Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-1272

CPTu-based identification of glauconite sand using Random Forest modeling

Mertcan Geyin, Asitha Senanayake, Federico Pisanò
Norwegian Geotechnical Institute, USA, mertcan.geyin@ngi.no

Zack Westgate
Department of Civil and Environmental Engineering, University of Massachusetts Amherst, USA

ABSTRACT: Offshore wind development in the U.S. has to date focused on the North and Mid-Atlantic Outer Continental Shelf (OCS). These regions contain glauconite-rich sediments, which exhibit unique shearing behavior, potentially impacting or preventing pile foundation installation. Detecting glauconite-rich sands through site investigation data, such as piezocone penetrometer testing (CPTu), is challenging due to the limitations of current classification frameworks. The Soil Behavior Index (I_c) derived from CPTu data has shown promise for characterizing glauconite sands; however, existing correlations often fail to identify interlayered glauconite. To address this, an interbeddedness index was recently introduced, quantifying soil type transitions over a given depth interval. Spatial frequency analysis of I_c applied to over 500 offshore CPTu and borehole datasets across 300 km² enabled the creation of interbeddedness maps, which identified critical I_c frequency thresholds, providing a systematic framework for detecting glauconite-rich layers. Building on this framework, this paper introduces a machine learning-based classification strategy using Random Forests trained on CPTu-derived features and borehole-confirmed glauconite presence. A Positive-Negative-Unlabeled (PNU) learning approach was adopted to handle incomplete labeling, and bootstrapped modeling was used to assess robustness across varying depth window lengths. SHAP analysis was employed to interpret feature contributions, revealing that normalized cone tip resistance, friction ratio, and pore pressure response are dominant predictors, while interbeddedness appears to play a more domain-specific role. The model's predictive performance was evaluated across training and test datasets, showing high AUC scores with acceptable generalization scatter. This paper applies the interbeddedness-based classification approach to new CPTu data from both authigenic and allogenic sites with glauconite presence, generating plausible soil layering scenarios and exploring its broader applicability. Findings highlight the identification of glauconite-rich layers might be possible with CPTu-based classification approaches for unconventional soil conditions and expand on unique behaviors of glauconite-rich sediments useful for the development of future models.

KEYWORDS: soil characterization, glauconite sand, CPTu, machine learning, Random Forest.

1 INTRODUCTION

Glauconite sands present notable challenges in offshore geotechnical engineering, particularly in shallow marine environments along the U.S. Atlantic Continental Shelf (DeGroot et al., 2023). Similar deposits also occur in the Belgian North Sea (e.g., Joustra and De Gijt, 1982; Perikleous et al., 2003) and eastern Canada (Philibert et al., 2024). While undisturbed glauconitic sands often demonstrate considerable strength, their friable nature makes them susceptible to brittle failure under moderate mechanical disturbance (Van Alboom et al., 2012; Westgate et al., 2023a, 2023b). This fragility can lead to abrupt particle disintegration during dynamic loading events such as pile driving or wave-induced cyclic stresses. Such breakdown can cause a rapid transition in soil behavior from coarse- to fine-grained, accompanied by increases in plasticity. Cementation in glauconite sediments further complicates their response, producing variable soil resistance to pile driving and uncertainty in long-term axial and lateral pile behavior. These factors underscore the importance of early glauconitic layer identification during offshore site characterization.

Despite their significance, detecting glauconitic materials in offshore settings remains a complex task. Sparse data availability has hindered a comprehensive understanding of their geotechnical behavior, which has to date mainly been based on experience with Belgian onshore construction (Van Alboom et al., 2012; De Nijs et al., 2015) and recent investigations associated with the Piling in Glauconitic Sand (PIGS) Joint Industry Project (Westgate et al., 2024, 2025; Pisanò et al., 2025). While geological surveys along the U.S. Atlantic coast (e.g., Trumbull, 1972; Poag, 1978) have mapped glauconite occurrences to some extent, these efforts primarily address formation processes rather than engineering implications. A deeper insight into their mechanical behavior is essential to assess their impact on offshore foundation performance.

To bridge this gap, there is a pressing need for a predictive framework that can estimate glauconitic presence prior to extensive laboratory testing and the development of detailed geological maps. This study proposes a pragmatic approach using conventional site investigation tools, specifically, piezocone penetration testing (CPTu) in conjunction with glauconite counts from adjacent borehole samples.

