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/ICSMGE2026-633

Data-driven framework for stress history prediction using machine learning with CPT data

Daeun Gwak
Department of Civil, Environmental and Plant Engineering, Konkuk University, Seoul, Republic of Korea

Taeseo Ku
Department of Civil and Environmental Engineering, Konkuk University, Seoul, Republic of Korea, tsku@konkuk.ac.kr

ABSTRACT: Geostatic stress history is influenced by various factors such as load changes, groundwater fluctuation, and environmental conditions. Understanding the stress history is essential for analyzing soil behavior and engineering properties, and it is typically characterized by the overconsolidation ratio (OCR). Laboratory methods, such as oedometer tests, have traditionally been used as the standard for directly determining the stress history. However, these methods are often considered time-consuming and unsuitable for certain soils, such as silt and sand. As a result, field-based approaches, particularly those utilizing Cone Penetration Test (CPT) data, have been widely adopted due to their ability to provide fast, practical insights and continuous soil profiling. Despite these advantages, concerns about the reliability and generalizability of CPT-based methods have been raised. To address these challenges, this study introduces a preliminary framework that integrates advanced machine learning techniques to improve stress history predictions using CPT data. This exploratory model combines unsupervised and supervised learning methods, aiming to establish a foundation for future advancements in geostatic stress prediction. A Random Forest (RF) algorithm is employed to analyze a comprehensive dataset compiled from well-documented literature and other global sources. Key CPT parameters, including cone tip resistance (q_t), porewater pressure (u_2), hydrostatic pressure (u_0), depth, normalized cone resistance (Q_t), and pore pressure ratio (B_q), are utilized as inputs for these models. Detailed data preprocessing, hyperparameter tuning, and 5-fold cross-validation are performed to ensure the model's robustness and accuracy. Finally, the proposed framework is validated through several case studies, designed to reflect a range of stress history conditions, allowing its performance to be assessed under diverse geotechnical scenarios.

KEYWORDS: cone penetration test, stress history, overconsolidation ratio, unsupervised learning, machine learning.

1 INTRODUCTION

In geotechnical engineering, stress history is a critical factor that influences various soil behaviors such as deformation, strength, and long-term stability. Stress history is commonly represented by the preconsolidation stress and the overconsolidation ratio (OCR), which are determined by a variety of geological and environmental factors including sedimentation, glaciation, seismic events, groundwater movements, and temperature variations (Chan and Poon, 2015, Leonards and Frost, 1988). Accurate estimation of these parameters is essential, especially in soils that exhibit significant variability or are sensitive to climatic changes.

Cone Penetration Test (CPT) is a widely recognized in-situ testing method, praised for its ability to continuously profile subsurface conditions efficiently. While CPT provides critical data, including corrected cone tip resistance (q_t), sleeve friction (f_s), and pore pressure (u_m), traditional approaches to estimating stress history often rely on empirical and/or analytical models that are dependent on site-specific assumptions. These methods, while useful, tend to have limited applicability in diverse or complex soil environments. To address these challenges, Gwak and Ku (2025) developed a supervised machine learning framework, incorporating selected models such as Deep Neural Networks (DNN), Support Vector Regression (SVR), Random Forest (RF), eXtreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LGBM). These models demonstrated strong performance, achieving high levels of accuracy and generalization by learning from labeled data.

This study extends the previous work by introducing a hybrid machine learning approach that integrates unsupervised learning techniques to reveal latent structures within the CPT data. Rather than focusing solely on improving predictive performance, the aim is to enhance the interpretability of the machine learning models by providing insight into the underlying patterns learned from the data. By incorporating unsupervised features into the supervised learning process, the proposed framework seeks to improve the understanding and transparency of stress history estimation in geotechnical applications.

2 BACKGROUND

2.1 *Stress history evaluation using CPT data*

The cone penetration test (CPT) provides continuous profiles of cone tip resistance (q_c), sleeve friction (f_s), and pore water pressure (u_m), offering high-resolution data for geotechnical characterization. When a pore pressure sensor is incorporated (CPTu), the measurement of excess pore pressure (u_2) becomes particularly valuable for evaluating stress-dependent soil behavior (Jamiolkowski et al., 1985).

