Utilizing Different Machine Learning Models for Predicting Soil Shear Strength

Priyanshu Jain[1][0009-0009-2915-6969] and Aali Pant[2][0000-0001-6949-9020] and Gaurav Bhatnagar[3][0000-0002-0282-3372]

[1]BTech Student Department of Civil and Infrastructure Engineering, Indian Institute of Technology Jodhpur, Rajasthan
[2]Assistant Professor, Department of Civil and Infrastructure Engineering, Indian Institute of Technology Jodhpur, Rajasthan
[3] Professor, Department of mathematics, Indian Institute of Technology Jodhpur, Rajasthan

Abstract
This paper presents a case study of using different machine learning algorithms for predicting the peak shear strength of soil, which is a critical soil parameter. The study was conducted using 309 results of direct shear test on granular soils, the data of which has been collected from published literature. The machine learning models used in the study included extreme gradient boosting (XGBoost), random forests, support vector machines (SVM), artificial neural networks (ANN), and their ensemble models. The performance of each model was evaluated based on several statistical metrics, including root mean squared error, and coefficient of determination (R^2). Influence of linear correlation between different factors such as normal stress, relative density of sample, grain size distribution characteristics etc. has been assessed on soil shear strength. Normal stress is observed to be the most influencing factor. It is observed that XGBoost gives the most accurate results with R^2 value of 0.99 in training dataset and 0.97 in testing dataset. The study demonstrates the potential of machine learning algorithms in predicting important soil parameters, which can have significant implications for engineering design and construction practices.

Keywords: Shear Strength, Machine Learning, XGBoost

1 Introduction

Soil is a complex and heterogeneous material that plays a vital role in many engineering applications. One of the key soil properties that affects its behavior and performance is the peak soil strength, which represents the maximum resistance of soil to shear deformation. Accurate prediction of peak soil strength is essential for geotechnical design and analysis, such as slope stability, foundation bearing capacity, and soil reinforcement.

However, measuring peak soil strength in the laboratory or in situ is often time-consuming, costly, and limited by the availability of soil samples[1], [2]. With the recent advances in machine learning (ML) algorithms and computing technology, there has been a growing interest in using these techniques to predict soil properties [3], [4]. Alternative methods that can estimate peak soil strength from easily obtainable soil parameters are desirable. ML is a branch of artificial intelligence that can learn from data and discover complex patterns and relationships[5]–[8]. ML has been widely applied to model various soil properties for geotechnical design, such as soil maximum dry density soil organic carbon², and soil water content[9]. Machine learning algorithms have shown promising results in predicting various soil properties, including shear strength, compressibility, and permeability.[10]–[13]

2 Literature Review

Several researchers have explored the use of machine learning algorithms for predicting soil properties, including the peak shear strength of soil. For instance, Khanlari (2012) used artificial neural networks (ANNs) to predict the undrained shear strength of clayey soils[10]. They trained the ANN using a dataset of 292 soil samples and found that the model achieved high accuracy in predicting the undrained shear strength of the soil.

Decision trees and random forests are other machine learning algorithms that have been used for predicting soil properties. For instance, Pacheco (2023) used decision trees to predict the residual shear strength of red clay soils [14]. They collected a dataset of 60 soil samples and found that the decision tree model achieved a high accuracy in predicting the residual shear strength of the soil. Meanwhile Pacheco (2023) used random forests to predict the shear strength of sandy soils[14]. They collected a dataset of 140 soil samples and found that the random forest model achieved high accuracy in predicting the shear strength of the soil.

Other machine learning algorithms, such as support vector machines (SVMs), have also been explored for predicting soil properties. For instance, Drucker, (1994) used SVMs to predict the shear strength of non-slum concrete[15] . They collected a dataset of 129 rockfill samples and found that the SVM model achieved high accuracy in predicting the shear strength of the materials.

2.1 XGBOOST

XGBoost stands for Extreme Gradient Boosting and is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. Gradient boosting is an ensemble method that builds a series of decision trees and optimizes a differentiable loss function using gradient descent [15]. XGBoost has several advantages over other gradient boosting algorithms, such as parallelization,

regularization, scalability, and handling of missing values . XGBoost has been widely used and recognized for its performance and speed in various regression tasks[16].

