Machine Learning-Based Cone Penetration Test (CPT) Data Interpretation

Boyu Wang*, Kelvin Tse, Clifford Phung

WSP (Asia) Limited

*Corresponding author
doi: https://doi.org/10.21467/proceedings.159.4

ABSTRACT

Ground investigations (GI) are essential prior to the design of construction projects. Among the different GI tasks, classifying soils into groups with similar properties is a fundamental geotechnical engineering process. Currently, experienced geotechnical engineers manually conduct soil classification using empirical tables based on laboratory or in-situ tests, which is labor-intensive and time-consuming. This study presents a machine learning (ML)-based approach to inferring soil types based on Cone Penetration Test (CPT) data. To identify an appropriate classification model, three classic algorithms, including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF), were built and validated on data collected from a reclamation project (The Project). Four important attributes from CPTs, including tip resistance q_c, sleeve friction f_s, pore-pressure u_2, and depth d, were used as input features, and six soil types in The Project were applied as labels. The different models were compared based on their prediction performance and required learning time. The best results for both targets were obtained using the RF classifier, achieving over 90% accuracy.

Keywords: Cone Penetration Test, Machine Learning, Soil Classification

1 Introduction

Classifying soils into groups with similar properties is a fundamental engineering task during the preliminary stages of a construction project. However, during the feasibility stage, project-specific soil properties are not yet determined without investing in project-specific ground investigations (GI), which can be high-risk and costly. As investigating subsoil conditions using a combination of field and laboratory tests is inevitably associated with high costs, this process is often designed to be minimal. In recent years, the Cone Penetration Test (CPT) has become a popular and cost-effective tool for investigating subsoil conditions (Rauter *et al*., 2021; Robertson, 2009, 2016).

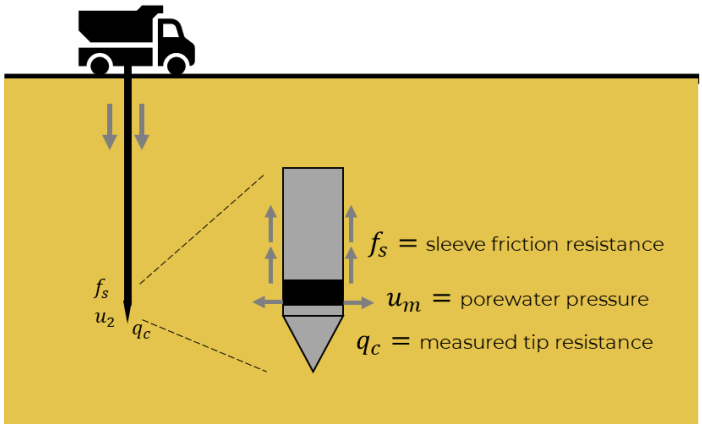

Figure 1: *Overview of the CPT*

Proceedings DOI: 10.21467/proceedings.159; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-965621-6-8

In a CPT, a cone with a specific diameter is pushed vertically into the ground at a constant rate, as shown in Figure 1. Based on the different measured data, such as tip resistance (q_c) and sleeve friction (f_s), various soil behavior charts have been developed to identify soil strata and behavior types. Additionally, various empirical correlations have been published to interpret CPT data quickly and easily (including parameter determination). CPTs are mainly performed to determine subsoil conditions, such as soil type and stratification, and to estimate shear strength parameters (e.g., effective friction angle ϕ', cohesion c, etc.). However, determining soil strata from CPT data requires manual processing using soil behavior type charts provided by Robertson (2009, 2016), as shown in Figure 2, which is a time-consuming and error-prone task.

