Are We Ready to Use AI Technologies for the Prediction of Soil Properties?

Ryan Yan*

AECOM Asia Co. Ltd., Hong Kong, China

*Corresponding author
doi: https://doi.org/10.21467/proceedings.133.35

ABSTRACT

Artificial intelligence (AI) has become a hot topic for different professions in which geotechnical engineering is no exception. It is anticipated that AI could perform tasks, solve complex problems and make decision by mimicking intelligence or behavioral pattern of humans or any other living entities. Attempts have been made to study and adopt AI technologies in geotechnical engineering. In this paper, a dataset of marine soil in South Korea is re-analyzed using different commonly adopted AI algorithms. The soil's compressibility is considered as the dependent variable (i.e., to be predicted) while other soil index and physical properties are regarded as the independent variables. The data are split into the training and validation set. While an algorithm learns from the training set, its prediction performance is examined using the validation set. Then, the Bayesian model class approach has been used to explain the potential problem of the use of AI algorithm to predict soil properties. At the end, by using this study as an example, the author discusses from a partitioner's perspective how AI could affect our professions. In particularly, the question "are we ready for using AI to predict soil properties" is addressed.

Keywords: Machine Learning, Compressibility, Prediction

1 Introduction

1.1 Overview

Artificial Intelligence (AI) can be referred to as the intelligence exhibited by machines or software capable of performing tasks, solving complex problems or making decision by mimicking intelligence or behavioral pattern of humans or any other living entity. There are many AI-related terminologies in which people often find confusing. Figure 1 shows a simple diagram to illustrate the relations among these terminologies. Machine Learning (ML), a subset of AI, is a technique by which a computer program can perform prediction without the use of any prescribed set of rules including mainly for instance statistical theories. This approach trains a predictive model from data. In a layman's term, the data will tell you the underlying pattern or governing rules, if any. Neural Network (NN) or in many studies people call this Artificial Neural Network (ANN), a subset of ML, is a technique to perform machine learning inspired by our brain's own network of neurons. In the case when multiple layers of neurons are used, it is called the Deep Neural Network (DNN).

In general, a key concept of AI is to convert data into value. AI has been embedded in our daily lives. Examples include Siri, automates driving, robot-advisors, email spam filtering, Netflix recommendations, facial detection and recognition, etc. It is believed that soon or later AI will dominate the area of data analytics.

1.2 Applications of AI Technologies to Geotechnical Engineering

AI technologies appear very appealing, and attempts have been made by geotechnical engineers to employ the techniques to solve engineering problems. Generally speaking, the development of an AI application involves the following key stages:

Proceedings DOI: 10.21467/proceedings.133; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-957605-1-0

- Collection of data which can be structured, unstructured, or both;
- Data conditioning including for example dimensionality reduction, outlier detection, looking for biases in the collection, highlighting incomplete data, etc.;
- Learning via algorithm;
- Validation and prediction;
- Publishing.

The applications of AI to geotechnical engineering can generally be divided into 4 categories. They are material behavior, system performance, classification, and automation. Material behavior refers to as the prediction of geomaterial properties using predominantly ML and ANN algorithms. System performance refers to the use of AI technologies to examine/predict the performance of an engineered or natural system, for example the prediction of damages due to natural disasters such as earthquakes and landslides. Classification refers to the categorization of features, profiles and systems. Typical examples include cracks and landslides recognition from images, classification of soil type and geological profile, etc. AI automation refers to the decision making based on defined rules and experience. In Hong Kong, AI automation has been applied mostly to site safety monitoring. Over years, increasing efforts have been spent to apply AI technologies in particularly ML and ANN to geotechnical engineering (Shahin, 2016; Ebid, 2021; Jaksa and Liu, 2021; Jong et al., 2021; Zhang et al., 2021a,b; and many more).

