

Soil Classification and Feature Importance of EPBM Data using Random Forests

Dayu Apoji, S.M.ASCE,¹ Yuji Fujita,² and Kenichi Soga, Ph.D., F.ASCE³

¹Department of Civil and Environmental Engineering, University of California Berkeley, Berkeley, CA 94720-1234; E-mail: dayu.apoji@berkeley.edu

²Enzan Koubou Co.,Ltd., 237-3 Yamazato-cho, Kamigyo-ku, Kyoto, 602-8268, Japan; E-mail: fujita@enzan-k.com

³Department of Civil and Environmental Engineering, University of California Berkeley, Berkeley, CA 94720-1234; E-mail: soga@berkeley.edu

ABSTRACT

This paper presents an implementation of Random Forest (RF), a supervised learning algorithm, to classify the encountered geologic conditions using continuous earth pressure balance machine (EPBM) operation data. This study was performed on a data set from State Route 99 (SR99) tunnel construction in Seattle, WA. Hyperparameter tuning was performed to investigate the effects of RF hyperparameters on the classification performance as well as to determine the best hyperparameter configuration. The role of the features in the classification model was investigated by evaluating the feature importance measures. This study demonstrates that, with straightforward hyperparameter tuning, RF could deliver good classification performance and could infer the geologic transition through the classification probabilities. This study indicates that although several EPBM features had relatively larger “weights” for the classification, it was the interactions among the features that contain the geologic information.

INTRODUCTION

Geologic uncertainty and variability along tunnel alignments strongly affect the tunneling performance and risk. To safely achieve the expected performance, tunnel boring machine (TBM) operators must be aware of the changing geologic conditions along the tunnel alignment (Sutcliffe 1996). A geologic map can be helpful. However, it can only provide a generalized guidance since it is interpreted from limited boreholes at discrete locations. Therefore, during tunneling, the operators have to continuously infer the geologic conditions based on signals produced by TBM sensor measurements (Garcia et al. 2021).

Developing more systematic methods to infer the encountered geologic conditions based on TBM data remains an open research question. Various data-driven methods have been implemented to tackle this problem, such as using probabilistic models (e.g., Sousa and Einstein 2012) and machine learning (e.g., Erharter et al. 2020; Zhang et al. 2019; Zhao et al. 2019). Despite the recent advances, more attention has been given to complex and less interpretable

models such as neural networks. Furthermore, less focus has been paid to understand the features (measured variables used as predictors) of the model (Sheil et al. 2020).

This study evaluated Random Forests (RF), a supervised learning algorithm, as a tool to systematically infer (i.e., classify) the encountered geologic conditions along a tunnel alignment. This study also investigated the role of each feature in the classification model using feature importance measures. This study was performed on a data set from State Route 99 (SR99) tunnel construction in Seattle, WA. The tunneling was performed using a 17.5 meters diameter earth pressure balance shield machine (EPBM).

DATA AND METHODS

Geologic Conditions. Geologic conditions along SR99 tunnel were obtained from the geotechnical baseline project reports (WSDOT 2010a; b). For the tunneling purpose, geologic deposits along the alignment were classified into several engineering soil units (ESU) based on their physical characteristics. The geologic conditions were dominated by over-consolidated glacial and non-glacial pre-Vashon geologic units, with ESU of Till Deposits (TD), Cohesionless Sand and Gravel (CSG), Cohesionless Silt and Fine Sand (CSF), Cohesive Clay and Silt (CCS), and Till-Like Diamict (TLD). Note that TD might have more complex engineering behavior since it was mixtures of gravel, sand, silt, and clay. TLD was generally cohesionless but might have layers and lenses of tills.

Supervised learning algorithms require labels as the “ground truth”. In this study, ESU at borehole locations along the tunnel alignment were chosen as the labels. The boreholes were envisaged to represent geologic conditions for a 50-ft (15.2 meters) radius from the drilled location. Based on the boreholes, the EPBM face might encounter multiple ESU in most of the alignment. Therefore, five labels were created to represent mixtures of the ESU at each borehole location (Table 1).

Table 1. Soil label for supervised classification, based on mixtures of ESU.