Characteristic CPTu signatures, such as elevated normalized tip resistance (Q_t), friction ratio (F_r), and pore pressure ratio (B_q) can be used to predict the presence of glauconitic soils. While the pore pressure response is still being explored as a basis for classification of glauconite deposition history, e.g. in situ, or *authigenic*, deposits versus reworked, or *allogenic*, deposits, the friction ratio is now well understood to be a key parameter in early glauconite identification, which can exceed 10% (Joustra and De Gijt, 1982; Van Alboom et al., 2012; Rogiers et al., 2017; Long et al., 2019). Additionally, the vertical variability of these parameters can be used to detect glauconitic traces. To quantify this variability, we apply the "interbeddedness" metric proposed by Geyin et al. (2023), defined as the frequency of inferred soil behavior type (SBT) transitions per unit depth. This simple proxy has shown promise in glauconite identification.

In this paper, the work of Geyin et al. (2023) is extended using machine learning methods, specifically Random Forest models, which were developed to predict glauconite presence using CPTu signatures from datasets collected from two distinct depositional environments. Beyond their predictive utility, the modeling process itself offers valuable insights into how glauconitic materials influence CPTu responses.

The following sections describe salient methodological aspects and obtained results. Given the intended geotechnical readership, familiarity with machine learning terminology is not taken for granted; therefore, a glossary of key terms is provided in the final appendix.

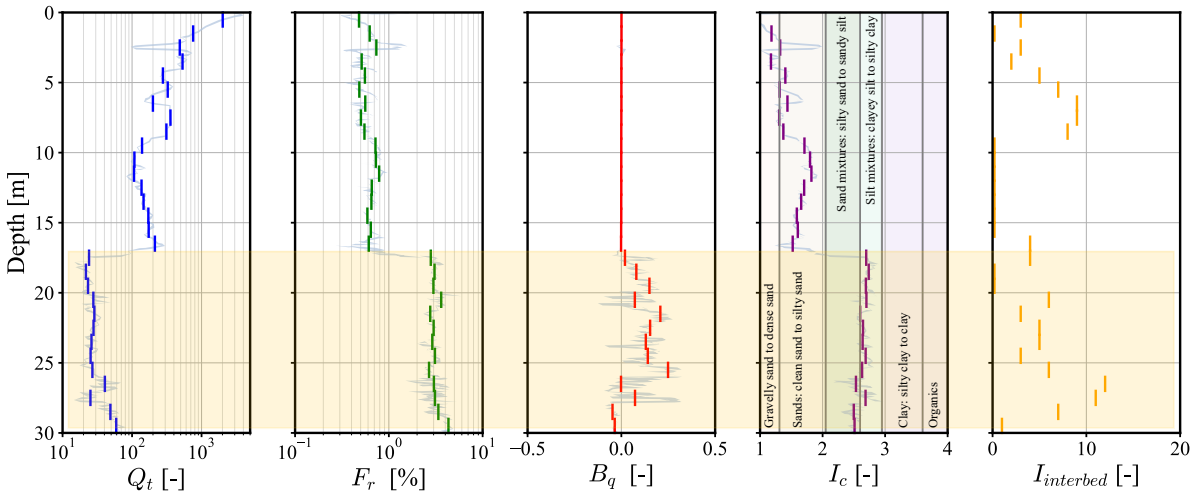

Figure 1. Representative CPTu profiles with median feature values and glauconite-rich interval highlighted.

2 DATA AND METHODOLOGY

2.1 *Dataset Integration and Feature Engineering*

This study integrates two geotechnical datasets representing distinct depositional environments, specifically authigenic and allogenic conditions. Each dataset contains CPTu records and borehole-based glauconite content data, based on the visual grain counting method. The CPTu records were concatenated to form a unified modeling corpus, enabling broader generalization across depositional settings. Borehole identifiers were used to align CPTu data with lab-derived glauconite content measurements, allowing depth-resolved feature extraction and labeling.

To observe the effects of vertical variability and stratigraphic transitions, CPTu data were segmented into overlapping depth windows ranging from 0.5 to 4 m. Within each window, six features were computed to characterize soil behavior: interbeddedness index (number of SBT transitions per meter, $I_{interbed}$), median and standard deviation of I_c, normalized pore pressure ($B_q = \Delta u / (q_t - \sigma_{vo})$ – where q_t is the cone tip resistance corrected for pore pressure; Δu is the excess pore pressure measured during penetration; and σ_{vo} is the estimated total vertical stress), and median values of Q_t and F_r (Figure 1). These features were selected based on prior sensitivity analyses and in light of their geotechnical relevance. For the current modeling effort, five features ($I_{interbed}$, median I_c, B_q, Q_t, and F_r) were retained, excluding the standard deviation of I_c due to its limited contribution in earlier trials (Geyin et al., 2023).