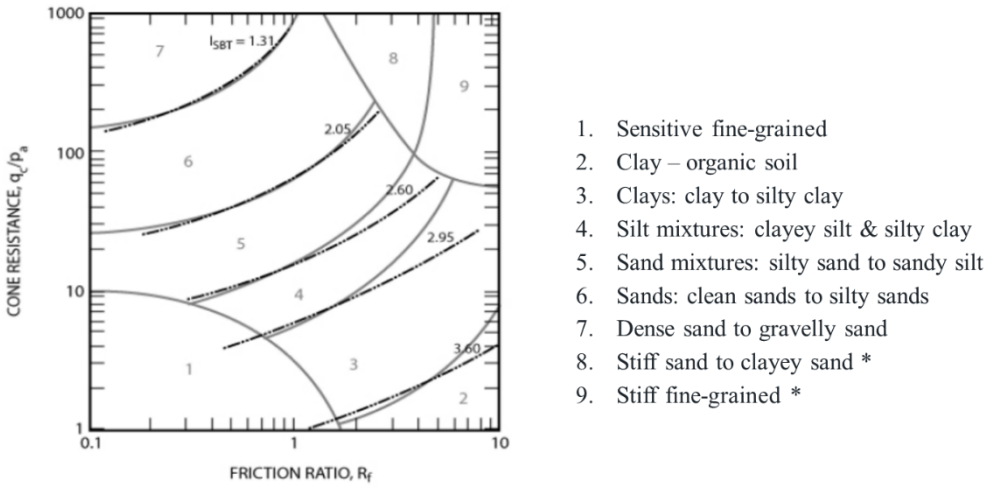

Figure 2: *Soil behavior type charts provided by Robertson, (2009, 2016).*

There have been some research efforts to improve the accuracy and efficiency of CPT data interpretation. Shi & Wang (2022a); Wang *et al.*, (2019) and Zhao *et al.*, (2020) focused on predicting soil stratification with limited data. For instance, Wang *et al.*, (2013) developed a Bayesian approach to explicitly model the uncertainty of the chart-based interpretation. Building on this work, Wang *et al.* (2019) extended the Bayesian approach from one dimension (1D) to two dimensions (2D) and proposed a subsurface soil stratification and zonation method in a 2D vertical cross section. Zhao *et al.*, (2020) combined Markov Chain Monte Carlo with Bayesian Compressive Sensing (BCS) to achieve fast non-parametric simulation of 2D multi-layer CPT data. Shi & Wang (2022a, 2022b) leveraged prior geological knowledge to construct three-dimensional (3D) subsurface geological models.

Meanwhile, with the rise of machine learning (ML), big data-driven approaches have also emerged as a promising trend. Over the last decade, the application of ML algorithms in civil engineering has gained more interest as it could play a key role in reducing costs and time needed for data analysis and prediction (Padarian *et al.*, 2020; Rauter *et al.*, 2021). Machine learning techniques have been employed in previous studies to classify soils based on CPT data and to accurately estimate soil and design parameters (Kurup & Griffin, 2006; Reale *et al.*, 2018; Zhang *et al.*, 2021). In contrast to previous studies, the present investigation is based on an extensive dataset of 700 CPT tests that were conducted under similar soil and environmental conditions in Hong Kong. Three classic ML models were evaluated in terms of accuracy and learning time. The results demonstrate that the ML approaches can achieve over 90% accuracy in predicting soil type and significantly reduce the inference time for engineers.

2 Dataset

The dataset consisting of 700 CPTs used for this study was collected from The Project. In total eight types of soil were considered including Marine Deposit (MD), Dumped Mud (DM), Dumped Sand (DS), Alluvium Sand (ALL-s), Intermediate Alluvium Clay (ALL-c(int)), Unweathered Alluvium Clay (ALL-c(unw)), Paleosol Alluvium Clay (ALL-c(pal)) and Grade V soil. As the soil properties of MD and DM are quite similar, this study did not differentiate the two classes purposely. The same strategy also applied to the soil types of DS and ALL-s. The soil strata information was obtained based on adjacent boreholes.

In this study, four crucial parameters obtained from CPTs - namely, tip resistance q_c, sleeve friction f_s, pore pressure u_2, and depth d - have been utilized as input features for predictive modeling. To ensure data quality, any rows with missing or null entries were removed since the sample size was sufficiently large. The dataset was then split into two subsets for training and validation using the scikit-learn tool "train_test_split" (Pedregosa *et al.*, 2011), which randomly divides the samples. In this study, 100 CPTs were used for training and the remaining 600 CPTs were used for validation. A sample of the CPT data is presented in Figure 3.