This study first presents the use of various AI algorithms to predict the soil compressibility of marine soils based on a database compiled from literature. Performance of the prediction is discussed. Based on the findings, the author then shares his thoughts on whether AI-based prediction of soil properties should be widely adopted in the profession.

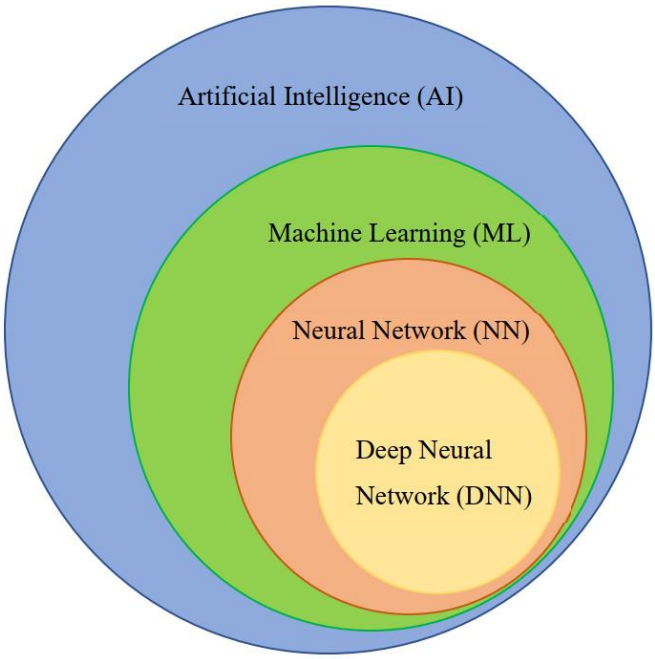

Figure 1: Relations among different terminologies in the family of AI.

2 Prediction of Soil Compressibility

2.1 Database

A comprehensive dataset containing the compression index C_c and other soil properties including the in-situ water content w_n, initial void ratio e_0, liquid limit LL, plasticity index PI, specific gravity G_s, and soil dry density ρ_d of marine clays in the coasts of South Korea is re-examined. The dataset contains

Proceedings DOI: 10.21467/proceedings.133
ISBN: 978-81-957605-1-0

Series: AIJR Proceedings
ISSN: 2582-3922

223, 274 and 298 complete sets of records from the east, south and west coast of South Korea, respectively. Figure 1 shows the variation of C_c with each soil parameter. Clearly, a large range of soil compressibility can be identified in the east and south coast data while that in the east coast covers a smaller range. More details of the sites and soils can be found in Yoon et al., (2004). In the following prediction analysis, the compression index is considered as the dependent variable while the remaining soil properties are considered as independent variables. In other words, the compression index will be predicted using the independent soil properties.

