Predicting Soil Compression Index Using Random Forest and Gradient Boosting Tree

Yu Huat.Chia
University of Malaya
Department of Civil Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, 50603, Malaysia. Email: 17107717@siswa.um.edu.my

Danial Jahed Armaghani
School of Civil and Environmental Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia. Email: danial.jahedarmaghani@uts.edu.au

Sai Hin. Lai
University of Malaya
Department of Civil Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, 50603, Malaysia. Email: laish@um.edu.my

ABSTRACT

In the construction on soft ground, settlement prediction is essential for effective treatment and mitigation. Soil compressibility is a crucial property for analyzing the settlement of soil layers under applied loads. The compression index is one of the crucial geotechnical parameters that quantifies soil deformation under load. The compression index represents the slope of the curve of the void ratio against the logarithm of the effective pressure. The conventional approach to determining the compression index through oedometer tests is both time-consuming and costly. This paper employs a dataset comprising laboratory soil data obtained from various tests, including Atterberg limit, oedometer, and moisture content tests, conducted on Alluvium soil samples from Malaysia. Additionally, relevant data from literature sources is incorporated to augment the analysis. The primary objective of this study is to develop predictive models for estimating the compression index using tree-based machine learning algorithms, namely random forest and gradient boosting tree. To evaluate the performance of these models, the results obtained from the machine learning models are compared with those derived from empirical formulas commonly used in the field. The findings show that the machine learning methods outperform the empirical formula in predicting compression index, indicating the potential of these techniques to determine geotechnical parameters.

KEYWORDS: Compression index, Random Forest, Gradient Boosting Tree, Machine Learning

1. Introduction

In geotechnical design for the foundations, it is crucial to determine the compressibility of the soils as this factor plays an important role for the settlement of the soil layers due to external load. Soil compressibility refers to the reduction in volume experienced by soils due to the drainage of pore water when subjected to a load (Craig and Knappett, 2012). To calculate the compressibility of soils, several key parameters need

to be determined, including the compression index (c_c), coefficient of compressibility (c_v), and coefficient of volume change (m_v). In order to determine the consolidation parameters of fine-grained soil, there is a need to conduct Oedometer tests. However, the process of carrying out this test is time-consuming and costly (Ozer et al., 2008). In addition, sample preparation is a real challenge for this test in the laboratory. In the field of geotechnical engineering, practising engineers employ correlations and empirical relationships to estimate geotechnical parameters, including the c_c. Many researchers have used various soil parameters to correlate with c_c such as water content, liquid limit, plasticity index and void ratio (Skempton and Jones, 1944; Huang et al., 2019; Mandhour, 2020). In addition, researchers also employed the Atterberg limit test to establish correlations between soil properties and geotechnical engineering characteristics (Shouka, 1964; Yoon et al., 2004)). This test is particularly useful as it serves as an initial characterization of the soils. Moreover, the plasticity-related parameters are influenced by the electro-chemical behavior of clay minerals, which subsequently impact water retention potential and mechanical properties of the soil (Carter and Bentley, 1991). Many studies have used traditional statistical methods to develop the empirical formulas with several assumptions. Notably, the Atterberg limit is consistently incorporated into these formulas as a significant contributing factor.

Due to the rapid development of machine learning (ML), many researchers have started to integrate ML application in geotechnical engineering such as prediction of pile capacity (Khanmohammadi et al., 2022), slope stability (Lin et al., 2018), determination of geotechnical parameters (Pham et al., 2021) tunnelling (Armaghani et al., 2017; Zhou et al., 2020) and many more. The use of ML for prediction of c_c has been explored by several techniques such as Artificial Neural Network (ANN) by Park and Lee (2011) and Expression Programming by Mohammadzadeh et al. (2014). However, these techniques include some drawbacks such as getting trapped in local minima and slow convergence (Gordan et al., 2016; Hasanipanah et al., 2016). On the other hand, tree-based ML techniques, such as Random Forest (RF) and Gradient Boosting Tree (GBT), have been successfully applied in many geotechnical applications with high level of performance (Li et al., 2022; Liu et al., 2022; Yari et al., 2023). This due to their ability to: (i) avoid subjective uncertainty, (ii) handle large data points at greater modelling speed, (iii) select the most influential factors on output, and (iv) visualize nonlinear data.

