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/ICSMGE2026-138

Machine learning for predicting soil penetrometric profiles: key challenges and feature engineering

Eduardo Martínez García, Marcos García Alberti
Departamento de Ingeniería Civil: Construcción, E.T.S. de Ingenieros de Caminos, Canales y Puertos, Universidad Politécnica de Madrid, Madrid, Spain, emartinez@menard.es

Antonio Alfonso Arcos Álvarez
Departamento de Ingeniería y Morfología del Terreno, E.T.S. de Ingenieros de Caminos, Canales y Puertos, Universidad Politécnica de Madrid, Madrid, Spain

ABSTRACT: Artificial Intelligence (AI) and Machine Learning (ML) have seen rapid advancements in recent years, with geotechnical engineering increasingly benefiting from these technologies. ML applications in this field have covered a wide range of topics, including bearing capacity, soil and rock characterization, landslides, and tunneling. However, a common challenge in geotechnical engineering is the scarcity and uncertainty of available data. Furthermore, during the drilling of rocks and soil, modern rigs measure the parameters of the process. These values are an invaluable source of information for applying ML. This study shows the findings of a research focused on using ML to obtain soil penetrometric profiles from drilling data collected during the installation of rigid inclusions across various sites. The study addresses the key challenges encountered when applying ML to geotechnical data, such as data formatting, anomaly detection, dataset preparation, and selecting optimal algorithms. Additionally, it discusses the creation of informative features that enhance algorithmic performance. The results show that ML can be applied successfully to obtain the penetrometric profile of the soil at each of the points where a borehole with parameter measurement is executed. They also show how the proposed additional features can significantly improve prediction accuracy.

KEYWORDS: Machine Learning, penetrometric profile, drilling data, rigid inclusions, feature engineering.

1 INTRODUCTION

The scarcity and uncertainty of in-situ data remain major obstacles for reliable geotechnical design. Modern rigs, however, record torque, rotation speed, penetration rate and thrust in real time. If these signals were converted into a continuous penetrometric profile, each perforation would act as its own quasi-CPT/DPSH test, reducing the need for dedicated soundings, improving site safety and enabling foundation optimisation.

This idea has been explored by the authors previously (e.g. (Martínez García, García Alberti and Arcos Álvarez, 2025)); here we emphasize the engineering of new features and their effect on predictions.

2 DATASETS

Data for this study were collected at five sites in Spain and France. Four sites employed dynamic penetrometers. The remaining site used cone penetration tests (CPTs).

Drilling data were obtained during the installation of displacement rigid inclusions. Recorded variables were torque, rotation speed, penetration rate, and thrust. Table 1 presents one representative data row (i.e. one sample). Note that the cone resistance unit depends on the type of penetrometer.

Table 1. Representative data row.

Torque (Nm)	Rotation speed (rpm)	Penetration rate (m/hr)	Thrust (kN)	Cone resistance
18548.6	14.1	143.6	151.2	7

After cleaning, the dataset contains several hundred thousand samples.

3 APPLICATION OF ML

A well-designed ML pipeline is required to extract maximum information from the data. The main steps to consider are listed below. Some steps may be reordered or applied iteratively.

- Exploratory data analysis (EDA).
- Data cleaning and processing.
- Target and metrics
- Train-Test split
- Machine learning model
- Feature engineering
- Error analysis
- Hyperparameters tuning
- Explainability

3.1 *EDA*

After data collection, EDA is the first, and one of the most important, stages of the ML process. The objective of the EDA is to know the data: their distribution, relationships, anomalies, missing values, etc.

3.2 *Data cleaning and processing*

Even with perfectly recorded data, they usually need some processing to be used in ML. Sadly, as Daktera and Janodet (2024) pointed out, geotechnical engineering may not be ready yet for ML in the sense that data is shared in heterogeneous, non-machine-readable formats.

A good practice is to convert data to a standard such as AGS(Association of Geotechnical & Geoenvironmental Specialists, 2023) before further processing. This conversion makes subsequent ML steps reproducible and reusable.

Another part of the preprocessing is to deal with anomalies. Adopting AGS eliminates purely technical anomalies (format or unit errors). Recording-related anomalies are harder to handle. For penetrometers, a value may be anomalous only within its specific layer, so global-mean techniques are unsuitable. In this case, an easy approach is to use a moving average of the penetrometer or consider the samples too far from that average to be anomalies.