To estimate the overconsolidation ratio (OCR), a key indicator of stress history, CPT-based approaches can be broadly classified into empirical and analytical methods. Empirical approaches rely on observed correlations between OCR and CPT parameters such as q_c, q_t, u_2, and normalized indices like Qt, B_q (Schmertmann, 1978, Smits, 1982). These methods are practical and widely applied in engineering practice but may be limited by site-specific or soil-dependent variability.

Analytical approaches, in contrast, use theoretical frameworks such as cavity expansion theory, stress path analysis, and dimensional analysis to interpret the cone response in relation to stress history. For instance, Mayne and Holtz (1988) integrated SHANSEP concepts with both cylindrical and spherical cavity expansion models to link CPT readings with preconsolidation stress. These methods offer a deeper understanding of stress mechanisms and improve generalizability across different soil conditions (Konrad and Law, 1987, Houlsby and Teh, 1988).

Although these approaches have proven useful within specific geologic contexts, they are often limited by their reliance on datasets derived from particular sites or soil types. As a result, their applicability to geotechnically diverse conditions can be constrained.

Table 1 summarizes representative empirical and analytical methods proposed in the literature, outlining their key parameters, theoretical background, and practical applicability.

Table 1. Summary of previous studies on OCR estimation using CPT data.

Empirical Method	**Reference**
$\frac{q_c - \sigma_{v0}}{\sigma'_{v0}}$	Schmertmann (1978)
$\frac{u_1}{q_c}$	Baligh et al. (1981)
$\frac{u_2 - u_0}{q_t}$	Campanella and Robertson (1981)
$\log OCR = \frac{u_1 - u_0}{\sigma'_{v0}}$	Azzouz et al. (1981)
$\frac{u_2 - u_0}{q_c - u_0}$	Smits (1982)
Analytical Approach	**Reference**
$2\left[\frac{(\frac{\Delta u}{\sigma'_{v0}})}{(\frac{2M}{3})\ln(I_R)}\right]^{(\frac{1}{\Lambda})}$	Schofield and Wroth (1968)
$2\left[\frac{\frac{\Delta u_2}{\sigma'_{v0}-1}}{(\frac{M}{2})\ln(\frac{G}{s_u})-1}\right]^{1/\Lambda}$	Mayne and Bachus (1988)
$\left[\frac{1}{1.95M+1}\left(\frac{q_t-u_2}{\sigma'_{v0}}\right)\right]^{1.33}$	Mayne (1991)
$\frac{0.46(q_t-u_2)}{\sigma'_{v0}}$	Chen and Mayne (1994)

Note: σ_{v0}=total vertical stress, σ'_{v0}=effective vertical stress, $M=6\cdot\sin\phi'/(3-\sin\phi')$, I_R=undrained rigidity index, $\Lambda=1-C_s/C_c$, C_s=swelling index, C_c=virgin compression index, s_u=undrained shear strength, G=shear modulus.

2.2 *Unsupervised learning*

Unsupervised learning refers to machine learning techniques that aim to identify hidden patterns or structures in data without relying on labeled outcomes. Among various unsupervised learning approaches, dimensionality reduction and clustering are commonly used to explore the intrinsic organization of high-dimensional datasets.

UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction method that preserves both local and global data structures while projecting high-dimensional data into a lower-dimensional space (McInnes et al., 2018). Its ability to maintain neighborhood relations makes it particularly effective for visualizing complex datasets and detecting inherent patterns.

Gaussian Mixture Model (GMM) represents another core unsupervised technique. GMM is a probabilistic clustering method that models the data as a mixture of multiple Gaussian distributions (Reynolds, 2015). This approach allows for soft clustering, where each data point can belong to multiple clusters with varying probabilities, offering flexibility in capturing complex or overlapping class boundaries.

Given these characteristics, UMAP and GMM present promising tools for exploratory data analysis in various domains. In geotechnical applications, such as interpreting CPT data, these methods may be leveraged to better understand subsurface variability and inform subsequent modeling efforts.

2.3 *Random Forest (RF)*

Random Forest (RF) is a versatile machine learning algorithm known for its robustness and ability to handle large datasets with multiple features. RF is an ensemble learning method that combines multiple decision trees to improve predictive performance (Breiman, 2001). Each tree in the forest is trained on a subset of the data, and the final prediction is made by aggregating the results from all trees. In the context of geotechnical applications, RF has been widely used for both regression and classification tasks, including the prediction of soil properties from CPT data. RF is particularly well-suited for handling non-linear relationships and interactions between features, which are common in geotechnical datasets (Feng et al., 2025, Ma et al., 2024). By using RF for unsupervised learning, it is possible to cluster CPT profiles into distinct groups, which can then be further analyzed for soil classification and property.