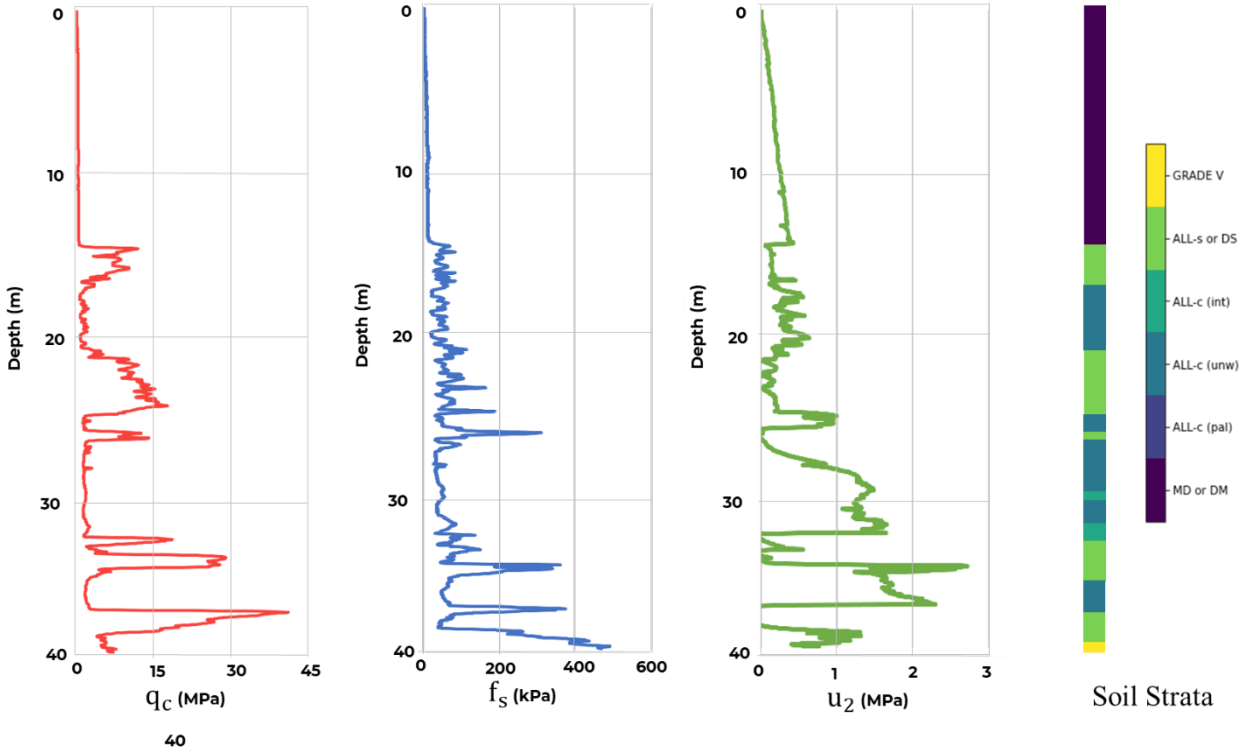

Figure 3: *An example of measured data from a CPT (tip resistance q_c, sleeve friction f_s as well as pore-pressure u_2) and the corresponding soil strata.*

To ensure that each input feature contributes equally to the training and prediction process, the "StandardScaler" module was used to scale the features. This scaling process rescaled the features to lie between -1 and +1 while keeping the median at the same level, thereby avoiding any bias. Additionally, the uneven distribution of data between the classes can pose challenges for many ML algorithms in making accurate predictions. In this study, a data resampling strategy was employed to address this issue. Specifically, an oversampling algorithm was used to fill the underrepresented classes with synthetic data generated from the available data, while overrepresented classes were removed to balance

Proceedings DOI: 10.21467/proceedings.159
ISBN: 978-81-965621-6-8

Series: AIJR Proceedings
ISSN: 2582-3922

the dataset. Throughout this process, statistical parameters such as the mean and median of the data were kept at the same level, ensuring that the resampled data was representative of the original dataset.

3 Machine Learning Models

To efficiently interpret the collected CPT data, learning-based method is utilized in this study. ML has emerged as a popular tool for analyzing large datasets in various scientific fields. Unlike traditional computing algorithms that compute results based on input and predefined solutions, ML models learn from input features and their corresponding outputs (targets) to find solutions, regardless of the specific algorithm used (Alpaydin, 2020). This study evaluates three distinct ML algorithms - the Support Vector Machine (SVM) (Schölkopf, 1998), Artificial Neural Network (ANN) (Hornik *et al.*, 1989), and Random Forest (RF) (Breiman, 2001) - each with their own unique function principles, which are briefly described in the following subsections.

3.1 Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning algorithm that can be used for classification, regression, and outlier detection (Schölkopf, 1998). For classification and regression tasks, the SVM algorithm aims to find separating hyperplanes in a high or infinite-dimensional space with the largest margin. A larger margin corresponds to a lower generalization error of the model. An example of a linear SVM is depicted in Figure 4(a), where the samples on the boundaries are referred as support vectors. In this study, radial kernel function was used for the employed SVM model to achieve higher prediction accuracy.

3.2 Artificial Neural Network

The Artificial Neural Network (ANN) is inspired by the function principle of the human brain (as shown in Figure 4(b)) and is composed of three distinct types of layers: the input layer, where input features are inputted into the model; the hidden layer(s), where information from the input layer is combined with weights; and the output layer, where results are computed (Hornik *et al.*, 1989). In this study, the backpropagation algorithm is used to train the neural network iteratively. During training, the model's output is compared with the real targets of the training set to calculate the error and update the weights in the hidden layers. This process continues until a minimum error is reached or the incremental improvement between iterations becomes negligible.