In this paper, we aim to compare the performance of RF, GBT and empirical formulas for predicting the c_c. A dataset comprising 116 observations from Oedometer, Atterberg limit, and moisture content tests conducted on the Alluvium formation in Malaysia will be utilized for predicting the c_c. Additionally, supplementary data from Mandhour (2020) will be incorporated to further enhance the analysis. The combined dataset will facilitate a comprehensive evaluation of the c_c prediction models, considering a broader range of samples.

2. Review of previous studies

Numerous researchers have conducted studies on the prediction of the c_c for various soil types, employing empirical formulas that utilize different parameters. These parameters include liquid limit, natural water content, plasticity index, void ratio, and multiple variables. A compilation of such empirical formulas, along with their corresponding parameters, is presented in Table 1, providing a general overview of the existing empirical formula in this area.

Table 1. Empirical formula computation for computation of c_c using various parameters

No	Author	Equation (s)	Type of Soils
Liquid Limit			
1	Shouka (1964)	$C_c = 0.017\ (LL - 20)$	All clays
2	Terzaghi and Peck (1967)	$C_c = 0.009\ (LL - 10)$	Normally consolidated clay
Plasticity index			
3	Wroth and Wood (1978)	$C_c = \frac{PI}{74}$	All clays
4	Yoon et al (2004)	$C_c = 0.014(PI) + 0.165$	Busan clay
Moisture Content			
5	Rutledge (1958)	$C_c = 0.0115\ (MC)$	Soft clays
6	Bowles (1979)	$C_c = 0.115\ (MC)$	Organic silts and clays
Void ratio			
7	Rendon-Herrero (1980)	$C_c = 0.3(e_o - 0.27)$	All soil types
8	Yoon et al. (2004)	$C_c = 0.39(e_o - 0.13)$	Busan clay
Multiple variables			
9	Koppula (1981)	$C_c = 0.009(MC) + 0.005(LL)$	All clays
10	Yoon et al. (2004)	$C_c = -0.194e_o + 0.0098LL - 0.0025PI - 0.256$	Busan clay

Table 1 reveals that most of the studies have employed empirical prediction equations derived from linear or multiple linear regression analysis. The utilization of regression analysis entails certain limitations and uncertainties due to the inherent simplification of the model. Conventional regression methods, such as linear and multiple regression, assume a predefined relationship between the input and output. This assumption inherently imposes constraints on the model's flexibility and may introduce potential limitations in capturing complex relationships within the data. Moreover, researchers have also explored the application of various ML models for predicting the c_c, as shown in Table 2**.** The outcomes indicate that different ML models exhibit diverse prediction capabilities and demonstrate promising potential in estimating c_c. This highlights the effectiveness of ML techniques in enhancing the prediction accuracy and performance compared to traditional empirical formulas or regression-based approaches.

Table 2. AI Method prediction of Compression index

Author (s)	ML	No Data	Predictor (s)	Performance of the prediction testing dataset
Mohammadzadeh S et al. (2016)	Multi-gene genetic programming	101	e_o, LL and PL	R^2 = 0.840
Mohammadzadeh S et al. (2019)	Gene Expression Programming	108	e_o, LL and PL	R^2 = 0.860
Park and Lee (2011)	Artificial Neural Network	947	MC, e_o, LL, PI, G_s , w_{sand}, w_{silt}, w_{clay}	R^2 = 0.885
Mohammadzadeh et al. (2014)	Expression programming	101	e_o, LL and PL	R^2 = 0.812
Fikret Kurnaz and Kaya (2018)	Extreme Learning Machine (ELM) Bayesian Regularisation Neural Network (BRNN) Support Vector Machine (SVM)	351	MC, e_o, LL, PI	ELM, R^2 = 0.890 SVM, R^2 = 0.915 BRNN, R^2 = 0.915

where G_s is specific gravity, w_{sand} is percentage of sand w_{silt} is percentage of silt w_{clay} is percentage of clay, R^2 is coefficient of determination.