3 DATABASE

3.1 *Data compilation*

The first step in developing the machine learning-based approach was compiling a comprehensive global CPT database focused on clay soils and stress history indicators. This database significantly expands the one created by Chen and Mayne (1994), incorporating additional CPT data from various domestic and international sites. Only datasets containing stress history information were selected, and geological details such as groundwater level and stratigraphy were included to ensure data quality. The final dataset includes 429 CPT results obtained from 84 sites across the world. These data were compiled based on 19 published references. The parameters used in this study include q_t, hydrostatic pressure (u_0), u_2, depth, normalized cone resistance (Q_t), pore pressure ratio (B_q), and OCR. For the unsupervised learning application, only OCR was used, as the wide range of preconsolidation stress values made it unsuitable for clustering. The data covers various cohesive soils, including clay, and all data were collected from existing studies where soil types were clearly identified through laboratory testing or field investigations. A few OCR values slightly below 1.0 were included in the verified experimental data reported in previous studies. The original authors had already validated the reliability of their testing procedures, and these values were retained as they have no meaningful influence on the overall trends or model performance.

The summary of data sources and parameter ranges used in the CPT-based stress history database is provided in Table 2.

Table 2. Summary of data sources and parameter ranges in the CPT-based stress history database.

Metrics	**max**	**min**
q_t [kN/m^2]	1912.60	113
u_0 [kN/m^2]	152	0
u_2 [kN/m^2]	1493.17	22.33
Depth [m]	15	2
Q_t	37.16	0.45
B_q	1.48	-0.22
OCR	9.98	0.47

3.2 *Exploratory Data Analysis (EDA)*

This section presents the exploratory data analysis (EDA) conducted to assess the characteristics of the compiled CPT dataset. The main objectives were to examine the distribution of input variables, evaluate their correlation with the target variable (OCR), and identify potential issues such as multicollinearity and inter-site heterogeneity. These insights guided the subsequent selection and transformation of input features for both unsupervised and supervised learning.

EDA revealed that most variables do not follow a normal distribution. In particular, variables such as q_t, u_2, Q_t, and OCR

exhibit strong right-skewed distributions, with high kurtosis values, indicating the presence of outliers and non-Gaussian characteristics. Logarithmic transformation was applied to these variables to mitigate skewness and stabilize variance. Conversely, u_0 and B_q showed approximately symmetric distributions and were retained without transformation.