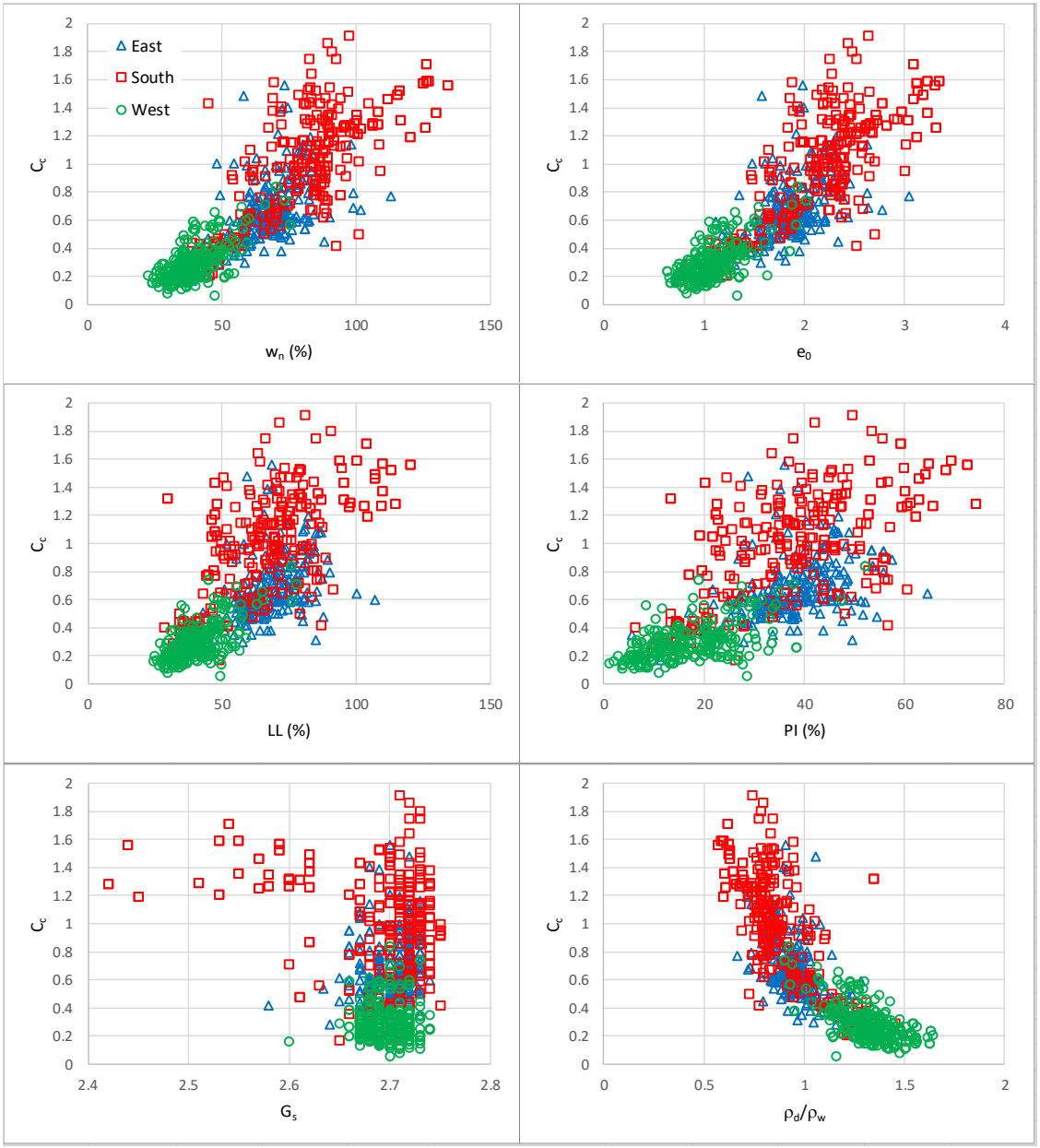

Figure 2: Variation of compression index with soil properties.

2.2 Readily Available Solutions using Microsoft Machine Learning Studio (classic)

Released in 2015, Microsoft Machine Learning Studio (classic), refers to as ML Studio hereafter, was the first drag-and-drop machine learning model builder in Microsoft Azure. It is a standalone service that offers a visual experience of ML. Microsoft Azure Machine Learning, however, is a separate service that delivers a complete data science platform. It is a cloud-based service to manage machine learning projects from model development, training, deployment and managing Machine Learning

Operations (MLOps). Users can create a model in Azure Machine Learning or use a model built from an open-source platform. The Azure Machine Learning Studio is a graphical user interface for a project workspace. In this study, ML Studio is employed to develop readily-to-be-used ML solutions for the prediction of soil compressibility.