The ANN model was constructed using the MLPClassifier module of the scikit-learn library (Pedregosa *et al.*, 2011). The best combination of hyperparameters was determined using grid search techniques, where a range of values for each parameter was defined. To ensure reasonable training time, the number of hidden layers was restricted to a maximum of three, with a maximum of 10 neurons per layer.

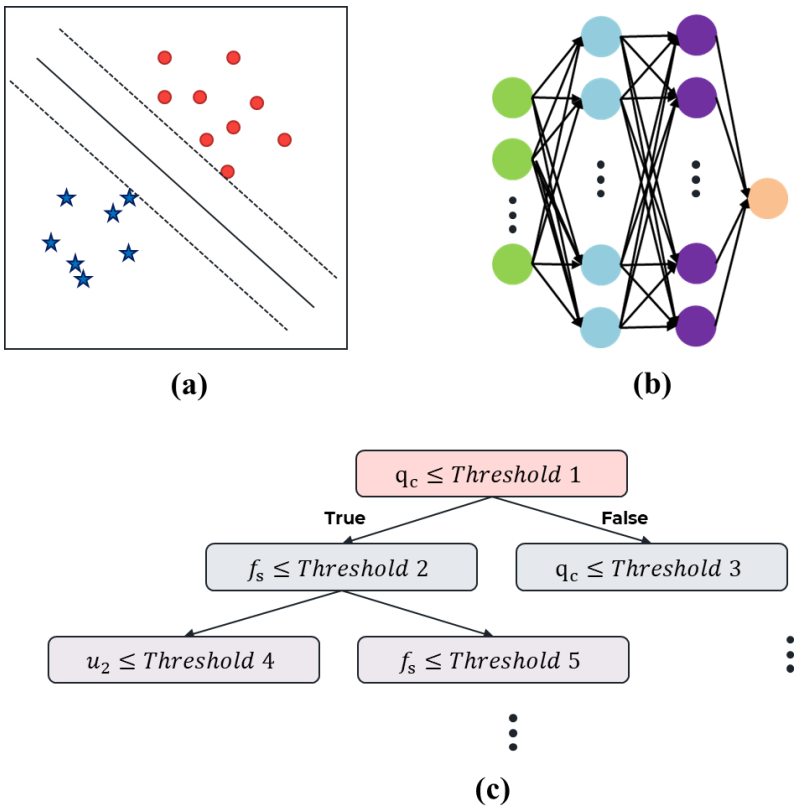

Figure 4: *ML algorithm visualization. (a) SVM. (b) ANN. (c) RF.*

3.3 Random Forest

The Random Forest (RF) is an ensemble of decision trees, where each decision tree is a non-parametric supervised learning method that summarizes decision rules from a series of data with features and labels (Breiman, 2001). Decision trees are capable of solving classification and regression problems by presenting the rules in a comprehensible tree structure that allows for the identification of each input feature's contribution to the model. As depicted in Figure 4(c), decision trees consist of nodes responsible for different decision-making steps.

In this study, the RF classifier was implemented using the ensemble learning module of scikit-learn. Similar to the ANN models, the best set of hyperparameters was determined through cross-validation. Learning and validation curves were used to analyze the RF models and visualize bias and variance, which indicate susceptibility to overfitting or underfitting. To obtain a robust model, bias and variance should be kept low. Variance is represented by the difference between training and validation accuracy. High variance results in a model that is not able to generalize well, leading to much higher training accuracy than validation accuracy. High bias, on the other hand, indicates that the data are too complex for the model. One of the primary hyperparameters that governed bias and variance in an RF model was the maximum size of each tree ("max_depth") in the forest.

3.4 Post Processing

To enhance the consistency of prediction results from machine learning (ML) models, a smoothing technique has been developed. For each predicted label at a specific position, the k neighboring labels are examined. If the most frequent labels above and below the query position are the same, and the

Proceedings DOI: 10.21467/proceedings.159
ISBN: 978-81-965621-6-8

Series: AIJR Proceedings
ISSN: 2582-3922

predicted label is different from the frequent label, the frequent label will be assigned to the query position. This approach helps to eliminate noise and improve the robustness of the prediction results.

4 Training, Validation & Testing

In this section, the performance of the three built ML models on the prepared dataset are reported. The experiment settings are described in detail.