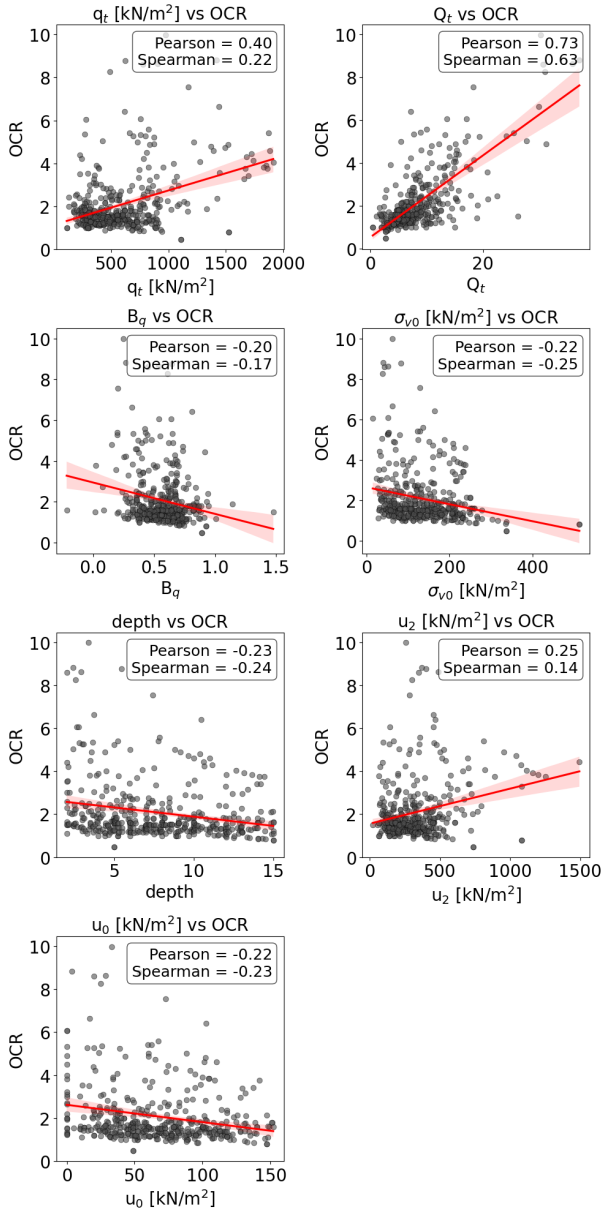

Figure 1. Scatter plots showing the relationship between OCR and key CPT variables. Each subplot includes both Pearson and Spearman correlation coefficients.

Based on these distributional characteristics, the relationships between input variables and OCR were then analyzed. The relationship between OCR and each input variable was explored through scatter plots, with Pearson and Spearman correlation coefficients computed to assess both linear and monotonic relationships (Figure 1). Among all variables, the Q_t demonstrated the strongest positive correlation with OCR (Pearson = 0.73, Spearman = 0.63), suggesting that it is the most influential predictor. Other variables, such as q_t, u_2, and B_q, showed weaker correlations and were considered as potential auxiliary features. Depth and total vertical stress (σ_{v0}) exhibited moderate negative correlations with OCR but were highly correlated with each other and with other variables, warranting careful selection or exclusion.

Multicollinearity among input variables was assessed through correlation coefficients and Variance Inflation Factor (VIF) analysis. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity among the predictors; a higher VIF indicates a stronger linear relationship with other variables. As shown in Table 3, severe multicollinearity was identified among q_t, u_2, σ_{v0}, depth, and u_0, all exhibiting VIF values exceeding 7. In particular, q_t and u_2 had the highest VIFs of 15.6 and 13.9, respectively, and showed strong mutual correlation ($r = 0.88$), suggesting substantial redundancy. σ_{v0}, depth, and u_0 also showed high pairwise correlations ($r > 0.84$), further confirming overlapping information. In contrast, Q_t and B_q exhibited low VIFs (3.27 and 3.24, respectively) and weak correlations with other features, supporting their relative independence and relevance as input variables in the subsequent modeling.

Table 3. Summary of data sources and parameters ranges in the CPT-based stress history database.

Feature	VIF	Correlated Variables (r)	
q_t	15.60	u_2 (0.88)	σ_{v0} (0.56)
u_0	7.70	depth (0.90)	σ_{v0} (0.84)
u_2	13.90	q_t (0.88)	σ_{v0} (0.60)
Depth	7.90	u_0 (0.90)	σ_{v0} (0.88)
Q_t	3.27	σ_{v0} (-0.40)	-
B_q	3.24	u_2 (0.40)	σ_{v0} (0.27)

4 METHODOLOGY

This study establishes a hybrid machine learning framework for estimating the OCR in cohesive soils using CPT data. The overall methodological procedure consists of three stages: (1) feature selection and preprocessing, (2) unsupervised learning using UMAP and GMM, and (3) supervised regression modeling augmented with clustering information. Each step was designed to enhance interpretability and predictive robustness while preserving geotechnical validity.

4.1 *Feature selection and preprocessing*

In the initial step, key input variables were selected based on their relevance to stress history and the results of the exploratory data analysis described in Section 3.2. The chosen parameters included depth, u_2, Q_t, and B_q. Among these, u_2 and Q_t exhibited strong positive skewness, and therefore underwent logarithmic transformation. The same transformation was applied to OCR to stabilize variance in the supervised regression phase.

Features such as q_t and σ_{v0} were excluded from the final model due to their high multicollinearity with other features, as confirmed through Variance Inflation Factor (VIF) analysis. All input features were then normalized using z-score standardization to ensure scale compatibility. Missing values, though infrequent, were handled by mean imputation to maintain data consistency across the dataset.

4.2 *Unsupervised learning: UMAP + GMM*

To extract latent structural patterns related to soil stress history, an unsupervised learning strategy was adopted that combines dimensionality reduction and probabilistic clustering. First, the selected input features were projected into a low-dimensional space using Uniform Manifold Approximation and Projection (UMAP). As aforementioned, this non-linear embedding method was chosen for its ability to preserve both local and

global data topology, allowing for effective structure discovery and visualization in a reduced dimension.