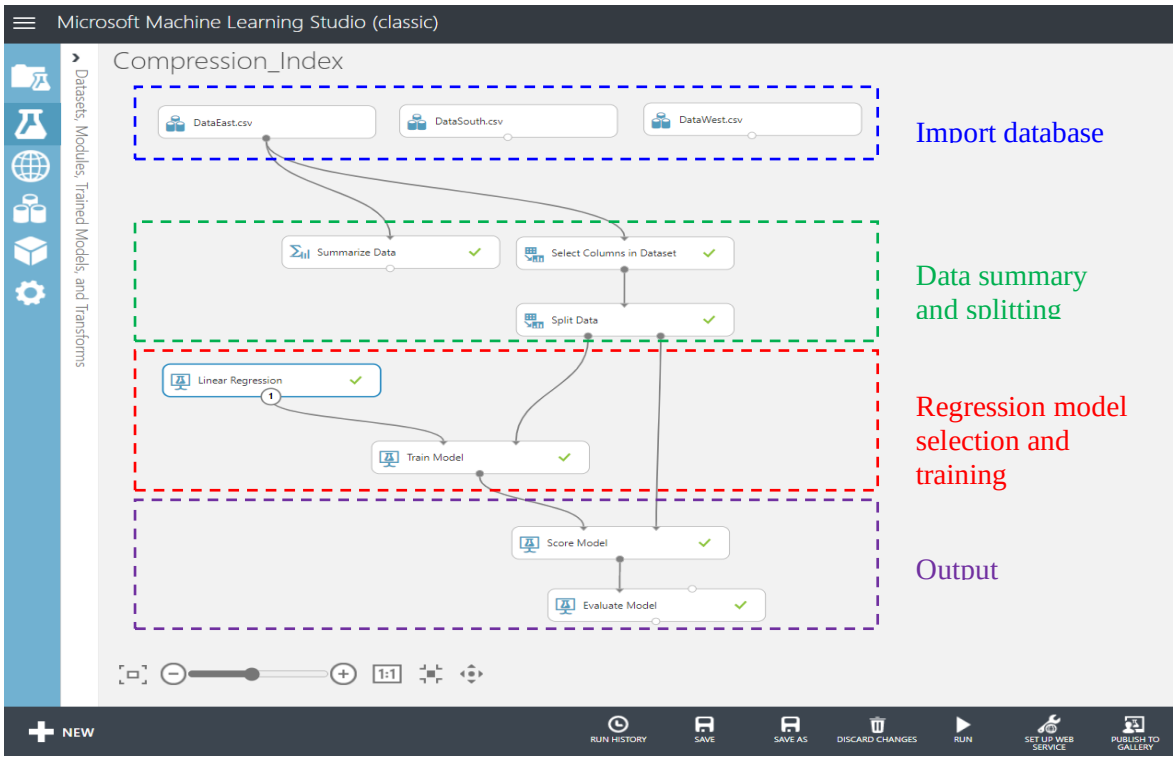

Figure 3: An example workflow in ML Studio.

The prediction of compression index falls into the regression category in ML. Many readily available regression models have been deployed in the ML Studio. For example:

- Linear Regression
- Bayesian Linear Regression
- Boosted Decision Tree Regression
- Decision Forest Regression
- Neural Network Regression
- Poisson Regression

The above regression models are used in this study to predict compression index from the independent soil properties. The dataset is divided into 2 groups: namely the training set which contains 70% of the entire dataset and the validation set containing the remaining 30%. This ratio of splitting is a common practice in ML. In most cases, default setting of the built-in models is adopted. It aims to mimic users with little experience or understanding of each ML algorithm. Its impact and implication will be discussed later in this paper. Figure 3 shows how the ML Studio graphic interface looks like. The drag-and-drop nature of the platform can be readily seen. As illustrated, the workflow can be divided into 4 parts: (i) import database, (ii) data summary and splitting; (iii) regression model selection and training; and (iv) output. It is worth noting that the use of ML studio requires nearly no experience of program coding.