2.2 Support Vector Machine Regressor

Support vector machine regressor is a machine learning algorithm that follows the principle of support vector machine to regression problems[17]. SVR is used to find a function that finds the relationship between the input variables and a continuous target variable at the same time minimizing the prediction error and the model complexity[18]. It uses kernels to map the input data into a high-dimensional feature space, where a linear function can be fitted to capture the nonlinear patterns in the data. It also uses an epsilon-insensitive loss function, which ignores the errors within a certain margin, and a regularization term, which penalizes the model for having large coefficients[19].

2.3 Random Forest Regressor

Random forest regressor (RFR) is a machine learning algorithm that applies the principle of random forest (RF) to regression problems[20]. RF is an ensemble method that builds a series of decision trees and optimizes a differentiable loss function using bootstrap aggregation, commonly known as bagging[20], [21]. RFR uses random sampling of features and data points to create diverse and uncorrelated decision trees, which reduces the variance and improves the generalization of the model. RFR also uses an out-of-bag error estimate, which measures the prediction error on the data points that are not used for training each tree. It has been applied to various geotechnical problems[1], [14], [22][23].

2.4 Artificial Neural Network Regressor

Artificial neural network regressor (ANNR) is a machine learning algorithm that applies the principle of artificial neural network (ANN) to regression problems. ANN is a computational model that simulates the structure and function of biological neural networks, such as the human brain. ANNR consists of interconnected processing units, called neurons, that are organized into layers. ANNR learns the relationship between the input variables and a continuous target variable by adjusting the weights and biases of the neurons based on a learning algorithm and a loss function. It has also been applied to various geotechnical problems[6], [24]–[27].

2.5 Ensemble Models

Ensemble models are machine learning algorithms that combine multiple basic models to produce a final output[13], [16], [28]. Ensemble models can be classified into two types: sequential and parallel. Sequential ensemble models, such as boosting, train the base models in a sequential manner, where each model tries to correct the errors of the previous ones. Parallel ensemble models, such as bagging, train the base models in a

parallel manner, where each model is trained on a subset of the data or features and then aggregated by voting or weighted averaging[29], [30].

3 Methodology

The shear strength of soil is affected by various factors, including mean diameter (D50), uniformity coefficient (Cu), coefficient of gradation (Cc), normal stress(N_S), and relative density. In this study, we analyzed a dataset of 309 values which contains 6 features, including D_{50}, C_u, C_c, normal shear stress(N_S), and peak shear strength(P_S_S). The data was collected from literature and published work, and no outliers were detected as all data points were deemed valid.

The dataset was pre-cleaned, and no missing values were present. The relative density feature was initially given in terms of dense, medium, and loose, which is a categorical feature. In order to incorporate this feature into the analysis, the density states were converted in to a numerical value using a coding system (dense=1, medium=2, loose=3).

Exploratory data analysis was conducted to gain insights into the relationships between different features. The mean, median, mode, standard deviation, and correlation coefficient between different features were then estimated and have been tabulated in Table 1. The correlation coefficients were calculated using the Pearson correlation coefficient, which measures the strength and direction of the linear relationship between two variables. The values of the correlation coefficient range from -1 to 1, where a value of -1 indicates a perfect negative correlation, a value of 0 indicates no correlation, and a value of 1 indicates a perfect positive correlation, as shown in Figure 1.

Table-1: Description of Data

	D_{50}, mm	**C_u**	**C_c**	**N_S (kPa)**	**P_S_S (kPa)**
Count	309	309	309	309	309
Mean	0.560	4.008	1.035	136.9	121.4
Std	0.406	2.712	0.289	157.0261	13.627
Min	0.086	1.06	0.34	4	2.2
25%	0.25	2.20	0.91	40	33.9
50%	0.44	3.12	1	100	78.5
75%	0.81	5.50	1.125	200	152.8
Max	2.18	15.20	2.03	910	800.4