2.2 *Labeling Strategy and PNU Framework*

Glauconite presence was labeled using a Positive-Negative-Unlabeled (PNU) framework. Borehole samples were considered positive if glauconite content exceeded a threshold of 30% (i.e., more than 30 glauconite particles counted out of 100 total representative particles from a given deposition sample) within the depth window, negative if below threshold, and unlabeled if no data were available. The glauconite threshold of 30%, is used here only as an illustrative example, but coincides with the notional geological classification boundary between glauconitic sand (<30% glauconite content) and glauconite sand (≥30% glauconite content) (Westgate et al., 2023). This approach accommodates the reality of incomplete lab testing, allowing the model to learn from confirmed positives and negatives while ignoring ambiguous samples. To further reduce label noise, we excluded CPTu data whose associated boreholes lacked any glauconite presence whatsoever. Such filtering steps ensure that the model is trained only on CPTu data with meaningful glauconite content.

2.3 *Model Selection, Performance Metric and Justification*

Two classifiers were evaluated: Logistic Regression (LR) and Random Forest (RF). LR offers simplicity and interpretability, modeling linear relationships between features and the log-odds of glauconitic presence. However, LR assumes feature independence and linearity, which may not hold in stratified subsurface environments. RF, on the other hand, is a non-parametric ensemble method that constructs multiple decision trees and aggregates their predictions. It is robust to multicollinearity, captures nonlinear interactions, and provides feature importance metrics via SHAP (SHapley Additive exPlanations, Lundberg and Lee, 2017).

Given the complexity of subsurface behavior and the need for flexible modeling, RF was ultimately selected as the primary classifier. Empirically tuned hyperparameters were applied, including a maximum tree depth of 3 to limit model complexity, a minimum of 10 samples required at each leaf node to ensure generalization, and an ensemble size of 100 decision trees to stabilize predictions.

To assess the model performance, Receiver Operating Characteristics (ROC) analyses were utilized. Area Under the Curve (AUC) was used to compare the performances. ROC curves illustrate how the model's sensitivity and false alarm rate change as the decision threshold varies. Specifically, they show how often the model correctly identifies glauconite (true positives) versus how often it incorrectly predicts its presence where it does not exist (false positives), across different probability cutoffs. While no single metric can perfectly summarize model performance, AUC is a widely accepted standard because it offers a statistically grounded and unbiased measure (see Fawcett, 2006). In this setting, AUC reflects how well the model assigns higher probabilities to locations where glauconite is present compared to those where it is absent. In summary, a model with a higher AUC is more effective at distinguishing sites with glauconite from those without.

Increasing tree depth or reducing leaf size led to overfitting, evidenced by inflated training AUCs and degraded test performance. The chosen configuration was found to be ample for capturing meaningful patterns without sacrificing generalizability.

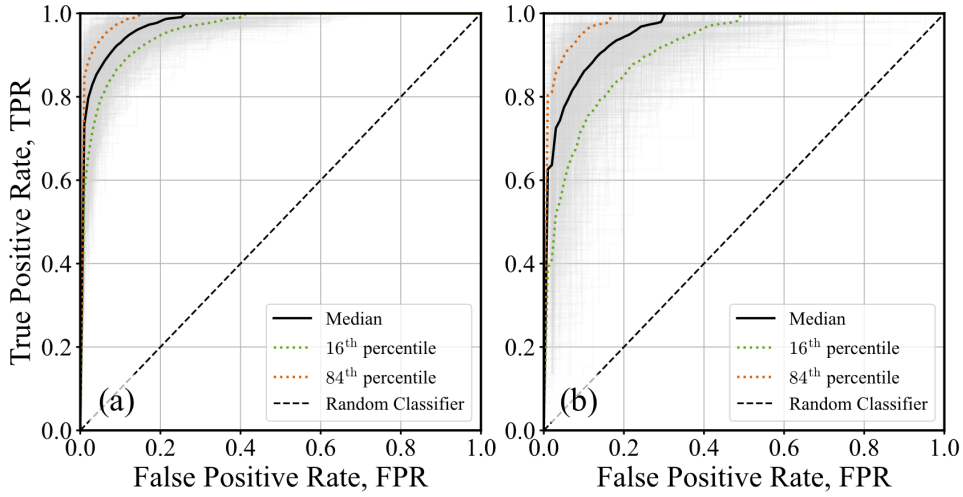

Figure 2. Bootstrapped AUCs from training (a) and testing (b) datasets.