The embedded data were then clustered using the Gaussian Mixture Model (GMM), which probabilistically assigns each sample to clusters based on likelihood. This soft clustering approach offers flexibility in representing geotechnically transitional behavior between soil groups.

The optimal number of clusters (k) and UMAP dimensionality (d) were determined through Bayesian optimization, guided by three internal clustering validation metrics: the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). Among these, higher Silhouette and CHI values indicate better-defined and well-separated clusters, while lower DBI values reflect tighter intra-cluster cohesion and clearer inter-cluster separation.

4.3 *Supervised learning with cluster-augmented features*

The final stage involved constructing a supervised learning model to predict OCR values using both the original input features and the cluster labels derived from the unsupervised learning phase. A Random Forest regression algorithm was selected due to its robustness to noise, ability to capture non-linear relationships, and interpretability through feature importance.

The input vector consisted of depth, log-transformed u_2 and Q_t, B_q, and the categorical cluster label. Bayesian optimization was conducted to fine-tune the model's hyperparameters, including the number of trees, maximum depth, minimum samples per split and leaf, and feature subsampling ratio. Five-fold cross-validation was used during tuning to ensure generalizability, and a fixed random seed was employed for reproducibility. Once the optimal configuration was obtained, the model was trained on the full dataset and evaluated using standard regression metrics including coefficient of determination (R^2) and root mean square error (RMSE). These metrics were calculated for both the training and test sets to assess generalization performance. The inclusion of cluster labels aimed to embed latent stress history-related structure into the prediction process, and its contribution was quantitatively evaluated by comparing models with and without this feature.

Through this integrated methodology, the study seeks to improve OCR prediction in cohesive soils by combining physical feature engineering, unsupervised pattern discovery, and supervised regression in a unified framework.

5 RESULT AND ANALYSIS

5.1 *Unsupervised clustering performance*

To identify the optimal structure in the CPT dataset, an unsupervised clustering approach combining UMAP with GMM was adopted. Given the importance of selecting appropriate dimensionality and cluster count, a Bayesian optimization process was conducted using Optuna to systematically evaluate combinations of UMAP dimensions (d) and GMM clusters (k).

As shown in Figure 2, UMAP with d=2 consistently outperformed d=3 across all three metrics. The best configuration was achieved at d =2 and k=3, which resulted in clearly separated and physically interpretable cluster groups. The resulting cluster labels were appended to the original dataset and subsequently used as a categorical input feature in the supervised learning phase, embedding latent structural information into the model. Based on this outcome, the final UMAP-GMM clustering was performed using d=2 and k=3. Figure 3 shows the CPT samples projected onto a two-dimensional embedded space, with each point colored by its assigned cluster label. The three clusters are clearly separated in the embedded space. The internal clustering validation scores—Silhouette Score of 0.514, Davies-Bouldin Index of 0.697, and Calinski-Harabasz Index of 573.8—indicate that the overall clustering quality is reasonably good.

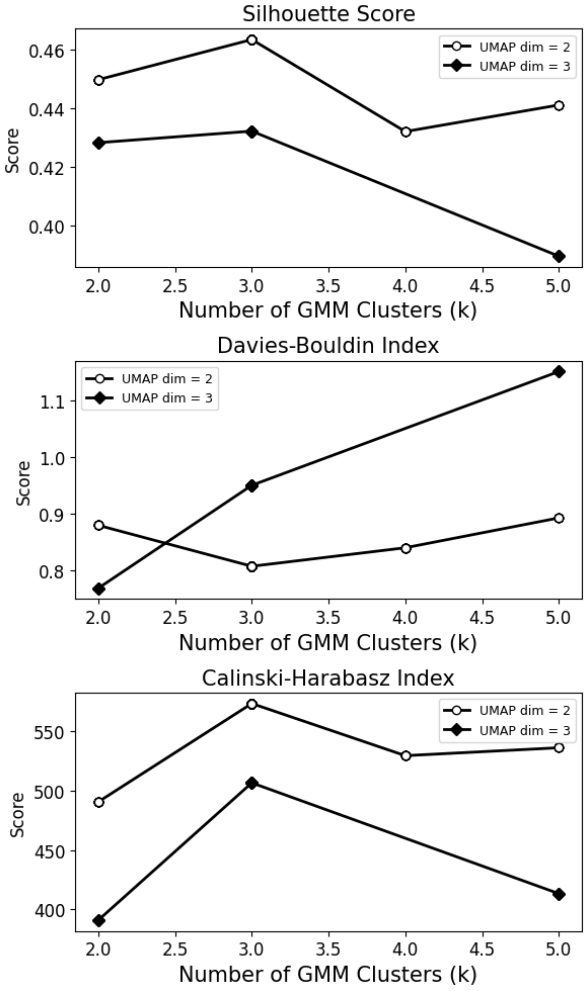

Figure 2. Clustering performance across UMAP dimensionalities (d=2 and d=3) and GMM cluster counts (k=2 to 5), evaluated using Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). Higher Silhouette and CHI values and lower DBI values indicate better clustering quality.