*Std refers to Standard Deviation

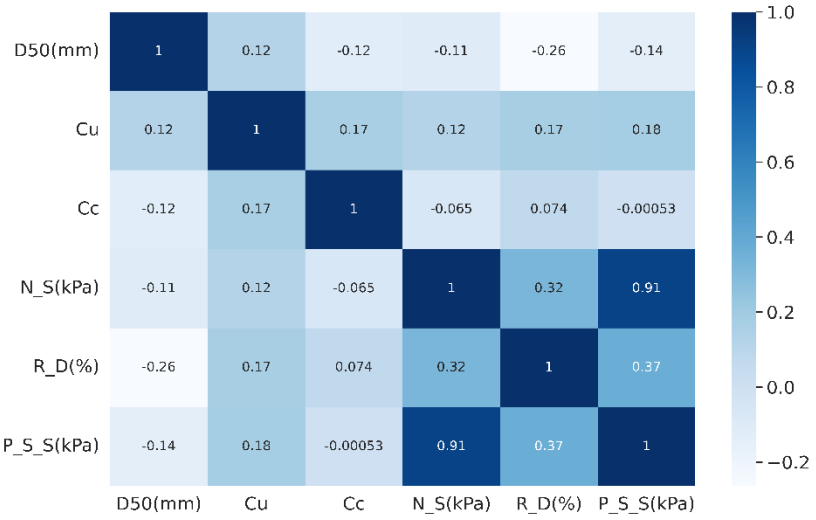

Fig-1: Heatmap of Linear Correlation

To prepare the data for modeling, two different scaling methods were used. The first method was min-max scaling, which scales the data to a range of 0 to 1. This was achieved by subtracting the minimum value from each feature and then dividing by the range of the feature. The second method was mean-variance scaling, which scales the data to have zero mean and unit variance. This was achieved by subtracting the mean from each feature and then dividing by the standard deviation. The purpose of these scaling methods was to ensure that all features were on the same range and had similar scales, which can improve the performance of some machine learning algorithms.

Next, the dataset was split into training and testing sets in an 80:20 ratio, respectively. The training set, which contains 80% of the original dataset, was used to train the machine learning models. The testing set, which contains 20% of the original dataset, was used to evaluate the performance of the models during training and to select the best performing model.

In this study, various machine learning models as discussed in the previous section were used to predict the peak shear strength of soil based on the given dataset. The models used in this study include support vector regression (SVR), random forest, XGBoost, Random Forest and 3 different ensemble approaches , and their corresponding scaled versions. In addition, an artificial neural network (ANN) model with a Relu activation function has also been used. Architecture of the 3 different ensemble models which are used in study are shown in Figure 3-5.