In this study, the ordinary least squares method is adopted as the solution scheme for the linear regression model. This method attempts to minimize the sum of the squared residuals to evaluate the values of the fitting coefficients. The Bayesian linear regression method in this study refers to the use

Proceedings DOI: 10.21467/proceedings.133
ISBN: 978-81-957605-1-0

Series: AIJR Proceedings
ISSN: 2582-3922

of Bayesian inference to evaluate the model fitting parameters. It is assumed that the errors of regression model possess a normal distribution and the posterior probability distributions of the model parameters are evaluated based on a prior distribution of the parameters. Weight regularization is used to reduce overfitting. Boosted decision tree method builds a series of trees in a step-wise fashion and then selects the optimal tree using an arbitrary differentiable loss function. Decision forest regression is a non-parametric approach that perform a sequence of simple tests traversing a binary tree data structure until a decision is reached. In this study, the bagging resampling method and a single parameter training method is adopted. Neural network regression is a multiple interconnected node approach. In this study, the min-max normalizing approach is used. Poisson regression assumes the output follows a Poisson distribution and the logarithm of its expected value can be modeled by a linear combination of the independent variables. Details of each algorithm is beyond the scope of this paper and readers are recommended to read corresponding learning materials which can be easily found in textbooks and/or from the internet.

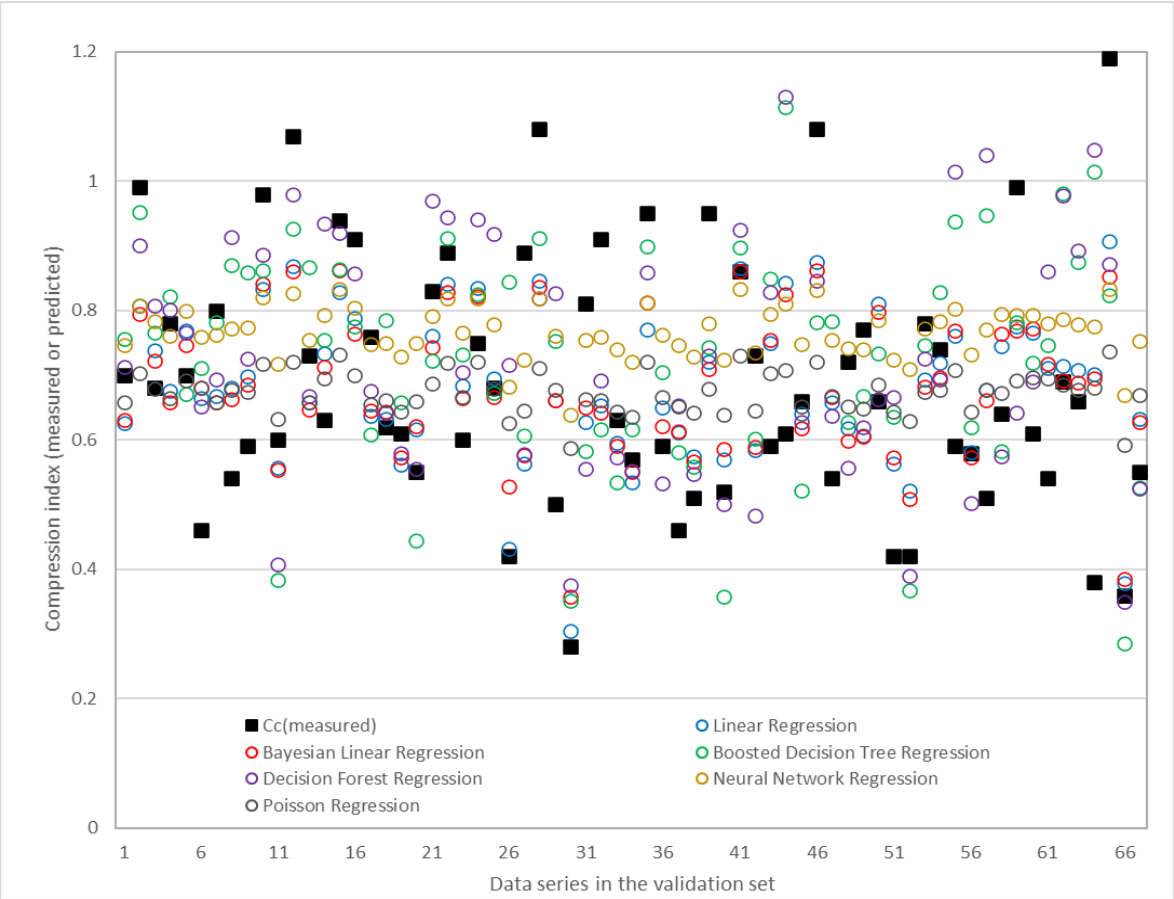

Figure 4: Prediction from different ML models (data from validation set of East Coast).

Figure 4 shows the comparison of prediction made using different ML algorithms for the East Coast validation dataset. There are 67 sets of date in this validation set (i.e., 30% of 223). The square symbol denotes the measured compression index and the circles having different colors represent prediction obtained from different algorithms. One can see clearly that the difference among the model predictions is noticeable.

Equation (1) is used to quantify the prediction error of the validation set where E denotes the mean absolute percentage error, y_i^m and y_i denote the measured C_c and corresponding prediction of entry i, respectively, N is the total number of data in the validation set.

$$E = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - y_i^m}{y_i^m}\right| \quad \textbf{(1)}$$

The larger the E, the more discrepancy between the measurement and the prediction is. Figure 5 summarizes the results for different algorithms at different location groups of the data. In generally, the linear and Bayesian regression algorithm give the smallest prediction discrepancy among other algorithms. They give an average absolute error of about 20%.

It is worth noting that fitting error is only one of the criteria to judge whether a prediction is good or not. Robustness of the fitting formula, for example, would also tell if the prediction is worth to be adopted. Overfitting is a common problem to observe. More details will be elaborated next