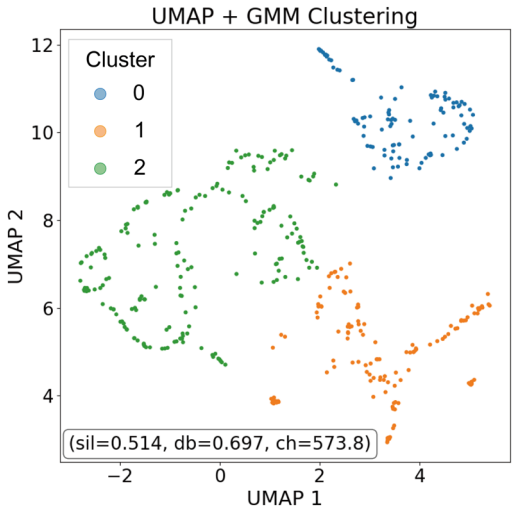

Figure 3. UMAP + GMM clustering result (UMAP dimension = 2, GMM clusters = 3). Each point represents a CPT sample, and colors indicate assigned cluster labels.

5.2 *Cluster characteristics and interpretation*

Figure 4 presents a pair plot of the input variables, where each data point is colored by cluster label, and diagonal plots show kernel density estimates (KDEs) for each variable by cluster. These plots provide insights into the distributional characteristics and separation of clusters across key CPT parameters.

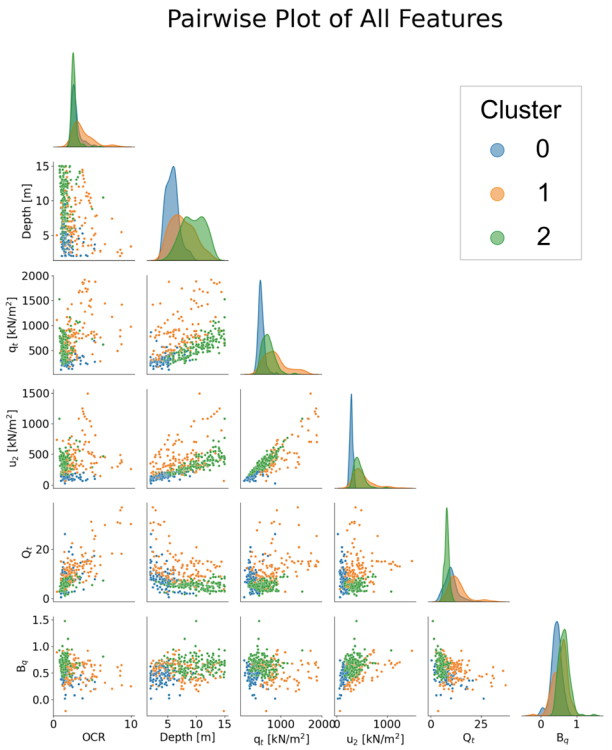

Figure 4. Pair plots of input variables colored by cluster labels. The clusters show meaningful separation in physical space, especially in q_t, u_2, and Q_t.