2.4 *Bootstrapping and Cross-Validation*

To assess model robustness, 1000 bootstrapped iterations were performed per depth window. In each iteration, CPTu data were sampled with replacement, features and labels were computed, and the model was trained and evaluated. Bootstrapping enables estimation of model variance and confidence intervals, following the principles established by Diaconis and Efron (1983). This approach is particularly valuable in geotechnical applications, where data heterogeneity and limited sample sizes can obscure model reliability.

Each iteration included a 70/30 train-test split, stratified by label. Additionally, 5-fold stratified cross-validation was performed on the training set to evaluate internal consistency. This approach ensures that each fold maintains the same proportion of samples with glauconite and those without, reducing variance in AUC estimates. Cross-validation also guards against overfitting by exposing the model to multiple training-validation splits, providing a more reliable estimate of generalization performance.

2.5 *Handling Class Imbalance*

Within the dataset analyzed in this study, glauconite-bearing samples are relatively rare. This rarity creates a class imbalance that can bias the model's learning process. To address this, the study used a technique called SMOTEENN, which combines two strategies:

- SMOTE creates synthetic examples of samples with glauconite by blending existing ones, helping the model see more of the minority class (Chawla et al., 2002).
- ENN (Edited Nearest Neighbors) then removes samples that are confusing or lie too close to the boundary between samples with and without glauconite, helping the model draw clearer distinctions (Wilson, 1972).

Other methods like SMOTE alone and SMOTETomek (Batista et al., 2003) (which removes borderline samples after SMOTE) were also tested, but they did not perform as well in maintaining a clear separation between classes. Additionally, the model was trained with balanced class weights, which means it penalizes mistakes on samples with glauconite more heavily to ensure they are not ignored.

Together, these techniques help the model stay sensitive to glauconite presence, even though most of the data comes from soils without glauconite.

3 RESULTS AND INTERPRETATION

3.1 *AUC Sensitivity to Depth Window*

Depth window over which the statistics take place (calculations regarding the $I_{interbed}$, median values of I_c, B_q, Q_t, and F_r) is investigated by changing the resolution from 0.5 to 4 m incrementally. All windows yielded high AUCs, typically exceeding 0.95. Shallower windows (e.g., 1 m) tended to produce slightly higher AUCs. This suggests that finer vertical resolution may enhance the model's ability to detect glauconite layer transitions, possibly due to sharper stratigraphic contrasts.

3.2 *Generalization Capability*

Figure 2 presents ROC curves for the 1-meter depth window, generated from 1000 bootstraps.

Each curve represents a model trained on a bootstrapped sample and evaluated on either the training or test set. The median, 16th, and 84th percentile curves are shown to illustrate the central tendency and variability in model performance. Higher areas under the curves (AUCs) reflect better model efficacies. As expected, the training ROC curves exhibit tighter bands and higher AUCs, while the test curves show broader variability, reflecting generalization uncertainty.

In the training dataset (Figure 2a), AUC values are tightly clustered above 0.95, with many approaching 1.0. The vertical spread is minimal, suggesting low variance across bootstraps and consistent learnability of the training data. These results reflect the expected behavior of ensemble models trained on resampled data: high performance with low dispersion.

In contrast, the test dataset (Figure 2b) reveals somewhat of a broader spread in AUC values, with more frequent dips into the 0.85–0.95 range and occasional values below 0.85. This increased scatter is deemed acceptable and expected when evaluating generalization on unseen data. The dispersion is noticeably higher than in the training set, as expected.

Together, these results confirm that the model performs well across a range of depth window lengths, with tighter and more consistent performance on training data and slightly more variable but still acceptable performance on test data. The use of bootstrapping and stratified splits has proven effective in quantifying this generalization gap and ensuring that the model's predictive capabilities are not overestimated.

3.3 *Feature importance via SHAP*

SHAP (SHapley Additive exPlanations) values were used to quantify the marginal effect of each feature on the model's

predicted probability of glauconite presence. Figure 3 presents a "SHAP beeswarm" plot summarizing results from 1000 bootstrapped models.