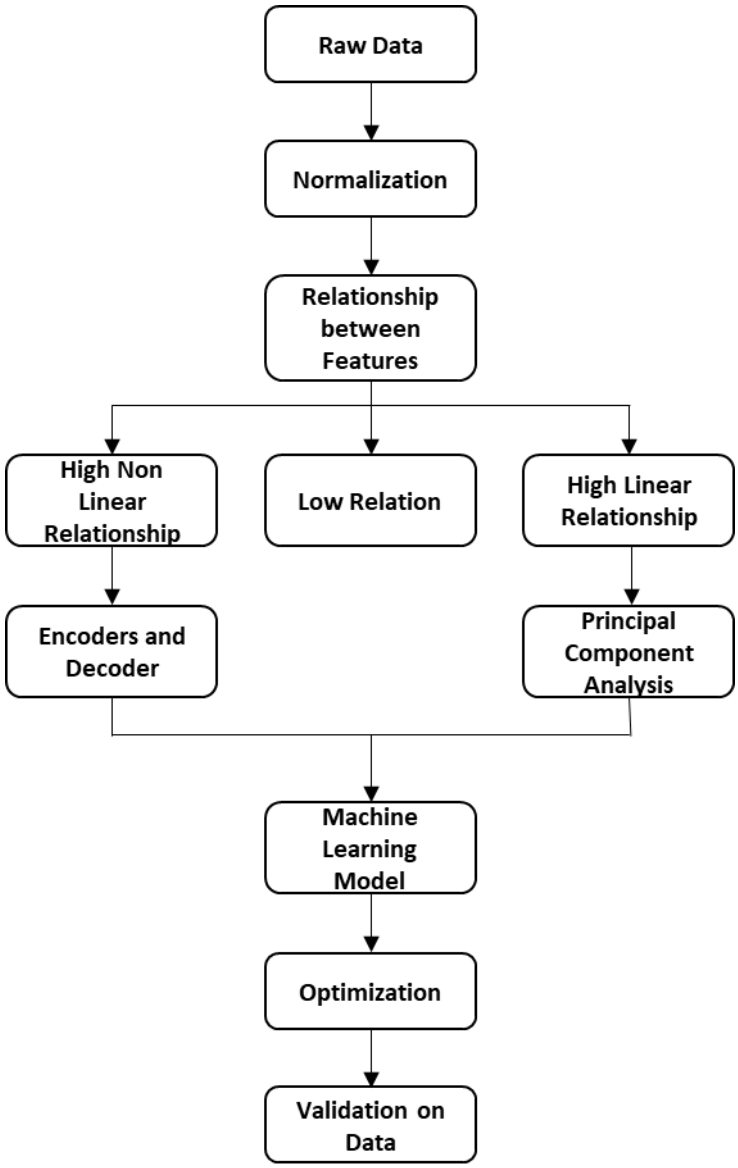

Fig-2: Methodology adopted in the study

Fig-3: Machine learning model of Ensemble-1

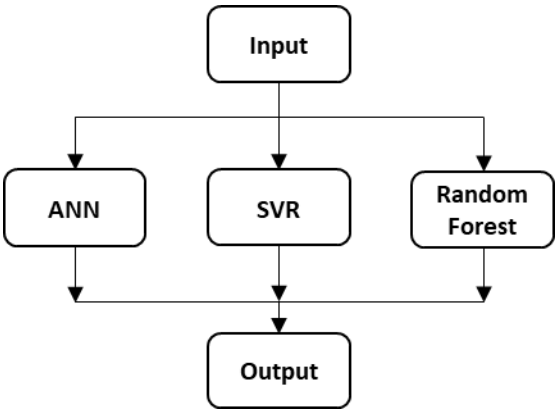

Fig-4: Machine learning model of Ensemble -2

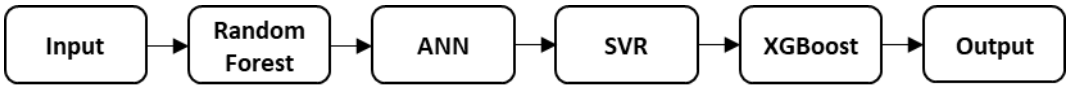

Fig-5: Machine learning model of Ensemble -3

Each model was trained on the training set and its performance was evaluated on the validation set using root mean squared error (RMSE) as the evaluation metric. The performance of each model was also compared with and without scaling the features using the two scaling methods: min-max scaling and mean-variance scaling. After training and evaluation, the best performing model based on its RMSE score on the testing set was selected.

4 Results

The study found that min-max scaling consistently resulted in the least RMSE across all models for the single dataset evaluated. Normalization and standardization scaling had mixed results, with some models showing an improvement in RMSE, while others showed an increase. However, min-max scaling produced the lowest RMSE as shown in Table 2, indicating its superiority over other scaling techniques for this particular dataset.

Table -2 : Different Models RMSE for different scaling on testing and training data

Model	RMSE on Testing			RMSE on Training		
	Without Scaling	Min- Max Scaling	Mean Variance Scaling	Without Scaling	Min- Max Scaling	Mean Variance Scaling
XG Boost	77	0.02	0.16	71	0.003	0.019
Support Vector Regressor	121	0.08	0.39	114	0.06	0.18
Random Forest	23	0.03	0.18	15	0.019	0.11
ANN	30	0.05	0.29	27	0.03	0.2
Ensemble -1	57	0.04	0.2	40	0.03	0.16
Ensemble -2	51	0.05	0.28	35	0.03	0.18
Ensemble -3	62	0.38	0.42	59	0.33	0.36

Furthermore, the study also found that the impact of scaling on RMSE varied depending on the model architecture. In some cases, scaling had a significant impact on RMSE, while in others, the impact was minimal. Nonetheless, min-max scaling consistently produced the best results across all models. The values of R2 for different models in training and testing dataset have been tabulated in Table 3.

Table -3 : Different Models R^2 Square for different scaling on testing and training data

Model	R^2 Square on Testing			R^2 Square on Training		
	Without Scaling	Min- Max Scaling	Mean Variance Scaling	Without Scaling	Min- Max Scaling	Mean Variance Scaling
XG Boost	0.96	0.97	0.97	0.99	0.99	0.99
Support Vector Regressor	0.65	0.74	0.84	0.70	0.76	0.96
Random Forest	0.96	0.96	0.96	0.98	0.98	0.98
ANN	0.80	0.88	0.96	0.94	0.81	0.91
Ensemble -1	0.82	0.96	0.97	0.91	0.90	0.16
Ensemble -2	0.92	0.88	0.18	0.97	0.96	0.28
Ensemble -3	0.96	0.95	0.62	0.97	0.95	0.68

Based on feature importance study, it was observed that normal shear stress, D_{50}, and C_u were consistently the top three features that influence peak shear strength of soil.