4.1 Experiment Procedure

Following data pre-processing, the construction of a ML model can be broken down into three distinct stages. The first stage is the training phase, where the ML algorithm learns from the training data through an iterative process that continues until a desired minimum error or maximum accuracy, or a predetermined maximum number of iterations, is achieved. The second stage is the validation phase, where the model's generalization properties, such as overfitting and underfitting, are analyzed. Overfitting occurs when the model fits the training data better than the validation data, while underfitting indicates that the model is too simplistic for the given data. To mitigate any issues related to data distribution, validation is commonly performed using cross-validation (CV) techniques such as k-fold cross validation, which involves splitting the training data into k subsets for training and testing. Through CV, learning and validation curves can be plotted to assess model performance. The third stage is the testing phase, where the model, with optimized hyperparameters, is evaluated on unseen data from the test dataset.

4.2 Experiment Configurations

All models were built on a DELL Precision 5570 (CPU: 12th Gen Intel (R) Core (TM) i7-12800H (20CPUs) ~2.4 GHz, RAM: 32 GB, GPU: NVIDIA RTX A1000 Laptop GPU) using the Anaconda python environment. The ML algorithms used are part of the open-source-software library of scikit-learn.

4.3 Evaluation Metrics

To assess the robustness and efficiency of the classification models, a confusion matrix was used to visualize their performance. The overall classification accuracy (OA) was then utilized to quantitatively evaluate the performance of the CPT data-based soil classification. In addition to OA, three other classification metrics, namely precision, recall, and F1-score, were applied to comprehensively evaluate performance. The definitions of these evaluation metrics are presented below.

A confusion matrix (CM) is a summary of a classifier's prediction results that provides a visual representation of its performance. Table 1 illustrates the contents of a confusion matrix, where true positive (TP) indicates a positive prediction that correctly matches the actual positive state. False negative (FN) represents a negative prediction that should have been positive, while false positive (FP) indicates a positive prediction that should have been negative. True negative (TN) represents a correct negative prediction. The confusion matrix not only provides a visual representation of the model's performance but also facilitates the identification of confusion between similar classes for error analysis.

Table 1: *Schema of a confusion matrix for the evaluation of results*

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The overall segmentation accuracy (OA) is the rate of accurately predicted labels out of the total number of data samples (Eq. (1)). However, OA alone is insufficient for evaluating a classifier's performance. Therefore, three additional classification metrics, including precision, recall, and F1-score, were used to quantitatively assess the performance of the classification models. Precision measures the accuracy of positive predictions (Eq. (2)), while recall refers to the ratio of positive data samples that are correctly predicted by the classifier (Eq. (3)). The F1-score represents the harmonic mean of precision and recall (Eq. (4)), integrating the precision and recall values into a single metric.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \textbf{Eq.(1)}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \textbf{Eq.(2)}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \textbf{Eq.(3)}$$

$$\text{F1} - \text{score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad \textbf{Eq.(4)}$$

5 Results

The RF algorithm achieved the best performance across all classification targets and feature sets, exhibiting high accuracy (92%) and minimal training time as shown in Table 2. In addition to the RF model, the SVM model was found to be unsuitable for this type of task and data, as it took significantly longer to train than the other two models and produced the worst overall accuracy of 81%. The three-layer ANN model was easier to train and performed better in determining soil classes from CPT data, achieving an accuracy of 85%.

Table 2: *Model performance reports*

Algorithms	Training Time (s)	Accuracy	Recall	Precision	F1-score
SVM	565	0.81	0.81	0.78	0.78
ANN	89	0.85	0.85	0.86	0.85
RF	71	0.92	0.92	0.92	0.92

Figure 5 displays a selection of comparison samples between the soil classification obtained from the RF model and the actual soil classification obtained from adjacent boreholes in The Project. It can be observed that the soil types in most positions were predicted correctly using the RF model with excellent consistency.