Geologic conditions based on ESU	Soil labels
Fully cohesive soils (CCS)	CCS
Predominantly cohesive soils (CCS) with layers of cohesionless soils (CSG, CSF)	CCS_CSGCSF
Predominantly cohesionless soils (CSG, CSF) with layers of cohesive soils (CCS)	CSGCSF_CCS
Predominantly cohesionless soils (CSG, CSF) with layers of till/till-like deposits (TD, TLD).	CSGCSF_TLTD
Mix of cohesionless soils (CSG, CSF), cohesive soils (CCS), and till/till-like deposits (TD, TLD)	TLTD_CSGCSF_CCS

EPBM Data. The EPBM was equipped with numerous sensors to record various measurements during tunneling operation. This recorded almost 6000 features in both spatial and temporal range at every few seconds. To limit the complexity of the data, the study was focused on the continuous data of the primary EPBM system, i.e., main features in the tunneling processes. Features that might contain biased information toward spatial location of the shield were not included (i.e., pitch and yaw angles of the shields, earth pressures in the chamber). To further condense the number of features and enable more interpretability of the analysis results, records from same sensor types were summed together. This produced a total of 36 features to be used in the analysis.

In this study, the observation points were represented in spatial series as EPBM chainage locations per ring advance. This means an observation point represents measurement records along one ring advance. Depending on the data characteristics, an observation point can be obtained by taking the final observation value of one ring advance (e.g., features related to volume, length), or taking the average value along one ring advance (e.g., features related to pressure, forces, speed). A few examples of the data with the soil labels are shown in Figure 1.

