.2017.08.001

Elbanna, M., Quinn, J., and Martens., S. (2011). "SPT – CPT correlations for oil sands tailings sand." *Proceedings Tailings and Mine Waste 2011*, Vancouver, BC.

Fernando, H., Nugroho, SA, Suryanita, R., and Kikumotoc, M. (2021). "Prediction of SPT value based on CPT data and soil physical properties using ANN with and without data normalization." *International Journal of Artificial Intelegence Research,* **5**(2), 123-131.

Hasan, S. (2023). *Development of Empirical Correlations between CPT and Other Soil Parameters within DMDP Area, Bangladesh*. BSc Engg Thesis, BUET, Dhaka, Bangladesh.

Jarushi, F., AlKaabim, S., and Cosentino, P. (2015). "A New Correlation between SPT and CPT for Various Soils," *World Academy of Science, Engineering and Technology, International Journal of Geological and Environmental Engineering*, **9**(2), 101-107.

Kara, O. and Gunduz, Z. (2010). "Correlation between CPT and SPT in Adapazari, Turkey." *2nd International Symposium on Cone Penetration Testing*, California.

Khan, Z., Yamin, M., Attom, M., Al Hai, N. (2022). "Correlations between SPT, CPT, and Vs for Reclaimed Lands near Dubai." *Geotechnical and Geological Engineering*, **40**, 4109-4120. https://doi.org/10.1007/s10706-022-02143-4

Khodaparast, M., Rajabi, A.M., and Derakhshan, M. (2020). "Development of practical correlations between cone penetration resistance and SPT values for various types of soils." *Iranian Journal of Science and Technology Transactions of Civil Engineering,* **44**(Suppl 1), 471-481. https://doi.org/10.1007/s40996-019-00319-2

Lingwanda, M.I., Larsson, S., and Nyaoro, D.L. (2015). "Correlations of SPT, CPT and DPL data for sandy soil in Tanzania." *Geotechnical and Geological Engineering*, **33**, 1221-1233. https://doi.org/10.1007/s10706-015-9897-1

Mayne, P.W. and Kulhawy, F.H. (1982). "K-OCR relationships in soil," *Journal of the Geotechnical Engineering Division*, ASCE, **108**(6), 851-872. https://doi.org/10.1016/0148-9062(83)91623-6

Robertson, P.K. and Campanella, R.G. (1983). "Interpretation of cone penetration tests: Sands and clays." *Canadian Geotechnical Journal*, **20**, 719-745. https://doi.org/10.1139/t83-079

Robertson, P.K., Campanella, R.G., Gillespie, D., and Greig, J. (1986). "Use of piezometer cone data." in *Use of in Situ Tests in Geotechnical Engineering*, 1263-1280. Reston, VA: ASCE.

Schmertmann, J.H. (1970). "Static cone to compute static settlement over sand." *Journal of the Soil Mechanics and Foundations Division,* ASCE, **96**(3), 1011-1043. https://doi.org/10.1061/JSFEAQ.0001418

Shahien, M.M. and Albatal, A.H. (2014). “SPT-CPT Correlations for Nile delta silty sand deposits in Egypt.” *3rd International Symposium on Cone Penetration Testing*, Las Vegas, Nevada, USA.

Suzuki, Y., Sanematsu, T., and Tokimatsu, K. (1998). “Correlation between SPT and seismic CPT.” *Proceedings of Conference on Geotechnical Site Characterization*, Balkema, Rotterdam, 1375-380.

Tarawneh, B. (2017). “Predicting standard penetration test N-value from cone penetration test data using artificial neural networks.” *Geoscience Frontiers*, **8**(1), 199-204, ISSN 1674-9871. https://doi.org/10.1016/j.gsf.2016.02.003

Zhao, X.L. and Cai, G.J. (2015). “SPT-CPT correlation and its application for liquefaction evaluation in China.” *Marine Georesources & Geotechnology*, **33**(3), 272-281. https://doi.org/10.1080/1064119X.2013.872740

Zhou, H., Wotherspoon, L.M., Hayden, C.P., McGann, C.R., Stolte, A., and Haycock, I. (2021). “Assessment of existing SPT-CPT correlations using a New Zealand database.” *Journal of Geotechnical and Geoenvironmental Engineering*, ASCE, **147**(11). https://doi.org/10.1061/(ASCE)GT.1943-5606.0002650