Normal shear strength had the highest weight, followed by D_{50} and C_u, indicating their importance in determining the outcome of the models.
Furthermore, the study also found that the impact of other features on model performance varied depending on the dataset.

5 Conclusion

The current study discusses the use of ML models for predicting peak shear strength of granular soils. The study indicates the importance of scaling on the performance of ML models. Normalization scaling and standardization scaling are common techniques that generally result in a decrease in RMSE, but their effectiveness depends on the dataset.

The results indicate the superiority of XGBoost model in predicting most accurate peak shear strength values. Furthermore, three different ensemble models were developed amongst which Ensemble 1 method produced the best performing model. In this method, the output of all three models were averaged to generate the final predicted value. Additionally, for each model, the output from one model was used as the label for the subsequent models during training. By doing so, a highly accurate and reliable model for the given dataset could be achieved. The study indicates the reliability of using hybrid models which include combining different mathematical models to develop a superior model.

References

[1] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geoscience Frontiers*, vol. 12, no. 1, pp. 469 – 477, 2021, doi: 10.1016/j.gsf.2020.03.007.

[2] S. C. Jong, D. E. L. Ong, and E. Oh, "State-of-the-art review of geotechnical-driven artificial intelligence techniques in underground soil-structure interaction," *Tunnelling and Underground Space Technology*, vol. 113, 2021, doi: 10.1016/j.tust.2021.103946.

[3] A. Mahmoodzadeh *et al.*, "Gaussian process regression model to predict factor of safety of slope stability," *Geomechanics and Engineering*, vol. 31, no. 5, pp. 453 – 460, 2022, doi: 10.12989/gae.2022.31.5.453.

[4] K. S. Hudson, K. J. Ulmer, P. Zimmaro, S. L. Kramer, J. P. Stewart, and S. J. Brandenberg, "Unsupervised machine learning for detecting soil layer boundaries from cone penetration test data," *Earthq Eng Struct Dyn*, 2023, doi: 10.1002/eqe.3961.

[5] V.-H. Nhu, B. T. Pham, and D. T. Bui, "A novel swarm intelligence optimized extreme learning machine for predicting soil shear strength: A case study at Hoa Vuong new urban project (Vietnam)," *Vietnam Journal of Earth Sciences*, vol. 45, no. 2, pp. 219 – 237, 2023, doi: 10.15625/2615-9783/18338.

[6] T. F. Kurnaz, C. Erden, A. H. Kökçam, U. Dağdeviren, and A. S. Demir, "A hyper parameterized artificial neural network approach for prediction of the

factor of safety against liquefaction," *Eng Geol*, vol. 319, 2023, doi: 10.1016/j.enggeo.2023.107109.

[7] F. K. Boadu, "A support vector regression approach to predict geotechnical properties of soils from electrical spectra based on Jonscher parameterization," *Geophysics*, vol. 85, no. 3, pp. EN39 – EN48, 2020, doi: 10.1190/geo2019-0256.1.

[8] Y. Li *et al.*, "Analyzing the shear strength of jointed magmatic rock mass excavatability using the hybridization of metaheuristic model of ELM-SVM," *Acta Geotech*, vol. 18, no. 4, pp. 1793 – 1819, 2023, doi: 10.1007/s11440-022-01596-4.

[9] L. O. Carvalho and D. B. Ribeiro, "A multiple model machine learning approach for soil classification from cone penetration test data," *Soils and Rocks*, vol. 44, no. 4, 2021, doi: 10.28927/SR.2021.072121.

[10] G. R. Khanlari, M. Heidari, A. A. Momeni, and Y. Abdilor, "Prediction of shear strength parameters of soils using artificial neural networks and multivariate regression methods," *Eng Geol*, vol. 131–132, pp. 11–18, Mar. 2012, doi: 10.1016/j.enggeo.2011.12.006.

[11] J. Rahman, K. S. Ahmed, N. I. Khan, K. Islam, and S. Mangalathu, "Data-driven shear strength prediction of steel fiber reinforced concrete beams using machine learning approach," *Eng Struct*, vol. 233, Apr. 2021, doi: 10.1016/j.engstruct.2020.111743.