3. Model background

3.1 Random Forest (RF)

Random Forest (RF) is a versatile algorithm capable of solving both classification and regression problems. Introduced by Breiman (2001) , RF combines decision trees by generating each tree from a random vector sample independently. This randomness helps overcome the limited generalization ability of a single decision tree. Bagging, which involves bootstrapping the training data, and random feature selection are integral parts of RF. In classification tasks, RF predicts the class by majority voting among the individual trees. For regression tasks, the predictions from each tree are averaged to obtain the result. By combining the predictions of multiple trees, RF improves accuracy and robustness. RF's performance is influenced by several hyperparameters, such as the number of decision trees (ntree) and the maximum depth of tree (max_depth). To identify the optimal hyperparameters, a grid search method can be employed. Grid search exhaustively explores all combinations of hyperparameters to find the best configuration of the hyperparameters. In the context of this study, the range of ntree values considered is (50, 100, 150, 200, 250, 300, 350, 400, 450, and 500), while max_depth values range from 2 to 10 at interval of 2. By systematically evaluating these hyperparameters based on grid search method, the optimal combination for these hyperparameter can be identified.

3.2 Gradient Boosting Tree (GBT)

Introduced by Friedman (2001), GBT builds decision trees in a sequential manner, where each subsequent tree is trained to correct the errors made by the previous trees. The key idea behind GBT is to iteratively fit the new trees to the negative gradient of the loss function of the previous trees' predictions. This process helps to improve the overall model's performance with each iteration. During the training phase, GBT starts with an initial model, usually a weak learner like a shallow decision tree. Subsequent trees are then added

to the ensemble by minimizing a loss function that quantifies the discrepancy between the predicted values and the true values. The learning process is guided by optimization techniques such as gradient descent, which updates the model's parameters to minimize the loss function. To make a prediction using the trained Gradient Boosting Tree, the individual trees' predictions are combined by weighted averaging or summation. The weights reflect the importance of each tree's contribution to the final prediction. The result is obtained by summing up the predictions from all the trees. GBT also involves hyperparameter tuning to achieve optimal performance. The hyperparameters that require tuning for GBT, such as ntree and max_depth have similar ranges as RF. Additionally, GBT incorporates an additional hyperparameter, learning rate, which will be explored within the range of 0.001, 0.01, 0.1, 0.2, and 0.3. Similarly with RF, this ML model will be using grid search.

4. Data and parameters

Liquid Limit (LL) is the moisture content (%) at which a soil transitions from a plastic state to liquid state whereas Plastic Limit (PL) is the moisture content (%) at which the soil changes from semi solid to solid state. The difference between liquid limit and the plastic limit is known as the plasticity index (PI = LL - PL), which provides an indication of the soil's plasticity. Void ratio (e) is the ratio of the volume of voids to the volume of solids in a soil sample. It is an indicator of the soil's porosity and its ability to compress under load. Soil compressibility is the volume reduction under load of pore water drainage. c_c is a measure of the compressibility of a soil. It can be obtained from the slope of the curve void ratio versus logarithm of effective pressure. Some of the factors affect the compression index such as grain size of the soils, plasticity, and overburden pressure. Fine-grained soils, such as clays, have a higher compression index than coarse-grained soils, such as sands. This is because fine-grained soils have larger surface area, which allow to absorb more water and swell. Soils with high plasticity index is prone to deformation by stress. Increasing overburden pressure leads to tighter packing of soil particles, resulting in a reduction of void ratio. Since the Atterberg limits is the measure of plasticity of the soil and void ratio amount of the voids in the soil, thus, these parameters affect the compression index. The natural moisture content (MC) refers to the moisture content of the soil in its undisturbed state, without any external addition or removal of water. Since water content affects the soil's particle fill, it can be one of the factors influencing compressibility. Furthermore, based on the information presented in Table 1 and Table 2 , it is evident that the authors in previous studies utilized parameters such as LL, PL, e_o, and MC to determine the c_c. Therefore, these parameters will be considered in these analyses. The analysis will incorporate a total of 116 laboratory test data points, including Atterberg limit, Moisture Content test, and Oedometer test data, obtained from the Alluvium formation of Malaysia for Analysis 1. In Analysis 2, soil data from Malaysia will be combined with the soil data from Al-Nasiriya city provided by Mandhour (2020) to augment the dataset for the analysis model and add variety range of soil. Table 3 summarizes two types of predictors that will be used to predict the c_c.

Table 3. Two different analyses of predictors used for the prediction of the c_c.