On the other hand, anomalies from measurement while drilling (MWD) are multivariate and thus harder to detect. A sample may be anomalous in one feature only when considered alongside others (e.g. high torque at high penetration rate). In

such cases, specialised algorithms like DBSCAN or Isolation Forest are appropriate.

3.3 *Target and metrics*

Identifying the metrics that best suit our target is not always easy. The prediction target is cone-penetration resistance (expressed as blows or q), a numerical quantity. Standard regression metrics—R^2, RMSE, MAE, etc.—can therefore assess model performance.

However, for penetrometers the shape of the predicted profile is also critical. To check if the shape is like the real one, we can use Dynamic Time Warping (DTW) which gives as a value of the distance between the two series. This approach for geotechnical data was proposed by Charles et al (2023) to compare synthetic CPTs.

3.4 *Train test split and dataset building*

For penetrometers, the main consideration when building the dataset is the distance between the sounding point (the penetrometer) and the drilled inclusion. As the perforation departs from the penetrometer location, soil conditions change, and the prediction error therefore increases. The magnitude of this effect depends on the spatial variability of the soil. If the ground is fairly uniform the effect is minor, but in highly variable soils the modelling error can be large.

Bunieski (Bunieski, 2022) tackled this issue at the previous edition of this conference, proposing the variability and continuous sampling methods to consider soil heterogeneity when building the dataset.

3.5 *Machine learning model*

Choosing the most suitable ML algorithm for a given case is not straightforward. In this study we relied on AutoML tools to select the most appropriate algorithm type. Ensemble trees algorithms such as Random Forest, ExtraTrees or XGBoost yielded the best results, as expected for tabular data, because they naturally handle mixed data types, nonlinear relationships, and feature interactions with minimal preprocessing. They are robust to missing values and outliers, work well even with limited data, and leverage ensemble techniques to reduce overfitting and boost accuracy.

3.6 *Feature engineering*

Feature engineering can refer to different kinds of techniques to adapt the features to obtain a better performance model. Several approaches exist:

- Normalize the values.
- Eliminate non-informative features, for example with principal component analysis (PCA).
- Create new features that capture relevant patterns, for example using a polynomial combination of the features.

Later we detail the creation of new features related to the drilling process.

3.7 *Error analysis*

Error analysis examines not only model performance but also the reasons behind it. Does the model meet the specified requirements? Is the model overly adapted to the training set (overfitting)?

More often than not, error analysis leads to hyperparameter tuning.

3.8 *Hyperparameters tuning*

An algorithm's hyperparameters are settings that must be tuned before training to constrain the set of possible models. This has two main objectives:

- Lower the computational cost of training the model.
- Avoid overfitting.

The tuning process can be a complex process that depends on the algorithm and on the data being fed to it. As noted earlier, AutoML tools such as TPOT(Olson et al., 2016) can simplify algorithm selection and tuning.

3.9 *Explainability*

In ML, explainability encompasses techniques that clarify why a model produces particular results. Individual Conditional Expectation (ICE) plots and SHAP values are useful in this regard. These tools reveal how the model makes decisions, but they do not necessarily reflect true causal relationships among features.

4 ADVANCED FEATURE ENGINEERING FOR RIGID INCLUSIONS

A typical ML challenge is a high feature-to-sample ratio or the presence of non-informative predictors. In our case the problem is the opposite: only four raw features—torque, rotation speed, penetration rate and thrust—are available. Are four features sufficient to model the complex relationship between the penetrometer and the soil? Probably not. Yet the parameters recorded during drilling are limited. Where, then, can we obtain additional information for the algorithm?

4.1 *Information from the drilling process of rigid inclusions*

This study uses data from five rigid-inclusion sites. These inclusions were drilled with soil displacement, without extraction. In this method, the rig's torque is transmitted along the drilling-tool shaft.