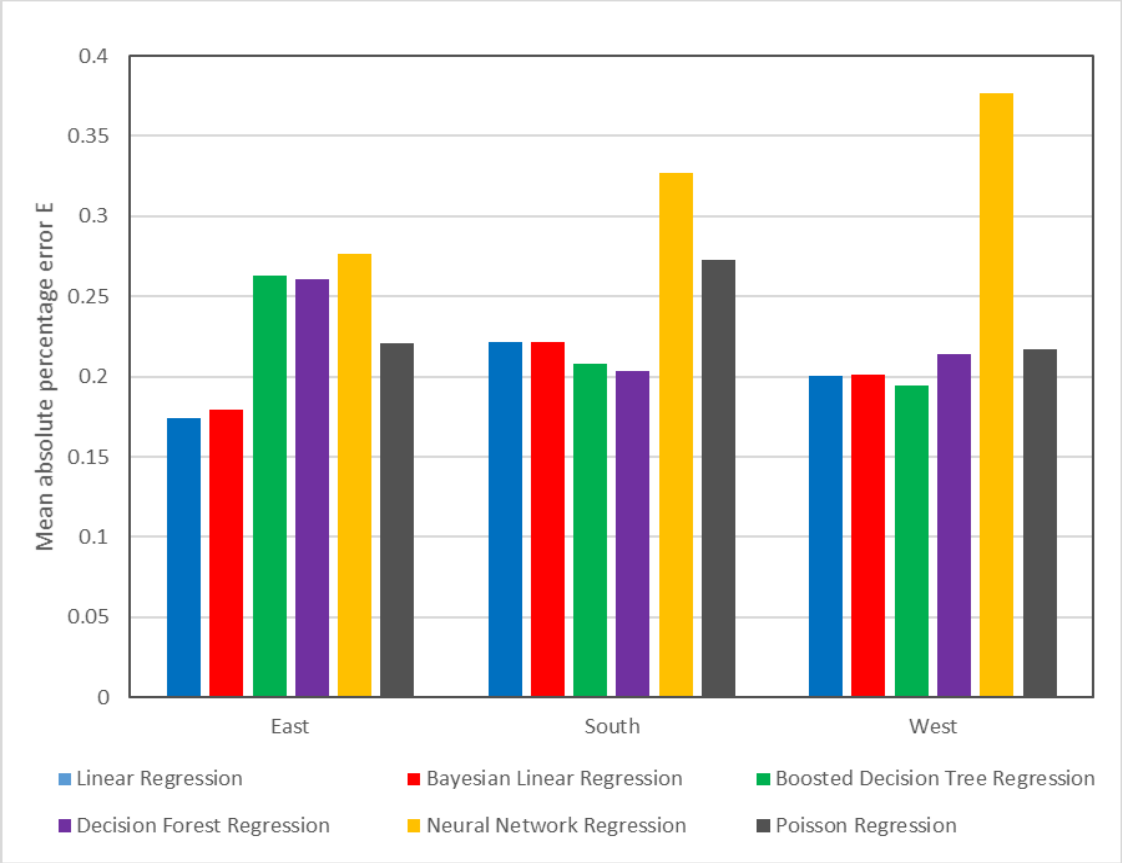

Figure 5: Mean absolute percentage error of each ML algorithm.

2.3 Parametric Bayesian Probabilistic-based Model Class Selection

Yan et al. (2009) presented a Bayesian probabilistic-based parametric approach to predict the soil's compression index. This approach has two major merits. First, by using the Bayesian probabilistic model class approach the most probable empirical prediction formula form is selected by achieving a balance between data fitting capability (i.e., likelihood) and sensitivity to modeling noise. This would mitigate the problem of overfitting. Second, an explicit form of empirical formula showing the optima fitting parameters is obtained which allows the formula to be examined or verified in the context of geomechanics. Though AI terminologies were not mentioned in their paper, the analysis was indeed falling into the family of ML according to the classification as presented in Section 1 of this paper. Based on their findings, the soil's compression index C_c can be expressed as

$$C_c = c_0 + c_1 e_0 + c_2 LL \quad \textbf{(2)}$$

where = c_0, c_1 and c_2 are fitting parameters calibrated from the dataset.

Proceedings DOI: 10.21467/proceedings.133
ISBN: 978-81-957605-1-0

Series: AIJR Proceedings
ISSN: 2582-3922

By using the parametric Bayesian probabilistic-based model class approach, the most probable empirical formula can be found. First, the problem of overfitting can be resolved using the data themselves. The most complex formula is not always the most probable one due to the problem of overfitting. Indeed, an over-complex prediction formula could bring in too much modelling noise. Second and more importantly, the formula offers geomechanics insights into the problem. In this compressibility prediction, for instance, the formula explicitly states that the soil compressibility would depend on a soil intrinsic index, the liquid limit, and another soil physical property, the initial void ratio. From a geomechanics perspective, the liquid limit quantifies the water content of a remolded soil for a specified undrained shear strength. Therefore, it is linked to the nature and the mineralogical composition of the soil and thus governs the compressibility. Besides, compressibility is affected by the soil structure which depends on its geological history. The initial void ratio is a suitable indicator to address this effect. A proper fundamental understanding of the rationale of the prediction formula helps to provide confidence when the formula is being used.