Analysis	Predictors	No of Data
1	e_o, LL , PL and MC	116
2	e_o, LL , PL	137

5. Analysis and results

For the purpose of RF and GBT analysis, the database will be divided into two subsets: 80% training dataset and 20% testing dataset. To address the limitations imposed by small-sized datasets when splitting into training and testing sets, a robust technique called k-fold cross-validation (CV) is employed. In k-fold CV,

the original dataset is randomly partitioned into k subsets or folds. Subsequently, k-1 folds are utilized for training the model, while the remaining fold is used for validation. This process is repeated k times, with each fold serving as the validation set once. By averaging the prediction errors across the k subsets, the model's performance can be accurately evaluated. K-fold CV maximizes the utilization of available data as every part of the original dataset is randomly assigned to both training and testing, enhancing the reliability of the model construction and validation process. In this study, a 10-fold CV approach was adopted. For each 10^{th} CVs have been carried out, different hyperparameter is used using grid search method to determine the optimum hyperparameter. Each of the hyperparameter will be assessed for the accuracy based on the coefficient of determination, R^2. Figure 1 illustrates the mechanism of the cross validations.

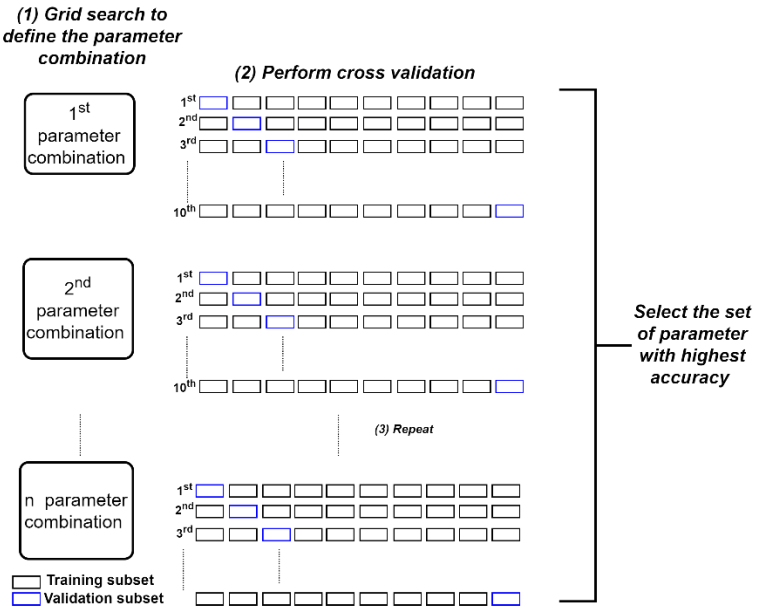

Figure 1. Illustration of the cross-validation mechanism

Using the grid search method, the optimal hyperparameters for Analysis 1 are found to be max_depth = 2 and n_tree = 100 for RF, and learning_rate = 0.01, max_depth = 2, and n_tree = 150 for GBT. Similarly, for Analysis 2, the optimal hyperparameters are determined as max_depth = 2 and n_tree = 100 for RF, and learning_rate = 0.01, max_depth = 2, and n_tree = 250 for GBT. These optimum hyperparameters shall be used for the testing datasets. Figure 2 and Figure 3 display the prediction capabilities of RF and GBT for Analysis 1 and 2 respectively. Additionally, Figure 4 illustrates the relationship between computed compression index using empirical formulas (Shouka, 1964; Wroth and Wood, 1978; Rendon-Herrero, 1980) and the measured compression index.