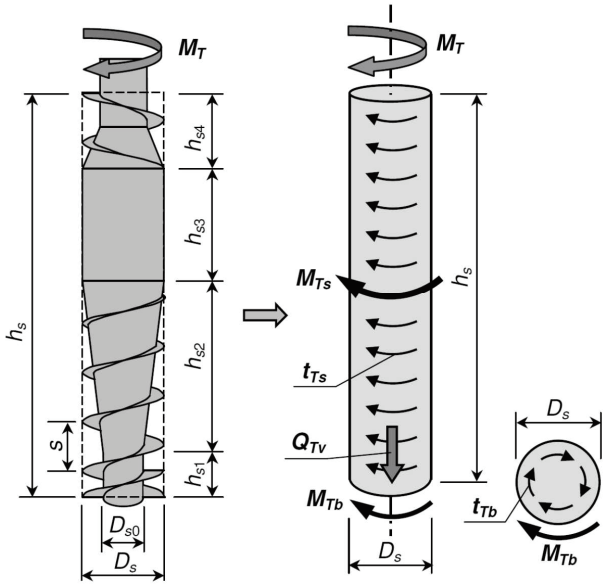

Figure 1. Simplified forces scheme in the drilling tool during perforation (Krasinski, 2015).

Therefore, the parameters of execution are affected by the soil above the tip along the length of the drilling tool. They are also affected by soil below the tip, but that information is unavailable when predicting the penetrometer value at the tip.

In our case, the tool geometry displaces ground upward to roughly 1.5 m above the tip, after which the soil moves laterally. We could include all signal values over that 1.5 m band, but doing so would inflate the feature set and risk the "curse of dimensionality".

Instead, we use the average values of the four signals (torque, rotation speed, penetration rate and thrust) over that interval.

Means alone may not capture terrain behaviour, so we also introduce a correlation vector. Sapronova and Marcher (Sapronova and Marcher, 2023) proposed the correlation vector as a way to extract patterns from aggregated data. The vector comprises the pair-wise correlation coefficients between the features. Table 2 lists the correlation vector correlation vector corresponding to the matrix in Figure 2.

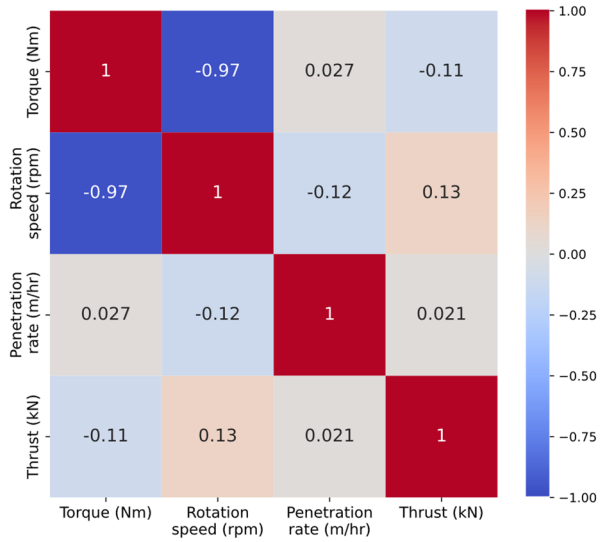

Figure 2. Example of correlation matrix obtained.

Table 2. Correlation vector correspondent to the correlation matrix of Figure 2.

Torque / Rotation speed	Torque / Penetration rate	Torque / Thrust
-0.97	0.027	-0.11

Rotation speed / Penetration rate	Rotation speed / Thrust	Penetration rate / Thrust
-0.12	0.13	0.021

The features used to fit the model and make predictions would be:

- Original signals: torque, rotation speed, penetration rate and thrust.
- The mean values of these four features along the 1.5 m above the tip of the tool.
- Six pair-wise correlation coefficients (correlation vector).

In total we use 14 features, still modest relative to the available sample size.

5 RESULTS

Table 3 reports test-set metrics for the default ExtraTrees implementation in scikit-learn, using Bunieski's continuous method and the four original features.

Table 3. Metrics for ExtraTrees with the four original features.

Site	R	R^2	RMSE	MAE
1	0.73	0.53	3.79	2.67
2	0.36	0.07	2.73	1.68
3	0.59	0.33	3.00	2.13
4	0.46	0.18	6.59	4.50
5	0.60	0.36	2.71	1.93

Where:

- R = Taylor correlation coefficient.
- R^2 = coefficient of determination.
- RMSE = root-mean-square error.
- MAE = mean absolute error.

On the other hand, Table 4 shows the metrics using the additional features from the mean values and the correlation vector.