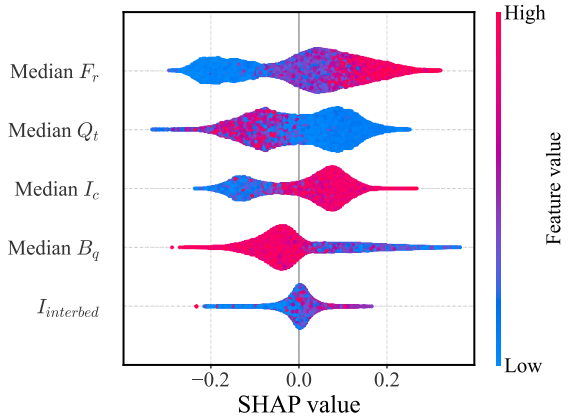

Figure 3. Aggregated feature contributions to glauconite probability using 1000 bootstraps on the combined dataset.

Each dot represents a single prediction:

- The horizontal axis shows the SHAP value, indicating how much that feature shifted the prediction relative to the average.
- The color encodes the actual feature value (from low to high).

Interpretation highlights:

- Bulges or dense regions indicate where many predictions had similar SHAP values. A bulge far to the right suggests that the feature consistently increased the predicted probability of glauconite presence; far left indicates a consistent decrease. This pattern is most evident in F_r and I_c.
- The spread of SHAP values reflects the variability in the feature's influence across samples. In this regard, the most influential factors are F_r and Q_t.
- Color gradients across the bulge reveal directional effects: for instance, if high feature values (red) cluster on the right, then higher values of that feature tend to increase glauconite presence probability. Using the combined dataset, high values of F_r, I_c, and $I_{interbed}$ (in red) tend to increase glauconite presence probability, while high values of Q_t and B_q generally reduce it.

This visualization provides both global and local interpretability, showing not only which features are important, but how their values influence model behavior across different conditions.

Most notably, the $I_{interbed}$ shows a flat and narrow SHAP distribution centered around zero. This indicates that, in the combined dataset, interbeddedness is not the strongest predictor of glauconite presence. This observation aligns with earlier findings that interbeddedness might be more relevant in allogenic settings (which are common offshore), and its predictive power diminishes in authigenic settings (which can be found offshore, but more often are onshore source deposits for allogenic offshore deposits). This reinforces the importance of domain-specific modeling and motivates the comparative analysis presented in Section 4.

4 DOMAIN-SPECIFIC BEHAVIOR: AUTHIGENIC VS ALLOGENIC

To explore domain-specific behavior, the models were retrained separately on the authigenic and allogenic datasets. While the same modeling pipeline was applied, SHAP analyses revealed a key divergence. In the offshore (which happens to be allogenic) dataset, interbeddedness index emerged a more dominant predictor. In the onshore (authigenic, in this work) dataset, $I_{interbed}$ showed minimal predictive value, behaving nearly randomly.

This finding aligns with earlier work (Geyin et al., 2023), where $I_{interbed}$ was identified as an indicator in offshore settings (Figure 4). The current study extends that insight by demonstrating its limited utility in authigenic environments, likely due to differing depositional processes or stratigraphic discontinuity, which may be linked to cementation, among other characteristics of authigenic deposition.

However, it is important to note that this contrast may not generalize to all allogenic or authigenic settings. The observed divergence could reflect specific characteristics of the datasets used here, rather than a universal rule. Further studies across varied geologic contexts are needed to confirm whether this pattern holds consistently.