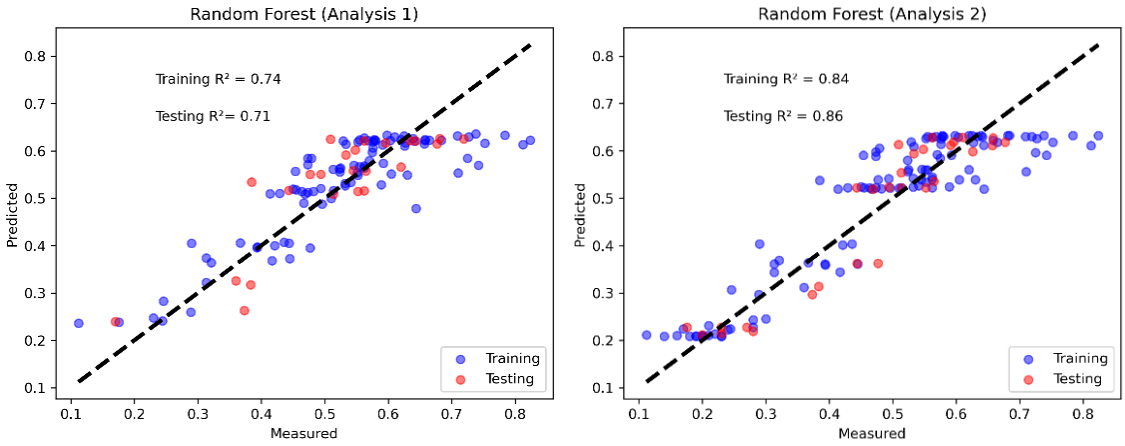

Figure 2. RF Prediction model for Analysis 1 and 2

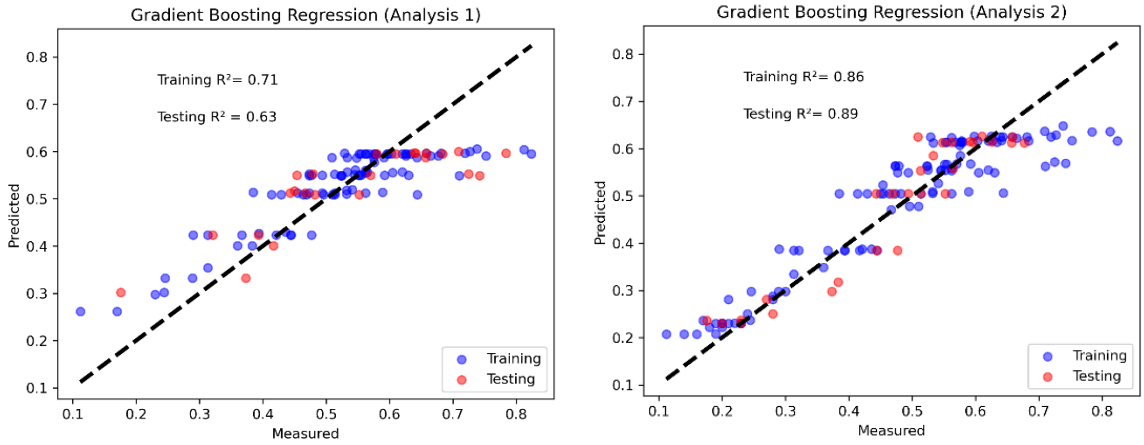

Figure 3. GBT Prediction model for Analysis 1 and 2

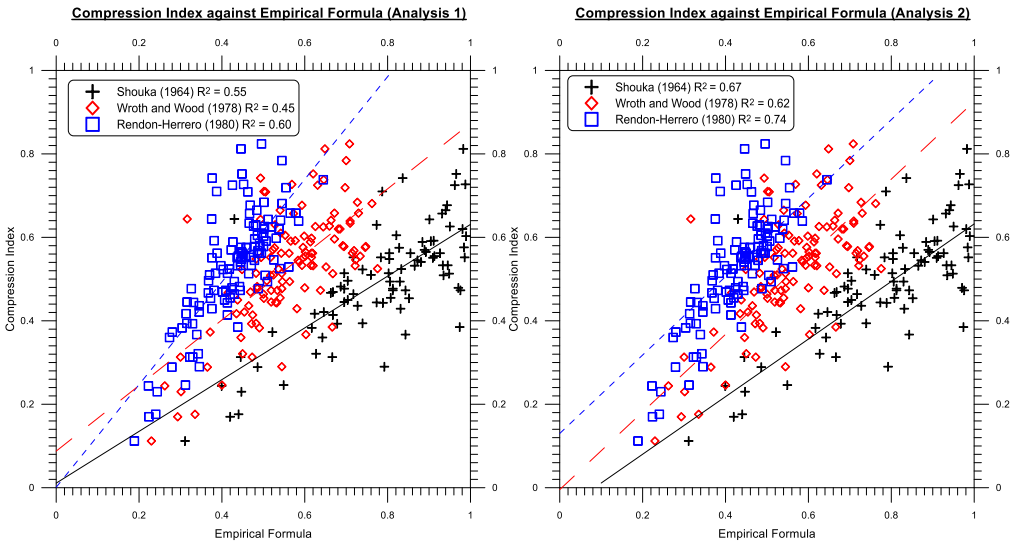

Figure 4. Measured and computed compression index

The comparison of prediction performance based on R^2, as observed from Figure 3, Figure 4 and Figure 4 clearly indicates that the RF and GBT models outperform the empirical formula in predicting the compression index. Moreover, when comparing RF and GBT, it becomes evident that GBT exhibits superior predictive capabilities. Furthermore, the predictors LL, PL, and e_o, when considered together, yield significant findings for computing the compression index. This implies that the inclusion of MC is not necessarily required for predicting the compression index. Notably, other authors (Mohammadzadeh et al., 2016, 2019) have also utilized these three parameters (LL, PL, and e_o) in combination with different ML models, and have achieved promising predictive results. Lastly, it was observed that Analysis 2, which utilized a larger dataset, demonstrated superior prediction performance compared to Analysis 1. This finding suggests that a greater amount of data is essential for establishing more accurate predictions using both ML models and empirical formulas. The improved performance in Analysis 2 highlights the significance of a comprehensive and diverse dataset in training ML models, enabling them to capture the underlying patterns and relationships more effectively. Additionally, it emphasizes the importance of considering larger datasets when developing ML models and empirical formulas to enhance the predictive capabilities. This insight underscores the value of data quantity and quality in achieving more reliable and precise predictions in geotechnical engineering applications.

6. Conclusions and future studies

In conclusion, this study explored the application of Random Forest (RF) and Gradient Boosting Tree (GBT) models for predicting the Soil Compression Index (c_c) and compared the performance against empirical formulas where the conclusion are as follows:

1) Both RF and GBT models exhibited better predictive capabilities compared to traditional empirical formulas. Notably, GBT displayed even stronger predictive power than RF, highlighting its potential as a preferred choice for determining the c_c parameter.
2) Furthermore, the findings suggest that tree-based machine learning models, particularly GBT, can effectively utilize parameters such as Atterberg limits (liquid limit and plastic limit) and void ratio to accurately predict the Soil Compression Index. This finding enhances our understanding of the essential factors contributing to soil compression behavior and emphasizes the importance of incorporating comprehensive soil properties in predictive models.