Table 4. Metrics obtained for ExtraTrees with the additional features.

Site	R	R^2	RMSE	MAE
1	0.94	0.88	1.89	1.18
2	0.85	0.70	1.54	0.90
3	0.89	0.78	1.73	1.03
4	0.83	0.68	4.15	2.47
5	0.83	0.67	1.93	1.30

Adding the new features yields substantial gains: correlations rise markedly, and errors fall. This stands out clearly in the Taylor diagram of Figure 3.

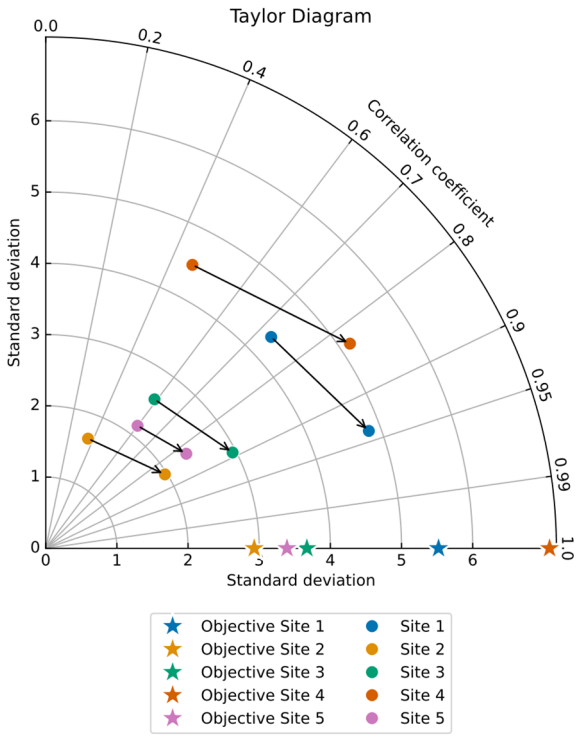

Figure 3. Taylor diagram showing the improvement from the new features.

The diagram plots, for each site, the model's correlation coefficient and standard deviation. Incorporating the proposed features brings predictions much closer to the observed penetrometer distributions. Please note that the centered RMSE normally shown in this kind of plot is not shown here, as it is different for every site, to maintain readability.

6 DISCUSSION

This study's main objective was to determine whether ML models trained solely on measurement-while-drilling (MWD) data can reliably reconstruct a continuous penetrometric profile at every inclusion location.

Results with the baseline feature set—torque, rotation speed, penetration rate and thrust—show that these four raw

signals alone are insufficient to capture the complex interactions during drilling. Although, the continuous method used to build the dataset introduces data that are several meters away from its penetrometers. Considering low distances, the metrics of the "baseline" feature set improve.

In any case, the addition of (i) running means over the 1.5 m of soil that actively resists the displacement tool and (ii) the six-element correlation vector produces a pronounced improvement in the metrics.

6.1 *Why do the new features help?*

Running means approximate the integrated effect of the overlying 1.5 m of ground that the screw displacement tool must displace upwards before lateral movement begins. This acts as a proxy for short-range stratigraphy and reduces the "measurement noise" coming from local heterogeneities at centimetre scale.

Correlation-vector elements encode the internal coupling between signals (e.g. high torque at high penetration rate), turning a purely pointwise description into a pattern-recognition problem that tree-based ensembles can exploit.

Magnitude of the gain (Table 4 vs Table 3): average R rose from 0.55 to 0.87, average R^2 from 0.29 to 0.74, while RMSE roughly halved.

6.2 *Limitations*

The models were trained and validated using data from individual sites, meaning that the predictive performance was assessed in a site-specific rather than cross-site manner. While good results were obtained for several locations, this does not yet demonstrate true generalisation. Geological conditions, soil behaviour, and operational practices can vary significantly between sites, and these variations may affect the relationship between MWD parameters and target geotechnical properties. Therefore, the models' ability to reliably transfer knowledge from one site to another remains unproven and should be tested using cross-site training and validation frameworks or leave-one-site-out approaches.

In the same way, all the data come from screw displacement inclusions. Other techniques, such as Continuous Flight Auger (CFA) piles or rotary flight augers, differ in tool geometry, energy transfer mechanisms, spoil extraction, and soil disturbance patterns. Consequently, the MWD signal signatures and their correlation with in-situ resistance or blows may change considerably.