Proceedings DOI: 10.21467/proceedings.159
ISBN: 978-81-965621-6-8

Series: AIJR Proceedings
ISSN: 2582-3922

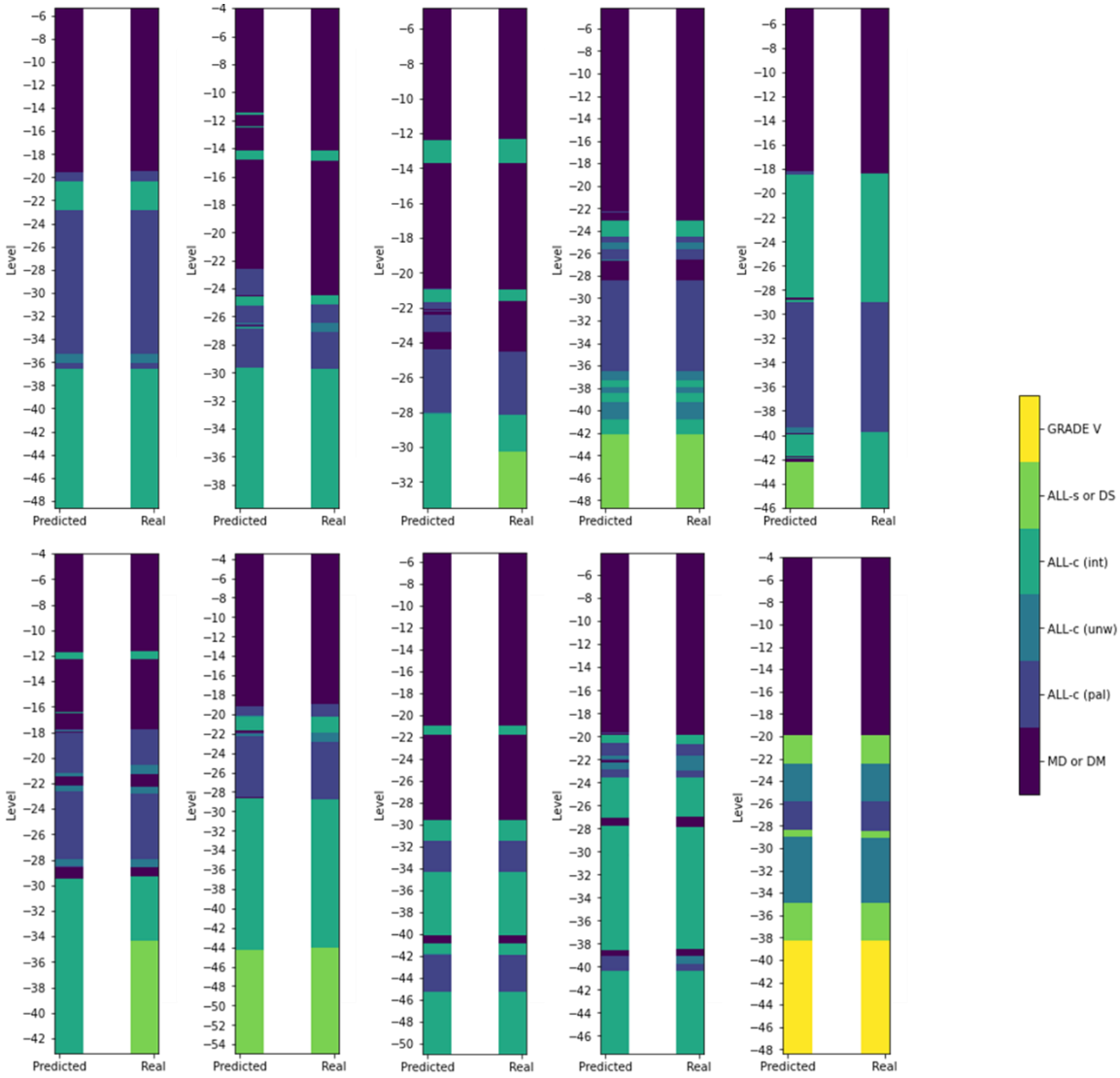

Figure 5: *Comparison between the soil classification obtained from RF and the actual soil classification from adjacent boreholes in The Project*

The confusion matrix for the RF model is presented in Figure 6. The results show that more than 94% of soil samples of types "ALL-c (unw)", "ALL-s or DS", and "MD or DM" were correctly predicted. The model exhibited less confidence in recognizing soil type "ALL-C (pal)", which was often classified as "ALL-s or DS", "ALL-c (unw)", or "ALL-c (int)". Similarly, "ALL-c (int)" could sometimes be misclassified as either of the other two "ALL-c" soil types.

Despite the possibility of misclassification, the ML methods offer a clear advantage over traditional workflows that rely on engineer judgement in CPT data interpretation, as illustrated in Figure 7. Using a machine learning approach, the analysis of 600 CPTs can be completed within half an hour, while the manual process can only handle less than 50 CPTs within the same time frame. ML can be used as a preprocessing tool to speed up the overall workflow without sacrificing accuracy.