3) It is crucial to acknowledge that further data collection and testing are necessary to optimize the predictive accuracy of the machine learning model for the c_c. The inclusion of more diverse and extensive datasets from various soil types and environmental conditions would help ensure the model's robustness and generalizability for real-world applications.

Lastly, other tree based techmiques such as Extreme Gradient Boosting Tree, Adaboost and Categorical Boosting can be considered to further explores by different researchers in order to establish robust ML moed. However, it is imperative to emphasize the importance of acquiring more diverse and extensive datasets before inputting these ML models.

References

Armaghani, D. J., Mohamad, E. T., Narayanasamy, M. S., Narita, N., & Yagiz, S. (2017). Development of hybrid intelligent models for predicting TBM penetration rate in hard rock condition. *Tunnelling and Underground Space Technology*, *63*, 29–43. https://doi.org/https://doi.org/10.1016/j.tust.2016.12.009

Bowles, J. E. (1979). *Physical and geotechnical properties of soils*.

Breiman, L. (2001). *Random Forests*. *45*, 5–32.

Carter, M., & Bentley, S. P. (1991). *Correlations of soil properties.* Pentech press publishers.

Craig, R. F., & Knappett, J. a. (2012). *Craig's Soil Mechanics, 8th Edition*.

Fikret Kurnaz, T., & Kaya, Y. (2018). The comparison of the performance of ELM, BRNN, and SVM methods for the prediction of compression index of clays. *Arabian Journal of Geosciences*, *11*, 1–14.

Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*.

Gordan, B., Jahed Armaghani, D., Hajihassani, M., & Monjezi, M. (2016). Prediction of seismic slope stability through combination of particle swarm optimization and neural network. *Engineering with Computers*, *32*, 85–97.

Hasanipanah, M., Noorian-Bidgoli, M., Jahed Armaghani, D., & Khamesi, H. (2016). Feasibility of PSO-ANN model for predicting surface settlement caused by tunneling. *Engineering with Computers*, *32*(4), 705–715. https://doi.org/10.1007/s00366-016-0447-0

Huang, C., Li, Q., Wu, S., Liu, Y., & Li, J. (2019). Assessment of empirical equations of the compression index of muddy clay: sensitivity to geographic locality. *Arabian Journal of Geosciences*, *12*, 1–13.

Khanmohammadi, M., Armaghani, D. J., & Sabri Sabri, M. M. (2022). Prediction and Optimization of Pile Bearing Capacity Considering Effects of Time. *Mathematics*, *10*(19), 3563.