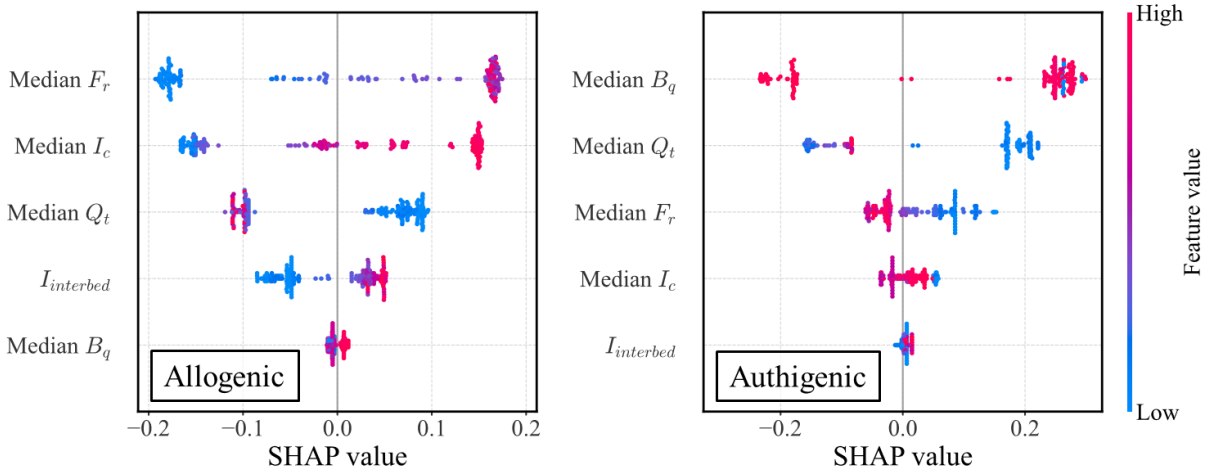

Figure 4 Allogenic vs. Authigenic Predictor Importances via SHAP Analysis

These results underscore the importance of domain-aware modeling and caution against overgeneralizing feature relevance across geologic settings. The combined dataset provides a robust foundation for general prediction, but domain-specific models may yield improved performance in targeted applications. The $I_{interbed}$, while valuable offshore, may

not reflect meaningful transitions in more homogeneous deposits. This insight has implications for feature selection and model interpretability in future glauconite prediction efforts.

5 CONCLUSIONS

This study presents the first known modeling framework aimed at predicting glauconite presence using CPTu data alone. By integrating CPTu-derived features with borehole-confirmed glauconite presence, and applying a Random Forest classifier within a Positive-Negative-Unlabeled (PNU) learning framework, the approach offers a practical and scalable method for identifying glauconite presence in offshore and onshore settings.

Key findings include:

- High predictive performance was achieved across all tested depth windows, with AUC values consistently exceeding 0.95. Shallower windows (e.g., 1 m) offered slightly better resolution, likely due to sharper stratigraphic contrasts.
- SHAP analysis revealed that F_r and Q_t are the most influential predictors. The interbeddedness index, while potentially useful in allogenic settings, exhibited limited predictive value in authigenic settings, a trend likewise observed for the B_q parameter.
- Domain-specific retraining highlighted the importance of depositional context. Interbeddedness was a reasonable indicator in the allogenic setting but behaved nearly randomly in the authigenic setting, suggesting that its utility is not universal and may depend on local stratigraphy and sedimentation processes.

While the results are promising, they should be interpreted with caution. The use of bootstrapping was essential to quantify model uncertainty and generalization variability. The observed patterns, particularly those related to domain-specific behavior, may reflect characteristics of the datasets used and not necessarily generalize to all environments containing glauconite.

Overall, this work demonstrates that CPTu-based classification, when combined with machine learning and interpretability tools like SHAP, can provide valuable insights into glauconite presence, which is important for site characterization and risk assessment for pile foundations. It also underscores the need for domain-aware modeling and careful feature selection when applying such models to new geologic settings. Future work should explore broader datasets and incorporate additional geotechnical parameters to further validate and refine the approach.

Further research in this area will support the development of CPT-based design methods for foundations in glauconite-rich deposits, where it is essential not only to identify glauconite but also to characterize key behavioral features—ranging from sand-like to clay-like—for analyzing soil–foundation interaction during both installation and operations.

6 REFERENCES

Batista, G.E., Bazzan, A.L., and Monard, M.C. 2003. Balancing training data for automated annotation of keywords: a case study. *Wob*, *3*, 10-18.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

DeGroot, D.J., Westgate, Z.J., and Yetginer-Tjelta, T.I. 2023. Geological and geotechnical challenges for US offshore wind farm development. Invited Keynote Paper – *Proceedings of 9th Offshore Site Investigation and Geotechnics (OSIG) International Conference*, London, England, UK.

De Nijs, R.E.P., Kaalberg, F.J., Osselaer, G., Couck, J.V., and Van Royen, K. 2015. Full scale field test (sheet)pile drivability in Antwerp (Belgium). *Geotechnical Engineering for Infrastructure and Development*, 1085-1090.

Diaconis, P., and Efron, B. 1983. Computer-intensive methods in statistics. *Scientific American*, 248(5), 116-131.

Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861–874.

Geyin, M., Madyarova, M.M., Dantal, V., and Kassa, H.M. 2023. An SBI-Based interbeddedness index for the prediction of glauconitic materials for offshore geotechnics. *Proceedings of 9th Offshore Site Investigation and Geotechnics (OSIG) International Conference*, London, England, UK.