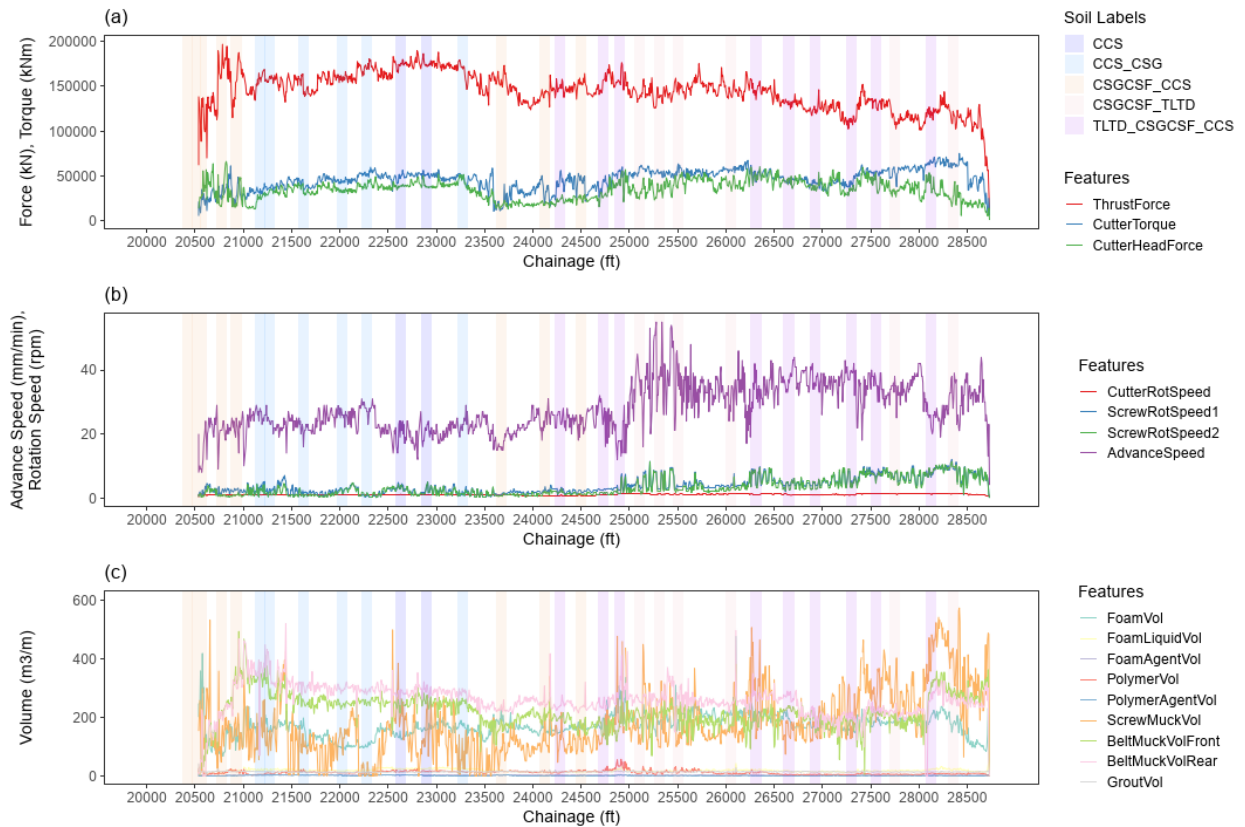


Figure 1. Examples of the data along the tunnel alignment: (a) features related to force and torque, (b) features related advance speed and rotation speed, (c) features related to volume. Background colors represent the soil labels.

Random Forests. RF is a supervised learning algorithm that produce the prediction by aggregating a large number of decision trees as the base learners (Breiman 2001). RF becomes popular due to its strong predictive performance in high-dimensional data, versatility in various feature types and scale, robustness to missing data, outliers, and noises, as well as its ability to measure importance of the features. Furthermore, RF is a nonparametric method. Thus, it is suitable for EPBM data analysis since the distribution of each sensor measurement may vary and cannot be predetermined.

Briefly, RF algorithm is performed as follows. (i) Generate different bootstrapped training data sets by drawing samples with replacement from the original training data set. Samples outside the bootstrapped data are called ‘out-of-bag’ (OOB) data. (ii) Construct decision trees using the bootstrap samples. A single decision tree is constructed using binary recursive partitioning into distinct subsets. The objective of partitioning is to minimize dissimilarity in terminal nodes. In classification problems, it is typically measured by the Gini impurity index. (iii) Obtaining the final result by aggregating all the constructed decision tree results. This study was performed in R programming language using fast implementation of RF in C++ (Wright and Ziegler 2017).

Data Preparation. The data preparation consisted of the following steps. (i) As previously discussed, EPBM chainage locations per ring advance were used as the observation points representation. The observations of non-advancing tunneling stages were not included. (ii) Features from same sensor types were summed together. Features that might contain biased information toward spatial location of the shield were not included. (iii) Observations with erroneous records and missing values were removed. This reduces the amount of observation but avoids further complexities that might arise from imputing missing data with assumptions. (iv) Observations located adjacent to available boreholes were labeled.

Model Setup. Classification models were developed based on two data splitting schemes, namely, static random model and dynamic sequential classification model. The static random model was developed as a benchmark model with randomness in training-testing data. In this model, data for training and testing were randomly selected from the labeled data with ratio of 80:20. To deal with imbalanced class distribution, the training data was up sampled. Beside the testing data (in labeled data set), the trained model was also used to classify the unlabeled data.

The dynamic sequential model was developed as a scheme that can be implemented in real tunneling cases. In this model, the classifications were performed at each observation point (testing data) based on the past observations (training data), sequentially. In other words, to classify the encountered soils at a particular chainage, the model was re-trained using observations from the previous chainages. The training data was also up sampled at every re-training stage.

Hyperparameter Tuning. Hyperparameter tuning was performed on the static random model to investigate the effects of RF hyperparameters on the classification performance as well as to determine the best hyperparameter configuration. The analysis was performed on the static random classification model. The overall value of correct classification in out-of-bag (OOB) data was used as the accuracy metric.

Repeated K-fold cross-validation (CV) analysis with 5 times repeat and 10-folds was performed in various range of RF hyperparameters: (i) splitting rule, (ii) possible number of features that are randomly selected to split at each node (mtry), (iii) minimum node size, and (iv) number of trees (ntrees). The splitting rule defines how a single decision tree to be constructed. Both mtry and minimum node size define the complexities of the constructed trees. The ntrees defines how the forests to be generated.