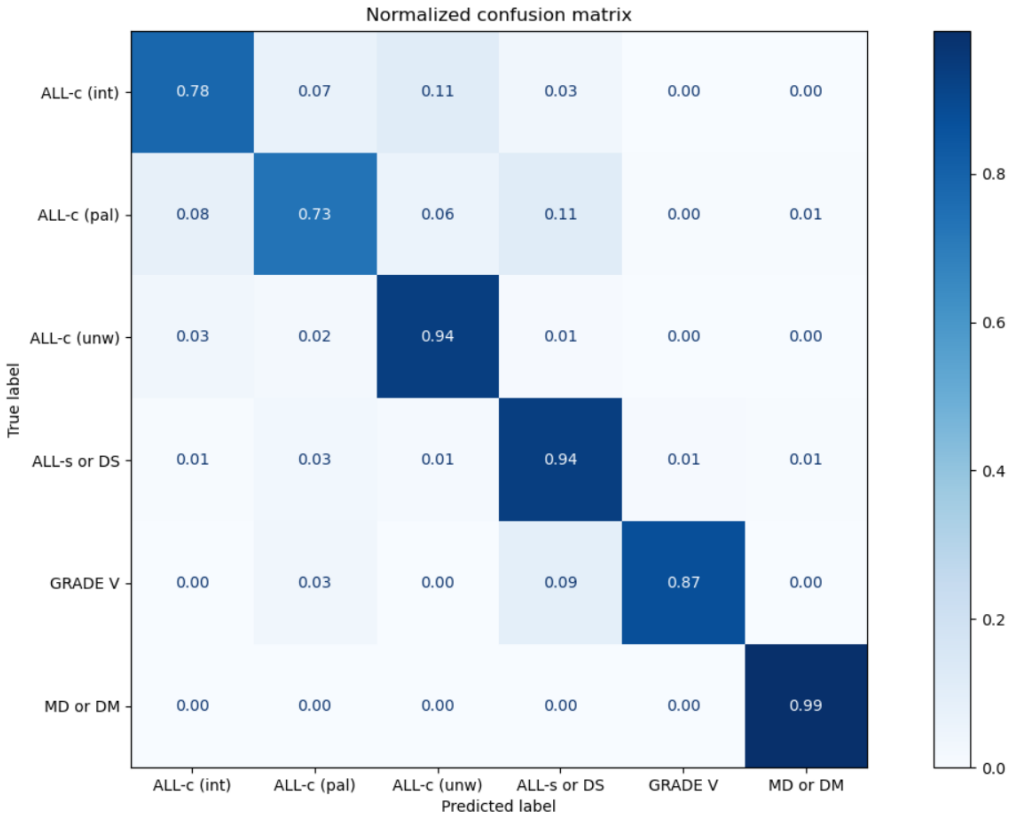

Figure 6: *Confusion matrix visualization for the testing results of the RF model.*

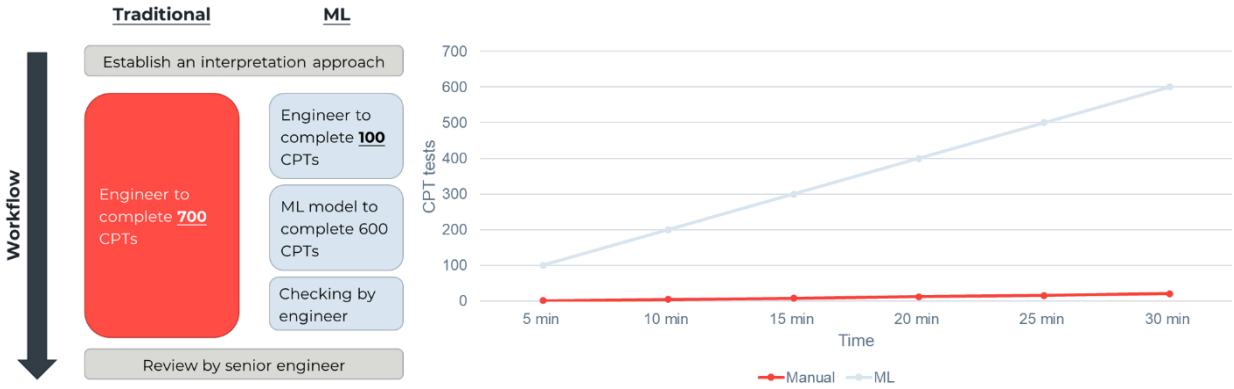

Figure 7: *Efficiency comparison between the ML approach and the traditional interpretation process.*

6 Conclusion & Discussion

This paper has demonstrated that ML algorithms can accurately classify soils based on measured CPT data. The best results in terms of prediction accuracy and learning time were achieved using a RF classifier. However, it should be noted that more advanced neural networks, such as deep neural networks (DNN), may lead to even better results. These investigations are part of ongoing research.