Joustra, K. and De Gijt, J.G. 1982. Results and interpretation of cone penetration test results in soils of different mineralogic composition. *Proc. 2nd Euro. Symp. Penetration Testing: ESOPT II*, Amsterdam, 24-27 May 1982, A.A. Balkema, Rotterdam, 615-626.

Long, X., Tucker, G., Gibbs, P., Westgate, Z., Diaz, A.T., and Senanayake, A. 2019. Soil classification and evaluation of preconsolidation stress of Atlantic Outer Continental Shelf OCS sediments from oedometer and cone penetration testing. *Proceedings of the Offshore Technology Conference (OTC 2019)*. Houston, TX, USA.

Lundberg, S.M., and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–77. NIPS'17. Red Hook, NY, USA: Curran Associates Inc

Perikleous, G., Meissl, S., Diaz, A.T., Stergiou, T., and Ridgway-Hill, A. 2023. Monopile installation in glauconitic sands. *Proceedings of 9th Offshore Site Investigation and Geotechnics (OSIG)* International Conference, London, England, UK.

Philibert, G., Todd, B.J., Campbell, D.C., King, E.L., Normandeau, A., Hayward, S.E., Patton, E.R. and Campbell, L. 2024. Updated surficial geology compilation of the Scotian Shelf bioregion, offshore Nova Scotia and New Brunswick, Canada. *Geological Survey of Canada*, Open File 8911 (revised).

Pisanò, F., Westgate, Z., Rahim, A., Maldonado, C., Komurka, V., Beemer, R., Stuyts, B., Hamre, L., Eiksund, G., Liedtke, E., Perikleous, Y., Ridgway-Hill, A., De Sordi, J., Roux, A., Jones, L., and Ghasemi, P., 2025. The Piling in Glauconitic Sand (PIGS) JIP: insights from axial and lateral pile load testing. *Proceedings of 5th International Symposium on Frontiers in Offshore Geotechnics (ISFOG)*, Nantes, France, 9–13 June 2025.

Poag, C.W. (1978). Stratigraphy of the Atlantic Continental Shelf and Slope of the United States. *Annual Reviews Earth Planet Science* 1978.6 251-280.

Robertson, P.K. 2011. Computing in geotechnical engineering - automatic software detection of CPT transition zones. *Geotechnical news*, 29(2), 33.

Robertson, P.K.; Wride, C.E. 1998. Evaluating cyclic liquefaction potential using cone penetration test. *Canadian Geotechnical Journal*. 35 (3), 442–459.

Rogiers, B., Mallants, D., Batelaan, O., Gedeon, M., Huysmans, M., and Dassargues, A. 2017. Model-based classification of CPT data and automated lithostratigraphic mapping for high-resolution characterization of a heterogeneous sedimentary aquifer. *PloS one*, 12(5), e0176656.

Trumbull, J.V.A. 1972. Atlantic Continental Shelf and Slope of the United States: sand-size fraction of bottom sediments, New Jersey to Nova Scotia. *US Government Printing Office*.

Van Alboom, G., Maertens, J., Dupont, H., and Haelterma, K. 2012. Glauconiethoudende zanden (in Dutch). *Geotechniek*, April, 32-37.

Westgate, Z.J., Beemer, R.D., and DeGroot, D.J. 2023a. Implications of glauconite sand on U.S. offshore wind development. *Proc. SUT Offshore Site Investigation and Geotechnics Conf.*, London.

Westgate, Z.J., DeGroot, D.J., McMullin, C. et al. 2023b. Effect of degradation on behaviour of glauconite sands from the U.S. Mid-Atlantic Coastal Plain. *Ocean Engineering*, Vol. 283, 115081.

Westgate, Z.J., Rahim, A., Senanayake, A., Pisanò, F., Maldonado, C., Ridgway-Hill, A., Perikleous, Y., De Sordi, J., Roux, A., Andrews, E., and Ghasemi, P. 2024. The Piling in Glauconitic Sands (PIGS) JIP: Reducing geotechnical uncertainty for U.S.

offshore wind development. *Proceedings of the Offshore Technology Conference (OTC 2024)*. Houston, TX, USA.

Westgate, Z.J., DeGroot, D., Zhang, G., Beemer, R., Miller, K., Browning, J., Coffman, R., Senanayake, A., and Maldonado, C., 2025. The Piling in Glauconitic Sand (PIGS) Joint Industry Project (JIP): Insights from site characterisation and laboratory testing. *Proceedings of 5th International Symposium on Frontiers in Offshore Geotechnics (ISFOG)*, Nantes, France, 9–13 June 2025.