The 1.5 m aggregation window is tool-specific; different tools may require recalibration.

The use of a 1.5 m aggregation window is specific to the mechanics and geometry of the tool used in this study. Different rigs or tools with alternative flight shapes, diameters, or penetration dynamics may require recalibration of this window to correctly capture representative behaviour. Likewise, perforation diameter was constant within each site and thus excluded as a feature. However, diameter directly influences stress distribution, torque demand, and penetration resistance. Future models should explicitly incorporate diameter to improve scalability across tools and contractors.

In addition to these methodological constraints, some parameters that influence drilling behaviour (tool wear, rig maintenance, operator decisions, soil moisture variations) are not recorded by the MWD system. These unmeasured factors introduce unavoidable uncertainty and may partially obscure the true soil–tool interaction. As a result, while the proposed approach shows strong potential, its robustness under diverse site conditions, drilling technologies, and operational environments requires further validation and enhancement.

6.3 *Practical implications*

With the refined feature set, every production borehole effectively becomes a quasi-continuous sounding whose quality ($R^2 \approx 0.7$–0.9) is sufficient for preliminary stratigraphic interpretation, optimisation of inclusion length, and near real-time quality control. The computational overhead is modest, meaning the workflow could be embedded in rig-side software.

7 CONCLUSIONS

Machine-learning models can transform standard MWD signals (torque, rotation speed, penetration rate and thrust) into a continuous profile of cone resistance or blows, turning each rigid-inclusion borehole into an additional penetrometric test.

Feature engineering is crucial. Simple, physically inspired aggregates (running means over the active displacement length) combined with a compact correlation vector raised the average coefficient of determination from ≈ 0.3 to ≈ 0.75 and approximately halved RMSE across five geologically distinct sites. However, these new features are tool-specific and should be calibrated if a tool with a different shape is used.

The enhanced models delivered consistently high correlations ($R \geq 0.83$) even at the most heterogeneous site, demonstrating the resilience of the approach.

Five sites are still a small sample, but the results obtained with the proposed methodology are very promising. With robust MWD acquisition and an appropriate ML pipeline, the approach could yield safer designs, optimised inclusion lengths, lower CO_2 emissions and reduced costs for this type of drilling.

8 REFERENCES

Association of Geotechnical & Geoenvironmental Specialists, 2023. *AGS Data Format*. [online] Available at: https://www.ags.org.uk/data-format/.

Bunieski, S., 2022. Practical methodologies to apply machine learning algorithms to ground improvement data. *20th International Conference on Soil Mechanics and Geotechnical Engineering*.

Charles, D.J., Axtell, M.D. and Gourvenec, S., 2023. Assessing the quality of synthetic CPT training data using time series similarity. *4th International Symposium on Machine Learning & Big Data in Geoscience*.

Daktera, T. and Janodet, L., 2024. Why geotechnical engineering isn't ready yet for machine learning? In: *Geotechnical Engineering Challenges to Meet Current and Emerging Needs of Society*, 1st edn. [online] London: CRC Press. pp.1539–1542. https://doi.org/10.1201/9781003431749-284.

Krasinski, A., 2015. The Analysis of Soil Resistance During Screw Displacement Pile Installation. *Studia Geotechnica et Mechanica*, 36(3), pp.49–56. https://doi.org/10.2478/sgem-2014-0026.

Martínez García, E., García Alberti, M.G. and Arcos Álvarez, A.A., 2025. Generation of Penetrometric Profile of the Soil Applying Machine Learning to Measure While Drilling Data from Deep Foundation Machinery. *Applied Sciences*, 15(3), p.1331. https://doi.org/10.3390/app15031331.

Olson, R.S., Bartley, N., Urbanowicz, R.J. and Moore, J.H., 2016. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. In: *Proceedings of the Genetic and Evolutionary Computation Conference 2016*. [online] GECCO '16: Genetic and Evolutionary Computation Conference. Denver Colorado USA: ACM. pp.485–492. https://doi.org/10.1145/2908812.2908918.

Sapronova, A. and Marcher, T., 2023. Correlational analysis for extracting patterns in geotechnical data. *4th International Symposium on Machine Learning & Big Data in Geoscience*.