Koppula, S. D. (1981). Statistical estimation of compression index. *ASTM Geotechnical Testing Journal*, *4*(2).

Li, D., Liu, Z., Armaghani, D. J., Xiao, P., & Zhou, J. (2022). Novel Ensemble Tree Solution for Rockburst Prediction Using Deep Forest. *Mathematics*, *10*(5), 787.

Lin, Y., Zhou, K., & Li, J. (2018). Prediction of slope stability using four supervised learning methods. *IEEE Access*, *6*(c), 31169–31179. https://doi.org/10.1109/ACCESS.2018.2843787

Liu, Z., Armaghani, D. J., Fakharian, P., Li, D., Ulrikh, D. V., Orekhova, N. N., & Khedher, K. M. (2022). Rock Strength Estimation Using Several Tree-Based ML Techniques. *CMES-Computer Modeling in Engineering & Sciences*, *133*(3).

Mandhour, E. A. (2020). Prediction of compression index of the soil of Al-nasiriya city using simple linear regression model. *Geotechnical and Geological Engineering*, *38*(5), 4969–4980.

Mohammadzadeh, D., Bazaz, J. B., & Alavi, A. H. (2014). An evolutionary computational approach for formulation of compression index of fine-grained soils. *Engineering Applications of Artificial Intelligence*, *33*, 58–68.

Mohammadzadeh S, D., Bolouri Bazaz, J., Vafaee Jani Yazd, S. H., & Alavi, A. H. (2016). Deriving an

intelligent model for soil compression index utilizing multi-gene genetic programming. *Environmental Earth Sciences*, *75*, 1–11.

Mohammadzadeh S, D., Kazemi, S.-F., Mosavi, A., Nasseralshariati, E., & Tah, J. H. M. (2019). Prediction of compression index of fine-grained soils using a gene expression programming model. *Infrastructures*, *4*(2), 26.

Ozer, M., Isik, N. S., & Orhan, M. (2008). Statistical and neural network assessment of the compression index of clay-bearing soils. *Bulletin of Engineering Geology and the Environment*, *67*, 537–545.

Park, H. Il, & Lee, S. R. (2011). Evaluation of the compression index of soils using an artificial neural network. *Computers and Geotechnics*, *38*(4), 472–481.

Peck, R. B. (1967). *Soil Mechanics in Engineering Practice, 2nd Edition. Edited by Karl Terzaghi and R.B. Peck*. John Wiley & Sons. https://books.google.com.my/books?id=AnbHzgEACAAJ

Pham, B. T., Nguyen, M. D., Nguyen-Thoi, T., Ho, L. S., Koopialipoor, M., Quoc, N. K., Armaghani, D. J., & Van Le, H. (2021). A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling. *Transportation Geotechnics*, *27*, 100508.

Rendon-Herrero, O. (1980). Universal compression index equation. *Journal of the Geotechnical Engineering Division*, *106*(11), 1179–1200.

Rutledge, P. C. (1958). *Study of deep soil stabilization by vertical sand drains*.

Shouka, H. (1964). Relationship of compression index and liquid limit of alluvial clay. *Proceedings of the 19th Japan Civil Engineering Conference. Touhoku*, 30–31.

Skempton, A. W., & Jones, O. T. (1944). Notes on the compressibility of clays. *Quarterly Journal of the Geological Society*, *100*(1–4), 119–135.

Wroth, C. P., & Wood, D. M. (1978). The correlation of index properties with some basic engineering properties of soils. *Canadian Geotechnical Journal*, *15*(2), 137–145.

Yari, M., Armaghani, D. J., Maraveas, C., Ejlali, A. N., Mohamad, E. T., & Asteris, P. G. (2023). Several Tree-Based Solutions for Predicting Flyrock Distance Due to Mine Blasting. *Applied Sciences*, *13*(3), 1345.

Yoon, G. L., Kim, B. T., & Jeon, S. S. (2004). Empirical correlations of compression index for marine clay from regression analysis. *Canadian Geotechnical Journal*, *41*(6), 1213–1221.

Zhou, J., Qiu, Y., Armaghani, D. J., Zhang, W., Li, C., Zhu, S., & Tarinejad, R. (2020). Predicting TBM penetration rate in hard rock condition: A comparative study among six XGB-based metaheuristic techniques. *Geoscience Frontiers*, https://doi.org/10.1016/j.gsf.2020.09.020.