Feature Importance. Feature importance analysis was performed on the static random model to investigate the role of each feature in soil classification. However, there are various methods to measure the feature importance and each method may produce different scores. Therefore, to capture the general tendency of the feature importance, three measurement methods were used in this study, namely (i) impurity (Breiman 2001), (ii) permutation (Nicodemus et al. 2010), and (iii) conditional permutation (Strobl et al. 2008).

RESULTS AND DISCUSSION

Effects of Hyperparameters. Figure 2(a) and Figure 2(b) present the effects of ntrees and mtry (at minimum node size of 10) on the OOB classification accuracy based on Gini (Breiman 2001) and Extratrees (Geurts et al. 2006) split rules, respectively. These figures show that ntrees was important to produce better classification accuracy. Both figures show that the accuracy increased to about 0.98 when ntrees was increased. However, the accuracy tends to be stable after ntrees reached a threshold value. To ensure both high accuracy and reasonable computation cost, ntrees of 500 were selected for the models.

Figure 2(a) shows that Gini split rule produced the lowest accuracy when mtry was equal to the number of features (mtry = p = 36). Generally, the accuracy increased at lower numbers of mtry. In contrast, Figure 2(b) shows that Extratrees split rule produced the lowest accuracy at mtry of 1. It appears that Extratrees split rule produced more stable accuracy in the variation of mtry, except the single one. Therefore, Extratrees split rule was selected for the models.

Figure 2(c) and Figure 2(d) present the effects of mtry and minimum node size (at ntrees of 500) on the OOB classification accuracy. Both split rules produced similar highest accuracy values. Figure 2(c) shows that Gini split rule achieved the peak accuracy at about mtry of 6, which is equal to $\sqrt{p}$. This result is in agreement with previous studies (Probst et al. 2019). At higher numbers of mtry, the accuracy dropped. In contrast, Figure 2(d) shows that Extratrees split rule produced relatively more stable accuracy at mtry of $\sqrt{p}$ to p . Both split rules obtained their lowest accuracy at minimum node size of 20. Beside this value, variation of the accuracy

was less significant. Therefore, mtry of $\sqrt{p}$ and minimum node size of 10 was selected for the models.

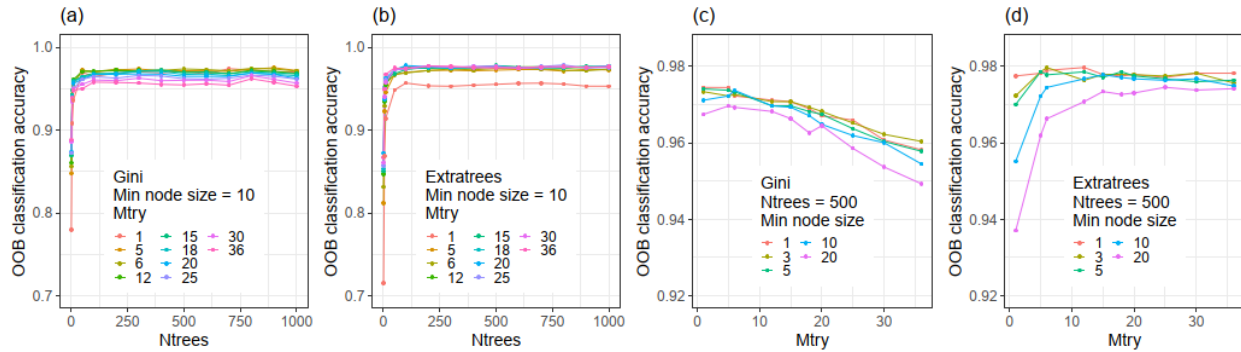


Figure 2. Hyperparameter sensitivity: ntrees vs. accuracy using Gini split rule (a) and Extratrees split rule (b), mtry vs. accuracy using Gini split rule (c) and Extratrees split rule (d).

Soil Classification. Figure 3(a) presents geologic map and boreholes along the tunnel alignment (WSDOT 2010a; b). The boreholes were colored based on the soil labels. The grey color denotes the unlabeled data, where no boreholes were available. Figure 3(b) presents the results of the RF static random classification. The first row of Figure 3(b) shows the classification results. The black bar denotes classification errors. The classification results were determined based on the highest classification probability produced at each data point, as shown in the second row of Figure 3(b). The static random model produced good classification accuracy of 0.933 (93.3%). However, this accuracy applies to the labeled data only. The classification results of the unlabeled data could only be qualitatively compared to the geologic map (which is also an interpretation by geologists).