Wilson, D. 1972. Asymptotic properties of nearest neighbor rules using edited data. In *IEEE Transactions on Systems, Man, and Cybernetrics*, vol. 2 (3), pp. 408-421, 1972.

APPENDIX - GLOSSARY

This appendix defines selected machine-learning and data-processing terms referenced in Sections 2–4, with the aim of supporting readers less familiar with these concepts.

Bootstrapping – A statistical resampling method in which datasets are sampled with replacement to create multiple "bootstrapped" versions of the data to account for the finite sample uncertainty. Each version is used to retrain and evaluate the model, allowing estimation of performance variability and confidence intervals if the model structure.

Cross-Validation – A model evaluation technique that partitions the available data into multiple training and validation subsets ("folds"). The model is trained and tested repeatedly across these folds to assess generalization and reduce overfitting risk. In k-fold cross-validation, the data are split into k subsets, each serving as a validation set once.

Receiver Operating Characteristic (ROC) Curve – A graphical plot showing the trade-off between a model's true positive rate (sensitivity) and false positive rate (1 – specificity) across different decision thresholds (e.g., predicted probabilities).

Area Under the Curve (AUC) – A numerical summary of a ROC curve, with values ranging from 0.5 (no better than random) to 1.0 (perfect discrimination of positive and negative classes). In this study, higher AUC values indicate better ability to distinguish between CPTu intervals with and without glauconite.

Positive–Negative–Unlabeled (PNU) Learning – A supervised learning framework used when some samples are labeled as "positive" (e.g., glauconite present) or "negative" (absent), while many samples are unlabeled. The model is trained using confirmed positive and negative examples, while ignoring unlabeled cases to avoid introducing uncertainty.

Random Forest (RF) – An ensemble learning algorithm that builds multiple decision trees using randomly selected subsets of the training data and features. Each tree outputs a probability based on how similar training samples are distributed in its decision path. The forest's final prediction is the average of these probabilities across all trees. RFs are robust to nonlinear relationships, multicollinearity, and noisy data, making them well-suited for complex classification tasks.

Logistic Regression (LR) – A classification method that estimates the probability of a binary outcome using a logistic (sigmoid) function applied to a linear combination of input features. LR assumes a linear relationship between the features and the log-odds of the outcome, producing a continuous probability between 0 and 1 that can be thresholded to make a class prediction.

Hyperparameters – Model settings that are fixed prior to training (e.g., number of trees, maximum tree depth in RF) and tuned to balance model complexity and generalization. Number of trees, each tree's depth, and the minimum samples left at each lef node are all hyperparameters.

Overfitting – A modeling problem in which the model learns patterns that are specific to the training data but fail to generalize to new, unseen data. This often results in inflated training performance and degraded test performance.

Class Imbalance – A condition where one class (e.g., CPTu intervals with glauconite) occurs much less frequently than the other. This imbalance can bias the model towards predicting the majority class if not considered well.

SMOTE (Synthetic Minority Oversampling Technique) – An oversampling method used to improve class balance in imbalanced datasets. Instead of duplicating existing minority class samples, SMOTE generates synthetic ones by interpolating between a sample and one of its nearest neighbors in feature space. This creates new, plausible examples that lie along the line segments connecting minority class samples, helping the model learn a more general decision boundary.

ENN (Edited Nearest Neighbors) – A data-cleaning technique that removes ambiguous samples lying close to the decision boundary between classes, helping the model make clearer distinctions.

SMOTEENN – A hybrid technique that combines SMOTE (oversampling) and ENN (cleaning) in sequence, increasing minority class representation while removing noisy or borderline samples.

Feature Engineering – The process of selecting, transforming, and constructing input variables ("features") from raw data to improve model performance. In this study, features included statistical descriptors of CPTu measurements and derived indices such as interbeddedness.

Feature Importance – A measure of how much each input variable contributes to a model's predictive accuracy.

SHAP (SHapley Additive exPlanations) – A model-agnostic method based on cooperative game theory that quantifies the marginal contribution of each feature to a given prediction, enabling both global and local interpretability.

Beeswarm Plot – A visualization of SHAP values where each point represents a sample, showing both the magnitude and direction of each feature's influence on model predictions, along with the feature's actual value.

Generalization – The ability of a trained model to maintain high predictive performance on new, unseen data, rather than only on the data used for training.

Domain-Specific Modeling – The practice of tailoring model training and interpretation to a particular subset of the data (e.g., authigenic or allogenic depositional settings) to capture patterns that may not generalize across all contexts.