Compared to traditional manual processes, ML methods can significantly expedite the CPT data interpretation process. It may serve as a useful tool within geotechnical engineering software packages to obtain fast and reliable soil classifications without relying on third-party solutions.

Since ML model training requires massive amounts of data, the authors propose that the relevant Hong Kong Government department(s) could take the lead to coordinate with local engineering firms and

Proceedings DOI: 10.21467/proceedings.159
ISBN: 978-81-965621-6-8

Series: AIJR Proceedings
ISSN: 2582-3922

developers / clients to create a centralized GI inventory. Prior to submitting the data to the Government, the submitting party would be responsible for checking the data format and cleansing the data. Additionally, a cloud-based data management system could be utilized to facilitate better data sharing and access. With a rich and comprehensive dataset, ML models can produce more accurate results. Moreover, larger and more reliable deep learning models can be used to perform sophisticated GI data analysis tasks, such as constructing three-dimensional subsurface geological models.

7 Publisher's Note

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

How to Cite

Wang *et al.* (2023). Machine Learning-Based Cone Penetration Test (CPT) Data Interpretation. *AIJR Proceedings*, 32-41. https://doi.org/10.21467/proceedings.159.4

References

Alpaydin, E. (2020). *Introduction to machine learning*. MIT Press. https://books.google.com/books/about/Introduction_to_Machine_Learning_fourth.html?id=uZnSDwAAQBAJ

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Kurup, P. U., & Griffin, E. P. (2006). Prediction of Soil Composition from CPT Data Using General Regression Neural Network. *Journal of Computing in Civil Engineering*, *20*(4), 281–289. https://doi.org/10.1061/(ASCE)0887-3801(2006)20:4(281)

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *SOIL*, *6*(1), 35–52. https://doi.org/10.5194/SOIL-6-35-2020

Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Rauter, S., Tschuchnigg, F., Jaksa, M., & Liu, Z. (2021). CPT Data Interpretation Employing Different Machine Learning Techniques. *Geosciences 2021*, *11*(7), 265. https://doi.org/10.3390/GEOSCIENCES11070265

Reale, C., Gavin, K., Librić, L., & Jurić-Kaćunić, D. (2018). Automatic classification of fine-grained soils using CPT measurements and Artificial Neural Networks. *Advanced Engineering Informatics*, *36*, 207–215. https://doi.org/10.1016/J.AEI.2018.04.003

Robertson, P. K. (2009). Interpretation of cone penetration tests — a unified approach. *Canadian Geotechnical Journal*, *46*(11), 1337–1355. https://doi.org/10.1139/T09-065

Robertson, P. K. (2016). Cone penetration test (CPT)-based soil behaviour type (SBT) classification system — An update. *Canadian Geotechnical Journal*, *53*(12), 1910–1927. https://doi.org/10.1139/CGJ-2016-0044

Schölkopf, B. (1998). SVMs - A practical consequence of learning theory. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–21. https://doi.org/10.1109/5254.708428

Shi, C., & Wang, Y. (2022a). Data-driven construction of Three-dimensional subsurface geological models from limited Site-specific boreholes and prior geological knowledge for underground digital twin. *Tunnelling and Underground Space Technology*, *126*, 104493. https://doi.org/10.1016/J.TUST.2022.104493

Shi, C., & Wang, Y. (2022b). Machine learning of three-dimensional subsurface geological model for a reclamation site in Hong Kong. *Bulletin of Engineering Geology and the Environment*, *81*(12), 1–18. https://doi.org/10.1007/S10064-022-03009-Y

Wang, Y., Hu, Y., & Zhao, T. (2019). Cone penetration test (CPT)-based subsurface soil classification and zonation in two-dimensional vertical cross section using Bayesian compressive sampling. *Canadian Geotechnical Journal*, *57*(7), 947–958. https://doi.org/10.1139/CGJ-2019-0131

Wang, Y., Huang, K., & Cao, Z. (2013). Probabilistic identification of underground soil stratification using cone penetration tests. *Canadian Geotechnical Journal*, *50*(7), 766–776. https://doi.org/10.1139/CGJ-2013-0004

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, *12*(1), 469–477. https://doi.org/10.1016/J.GSF.2020.03.007

Zhao, T., Xu, L., & Wang, Y. (2020). Fast non-parametric simulation of 2D multi-layer cone penetration test (CPT) data without pre-stratification using Markov Chain Monte Carlo simulation. *Engineering Geology*, *273*, 105670. https://doi.org/10.1016/J.ENGGEO.2020.105670