The second row on Figure 3(b) shows that the classification probability appears to reveal the geologic transition along the tunnel horizon. For example, both the geologic map and the classification probability show transition from predominantly CCS to predominantly CSG between borehole TB108 and TB223. In further details, the model could also capture localized geologic conditions. For example, the model captured increasing probability of mixed soils just before borehole TB312, where the geologic map indicates the beginning of the CCS layer on top of the tunnel horizon, creating mixed geologic conditions.

Figure 3(c) presents the results of the RF dynamic sequential classification model. It can be observed that more errors were produced by the dynamic model, with the accuracy decreased to 0.913 (91.3%). This lower accuracy was expected since the training data for the dynamic model was not randomized, hence contained more bias. Furthermore, the model was only trained with very limited data and labels in the early parts of the tunnel alignment. This caused weak performance in classification analysis. However, the performance seems improved (i.e., similar to the static model) after the model received adequate training data and labels.

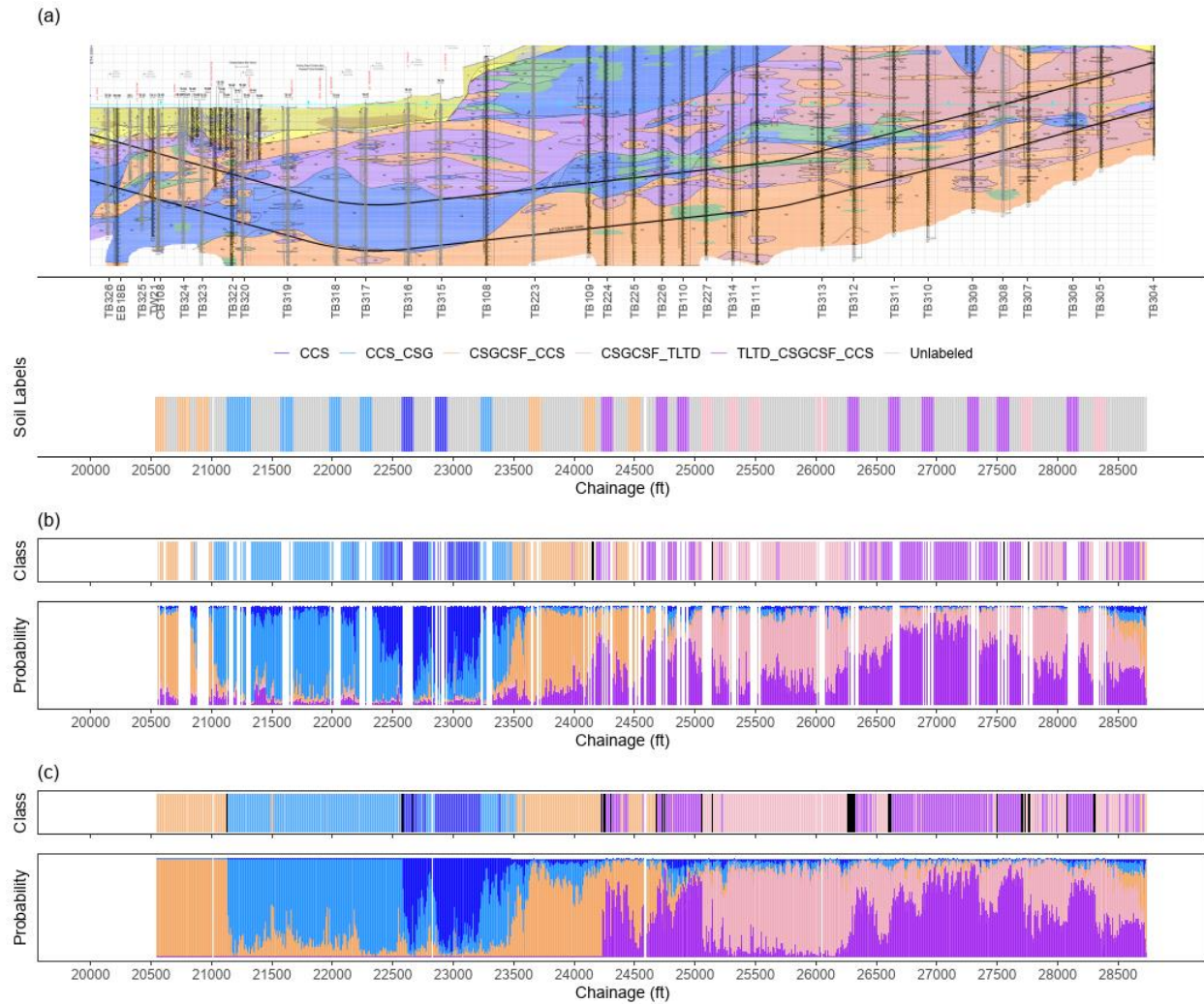


Figure 3. Classification results (a) Geologic map and boreholes (WSDOT 2010a; b), (b) static random classification model results, and (c) dynamic sequential classification model results. Black bars denote classification errors of labeled data.

Feature Importance. Figure 4 shows the results of feature importance analysis. The features are shown in color according to the EPBM feature groups. Features with higher rank imply that they had stronger “weights” in the RF soil classification model. The relative importance was standardized to maximum score of 1.0. Note that the relative comparison of the feature ranks is more important than the importance scores.