3 Observations and Lesson Learnt

Commercially available ML platforms offer a catalyst for the adoption or application of AI to our profession. The ML studio is adopted in this study. There are many built-in algorithms in the platform and simple drag-and-drop interface has been developed to facilitate the ML applications. On the one hand, different from the traditional structural coding/programing approach, this visual programming offers a mostly painless environment to develop the ML applications. On the other hand, users may overlook the default setting of each mathematical algorithm and make unintentional mistakes

This study has clearly demonstrated that various algorithms could give essentially “promising” or “non-promising” predictions. Adopting the ML algorithm like a black-box appears to give prediction without too much physical support. As shown in this study, noticeable difference in prediction could be obtained from various ML algorithms. Without going into details of the solution scheme one could not resolve the problem of overfitting and would be very difficult to judge which algorithm(s) outperforms the others. In the data science discipline, people believe that only a massive amount of good quality data could provide us more confidence on the results. Unfortunately, in geotechnical engineering massive amount of data is often impossible. Nevertheless, how to deal with sparse geotechnical data, particularly for modeling spatial variability of soil properties (e.g., Wang and Zhao 2017) and subsurface stratigraphy (e.g., Shi and Wang 2021), has been an active research area in recent years. The uncertainties associated with ML results of sparse data can be quantified using Bayesian methods and stochastic simulations (e.g., Wang et al. 2022). Promising development in this area is expected.

An independent study of Bayesian-based model class selection has demonstrated the importance of avoiding over-complex formula. Besides, physical significance of any prediction formula should be examined which could shed light on its validity.

To answer the statement “are we ready to use AI technologies for the prediction of soil properties”, the author believes that we are still at the starting point of the race. Collaboration between data scientists and geotechnical engineers would be required to bring forward this idea. Predicting soil properties should not be considered a purely mathematical issue but would need the knowledge of geomechanics.

4 Concluding Remarks

This paper presents the performance of predicting the compression index of marine clays found in different areas of South Korea using various built-in ML algorithms available in a commercial ML platform. The prime aim of this study is by using the above application as an example to examine if the use of ML for soil properties prediction can be readily adopted in the industry without much a concern. It is concluded that without a proper understanding of the basics of these algorithms any prediction

might be misleading and dangerous. Fundamental geomechanics still plays a key role in geotechnical engineering applications and any advanced tools or algorithms should be used with caution.

5 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ebid, A.M. 2021. 35 years of AI in geotechnical engineering: state of the art. Geotechnical and Geological Engineering, 39: 637-690.

Jaksa, M., Liu, Z. 2021. Editorial for special issue "application of artificial intelligence and machine learning in geotechnical engineering". Geoscience, 11(10): 399.

Jong, S.C., Ong, D.E.L., Oh, E. 2021. State-of-the art review of geotechnical-driven artificial intelligence techniques in underground soil-structure interaction. Tunnelling and Underground Space Technology, 113: 103946.

Shahin, M.A. 2016. State-of-the-art review of some artificial intelligence applications in pile foundations. Geoscience Frontiers, 7: 33-44.

Shi, C., Wang, Y. 2021. Development of subsurface geological cross-section from limited site-specific boreholes and prior geological knowledge using iterative convolution XGBoost. Journal of Geotechnical and Geoenvironmental Engineering, 147(9): 04021082.

Wang, Y., Hu, Y., Phoon, K.K. 2022. Non-parametric modelling and simulation of spatiotemporally varying geo-data. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(1): 77-97.

Wang, Y., Zhao, T. 2017. Statistical interpretation of soil property profiles from sparse data using Bayesian Compressive Sampling. Géotechnique, 67(6): 523-536.

Yan, W.M., Yuen, K.V., Yoon, G.L. 2009. Bayesian probabilistic approach for the correlations of compression index for marine clays. Journal of Geotechnical and Geoenvironmental Engineering ASCE, 135(12): 1932-1940.

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.

Zhang, P., Yin, Z., Jin, Y. 2021a. Machine learning-based modelling of soil properties for geotechnical design: review, tool development and comparison. Archives of Computational Methods in Engineering, 29(2): 1229-1245.

Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., Ding, X. 2021b. Application of deep learning algorithms in geotechnical engineering: a short critical review. Artificial Intelligence Review, 54: 5633-5673.

Proceedings DOI: 10.21467/proceedings.133
ISBN: 978-81-957605-1-0

Series: AIJR Proceedings
ISSN: 2582-3922