[12] A. Gajurel, B. Chittoori, P. S. Mukherjee, and M. Sadegh, "Machine learning methods to map stabilizer effectiveness based on common soil properties," *Transportation Geotechnics*, vol. 27, 2021, doi: 10.1016/j.trgeo.2020.100506.

[13] P. Zhang, Y.-F. Jin, and Z.-Y. Yin, "Machine learning–based uncertainty modelling of mechanical properties of soft clays relating to time-dependent behavior and its application," *Int J Numer Anal Methods Geomech*, vol. 45, no. 11, pp. 1588 – 1602, 2021, doi: 10.1002/nag.3215.

[14] V. L. Pacheco, L. Bragagnolo, F. Dalla Rosa, and A. Thomé, "Cone Penetration Test Prediction Based on Random Forest Models and Deep Neural Networks," *Geotechnical and Geological Engineering*, 2023, doi: 10.1007/s10706-023-02535-0.

[15] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and Other Machine Learning Algorithms," in *Proceedings of the 11th International Conference on Machine Learning, ICML 1994*, 1994, pp. 53 – 61. doi: 10.1016/B978-1-55860-335-6.50015-5.

[16] T. G. Dietterich, "Ensemble methods in machine learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1857 LNCS, pp. 1 – 15, 2000, doi: 10.1007/3-540-45014-9_1.

[17] F. Pérez-Cruz and A. Artés-Rodríguez, "A new optimizing procedure for ν-support vector regressor," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2001, pp. 1265 – 1268. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034854360&partnerID=40&md5=b9ae93f5df17e1b90cb1ea1b97dc9a84

[18] V. Roth, "The generalized LASSO," *IEEE Trans Neural Netw*, vol. 15, no. 1, pp. 16 – 28, 2004, doi: 10.1109/TNN.2003.809398.

[19] X. Wang and D. J. Brown, "Boosting orthogonal least squares regression," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3177, pp. 678 – 683, 2004, doi: 10.1007/978-3-540-28651-6_100.

[20] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 23, no. 1, pp. 20 – 31, 2015, doi: 10.1109/TASLP.2014.2367814.

[21] A. Criminisi *et al.*, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med Image Anal*, vol. 17, no. 8, pp. 1293 – 1303, 2013, doi: 10.1016/j.media.2013.01.001.

[22] R. Zhang, Y. Li, A. T. C. Goh, W. Zhang, and Z. Chen, "Analysis of ground surface settlement in anisotropic clays using extreme gradient boosting and random forest regression models," *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 13, no. 6, pp. 1478 – 1484, 2021, doi: 10.1016/j.jrmge.2021.08.001.

[23] S. Kaymak and I. Patras, "Multimodal random forest based tensor regression," *IET Computer Vision*, vol. 8, no. 6, pp. 650 – 657, 2014, doi: 10.1049/iet-cvi.2013.0320.

[24] T.-A. Nguyen, H.-B. Ly, and B. T. Pham, "Backpropagation Neural Network-Based Machine Learning Model for Prediction of Soil Friction Angle," *Math Probl Eng*, vol. 2020, 2020, doi: 10.1155/2020/8845768.

[25] T.-A. Nguyen, H.-B. Ly, and B. T. Pham, "Backpropagation Neural Network-Based Machine Learning Model for Prediction of Soil Friction Angle," *Math Probl Eng*, vol. 2020, 2020, doi: 10.1155/2020/8845768.

[26] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/4832864.

[27] P. Samui, N.-D. Hoang, V.-H. Nhu, M.-L. Nguyen, P. T. T. Ngo, and D. T. Bui, "A new approach of hybrid bee colony optimized neural computing to estimate the soil compression coefficient for a housing construction project," *Applied Sciences (Switzerland)*, vol. 9, no. 22, 2019, doi: 10.3390/app9224912.

[28] Z. H. Zhou, *Ensemble methods: Foundations and algorithms*. CRC Press, 2012. doi: 10.1201/b12207.

[29] A. Rabbani, P. Samui, and S. Kumari, "Implementing ensemble learning models for the prediction of shear strength of soil," *Asian Journal of Civil Engineering*, 2023, doi: 10.1007/s42107-023-00629-x.

[30] K. Mamudur and M. R. Kattamuri, "Application of Boosting-Based Ensemble Learning Method for the Prediction of Compression Index," *Journal of The Institution of Engineers (India): Series A*, vol. 101, no. 3, pp. 409 – 419, 2020, doi: 10.1007/s40030-020-00443-7.