All measurement methods show that features related to cutter (i.e., cutter torque and cutter rotation speed), thrust (i.e., thrust force and advance speed), and ground conditioners (i.e., foam and polymer) were consistently at high importance ranks. This was expected as cutter and thrust directly interact with the encountered soils (Sousa and Einstein 2012). In contrast, another feature from the thrust group, i.e., thrust stroke, was consistently at the lowest importance rank. Thrust stroke is typically pushed to its maximum stroke length or to the specified length based on

the shield trajectory and therefore the values should not depend on the encountered soils. This can be a sign that the feature importance measurements returned sensible results.

The association between ground conditioners and the encountered soils seems obvious. Ground conditioners are typically injected into the encountered soils at the cutterhead to achieve favorable soil flow characteristics. However, applying actual ground conditioners during tunneling is not straightforward as effective characteristics of the conditioners depends on the injected soils properties (Thewes et al. 2012). Therefore, during tunneling, EPBM operators have to continuously adjust the conditioners to achieve favorable soil flow characteristics (JSCE 2016).

Figure 5 can be evidence. This figure shows soil classification accuracy when RF models were trained using a single feature only. The figure shows that a single feature could only produce classification accuracy of about 0.5 or below, even if the feature were at high rank in the importance measures (e.g., ground conditioners). The low accuracy of every feature indicates that information contained in only a single feature may not be adequate for soil classification and it was the interactions among the EPBM features that contain the “true” information about the geologic conditions.

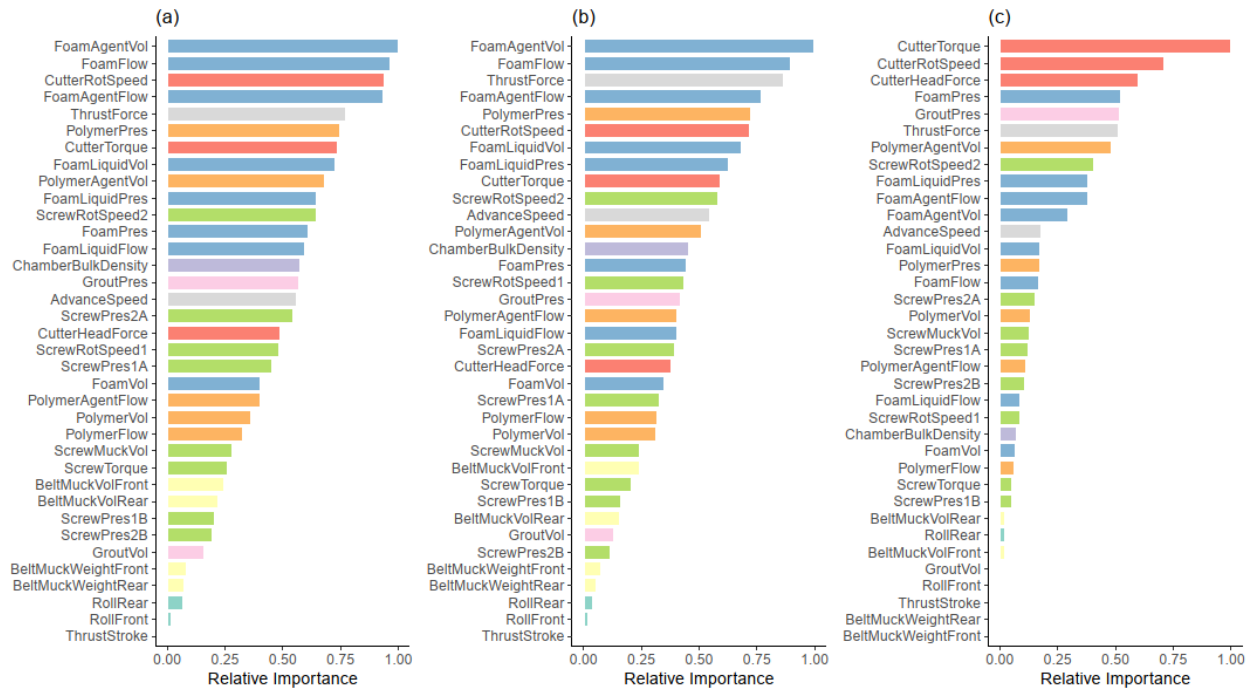


Figure 4. Feature importance measured based on (a) impurity, (b) permutation, and (c) conditional permutation.

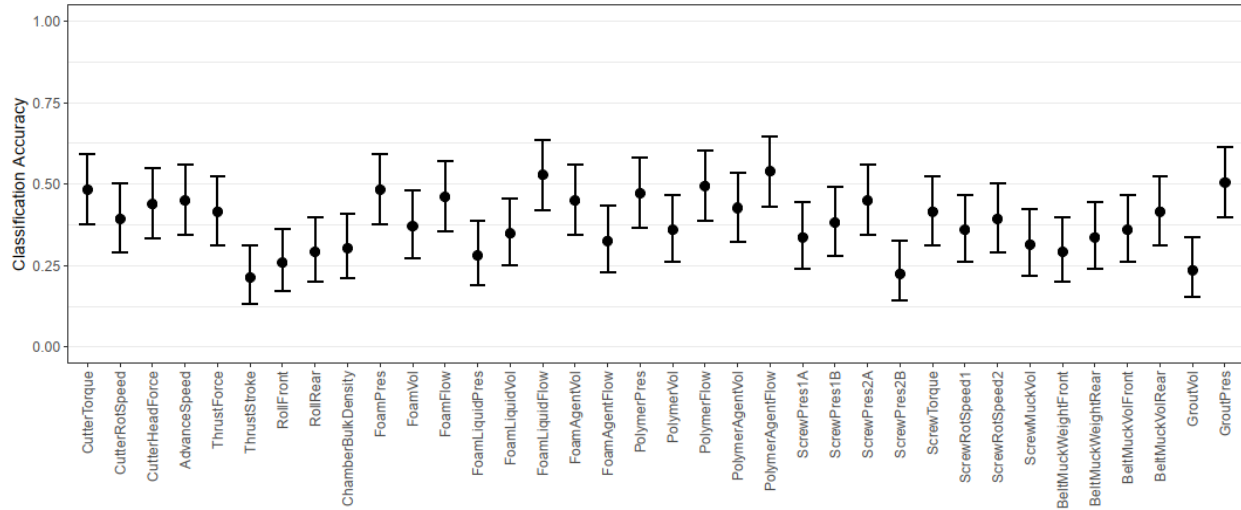


Figure 5. Soil classification accuracy when RF models were trained using a single feature only.

CONCLUSIONS

This study has demonstrated that RF can potentially be employed as a tool to systematically infer the encountered geologic conditions along tunnel alignments. In this case, RF could deliver stable and good classification performance with straightforward hyperparameter tuning. In adequate data and the corresponding labels, the RF models could also infer the geologic transition through the classification probabilities. This provides evidence that the EPBM data contained information about the geologic conditions.

The feature importance analysis has shown that features related to the cutter, thrust, and ground conditioners contained relatively larger “weights” in the classification model. However, information contained in a single feature, even if the feature was at high importance rank, may not be adequate for soil classification. This indicates that it was the interactions among the EPBM features that contain the information about the geologic conditions. Therefore, it is important to capture these interactions to infer the encountered soils during tunneling.

ACKNOWLEDGEMENTS

The first author gratefully acknowledged scholarship from Indonesia Endowment Fund for Education (LPDP) and research support from Department of Civil and Environmental Engineering, University of California at Berkeley. The authors thank Hiroshi Kogi, Hidetoshi Maeda, Takehiko Nakaya (Shimizu Corp.), Zhangwei Ning, Ph.D. (Sixense Inc.), and Renjie Wu (UC Berkeley) for valuable discussion during this research. The data used in this study was obtained from Washington State Department of Transportation (WSDOT).

REFERENCES

- Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5–32.
- Erharter, G. H., Marcher, T., and Reinhold, C. (2020). "Artificial Neural Network Based Online Rockmass Behavior Classification of TBM Data." *Information Technology in Geo-Engineering*, Springer Series in Geomechanics and Geoengineering, A. G. Correia, J. Tinoco, P. Cortez, and L. Lamas, eds., Springer International Publishing, Cham, 178–188.
- Garcia, G. R., Michau, G., Einstein, H. H., and Fink, O. (2021). "Decision support system for an intelligent operator of utility tunnel boring machines." *Automation in Construction*, 131, 103880.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). "Extremely randomized trees." *Machine Learning*, 63(1), 3–42.
- JSCE. (2016). *Standard Specifications for Tunneling: Shield Tunnels*. Japan Society of Civil Engineers.
- Nicodemus, K. K., Malley, J. D., Strobl, C., and Ziegler, A. (2010). "The behaviour of random forest permutation-based variable importance measures under predictor correlation." *BMC Bioinformatics*, 11(1), 110.
- Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). "Hyperparameters and tuning strategies for random forest." *WIREs Data Mining and Knowledge Discovery*, 9(3), e1301.
- Sheil, B. B., Suryasentana, S. K., Mooney, M. A., and Zhu, H. (2020). "Machine Learning to Inform Tunnelling Operations: Recent Advances and Future Trends." *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction*, 1–18.
- Sousa, R. L., and Einstein, H. H. (2012). "Risk analysis during tunnel construction using Bayesian Networks: Porto Metro case study." *Tunnelling and Underground Space Technology*, 27(1), 86–100.
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). "Conditional variable importance for random forests." *BMC Bioinformatics*, 9(1), 307.
- Sutcliffe, H. (1996). "Tunnel Boring Machines." *Tunnel Engineering Handbook*, J. O. Bickel, T. R. Kuesel, and E. H. King, eds., Springer US, Boston, MA, 203–219.
- Thewes, M., Budach, C., and Bezuijen, A. (2012). "Foam conditioning in EPB tunnelling." *Geotechnical Aspects of Underground Construction in Soft Ground*, CRC Press Hoboken.
- Wright, M. N., and Ziegler, A. (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software*, 77(1).
- WSDOT. (2010a). *Interim Report CT-6: Geologic Characterization*. SR 99 Bored Tunnel Alternative Design-Build Project, Prepared by Shannon & Wilson, Inc.
- WSDOT. (2010b). *Revised Geotechnical Baseline Report*. SR 99 Bored Tunnel Alternative Design-Build Project, Prepared by Parsons Brinckerhoff, Shannon & Wilson, Inc.
- Zhang, Q., Liu, Z., and Tan, J. (2019). "Prediction of geological conditions for a tunnel boring machine using big operational data." *Automation in Construction*, 100, 73–83.
- Zhao, J., Shi, M., Hu, G., Song, X., Zhang, C., Tao, D., and Wu, W. (2019). "A Data-Driven Framework for Tunnel Geological-Type Prediction Based on TBM Operating Data." *IEEE Access*, 7, 66703–66713.