OCR shows the clearest separation among clusters. Cluster 0 is concentrated around low OCR values with a right-skewed distribution, consistent with normally consolidated soils. Cluster 2 displays a sharp, unimodal peak at higher OCR values, indicating a relatively high degree of overconsolidation. In contrast, Cluster 1 spans the full range of OCR values with a broad and relatively flat distribution, suggesting internal variability in stress history among samples. Depth also reveals distinct central tendencies for each cluster, with all three exhibiting multi-modal distributions. Cluster 0 is associated with shallow depths, Cluster 2 with deeper regions, and Cluster 1 with intermediate depths. This stratification aligns with typical stress history profiles and supports the geotechnical plausibility of the clustering. q_t and u_2 exhibit similar patterns and can be interpreted together. Interestingly, both q_t and u_2 exhibit a positive correlation, which suggests that higher cone resistance is accompanied by higher pore pressure in this dataset. Consequently, Cluster 0, characterized by low values in both variables, may represent a low-pressure, low-stiffness condition, rather than typical soft clays with high excess pore pressure. Cluster 2 occupies the middle range with moderate values and relatively compact distributions. Cluster 1 has the widest spread, covering both low and high values, and exhibits the highest variance, indicating significant heterogeneity or transitional behavior. Q_t shows clearer differentiation across clusters. Cluster 2 has low values with very narrow variance, resulting in a sharp unimodal density curve. Cluster 1 has relatively higher Q_t values and a broader spread, while Cluster 0 falls in between the two. Given that Q_t reflects normalized cone resistance, this variable further reinforces the stress history distinctions among clusters. B_q demonstrates the weakest separation among clusters, with substantial overlap in the KDE curves. Notably, Cluster 1 displays a multi-modal distribution, which may reflect internal diversity in pore pressure responses. Clusters 0 and 2 occupy different value ranges, but overall, inter-cluster distinction for B_q is less pronounced than for other variables.

Taken together, the combined interpretation of KDE and scatter plots confirms that the unsupervised clustering process successfully identified groups with statistically and geotechnically meaningful distinctions across multiple CPT variables.

5.3 *Supervised regression results*

The final Random Forest model incorporated cluster labels as a categorical feature. Performance was assessed using R^2 and RMSE on training, testing, and full datasets. Figure 5 presents the predicted vs. actual OCR values, where most predictions align closely with the 1:1 line, indicating strong generalization. To examine the added value of clustering, Table 4 compares the model performance with and without the use of cluster information. While the overall improvement in prediction metrics was moderate, the incorporation of clustering helped reveal latent patterns in the CPT data and provided a more interpretable structure for understanding stress history. These findings suggest that combining physical features with data-driven clustering can support not only predictive modeling but also geotechnical interpretation.

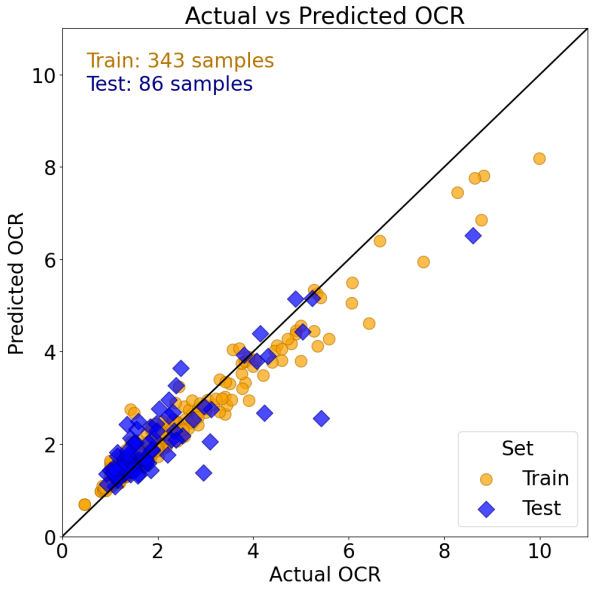

Figure 5. Predicted vs. actual OCR values for training and testing sets. The 1:1 reference line indicates accurate prediction across a wide OCR range.

Table 4. Performance metrics of the final RF model using hybrid learning approach

Dataset	**R^2**	**RMSE**
Train	0.94	0.36
Test	0.77	0.60
All	0.91	0.42

6 CONCLUSIONS

This study proposed a hybrid machine learning framework to enhance the prediction of OCR from CPT data. By integrating unsupervised clustering and supervised regression, the framework leveraged both physical features and latent data structure to improve prediction accuracy.

An unsupervised clustering process was conducted using UMAP for dimensionality reduction and GMM for probabilistic classification. A systematic parameter search using Optuna revealed that UMAP with 2 dimensions and GMM with 3 clusters provided the most effective separation, as validated by internal metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The resulting clusters were shown to align with geotechnically meaningful differences in stress history, as confirmed by the distributions of q_t, u_2, Q_t, and depth.

The derived cluster labels were then incorporated as categorical features into a Random Forest regression model for OCR prediction. The hybrid model achieved strong generalization performance, with R^2 values of 0.94 and 0.77 on the training and testing sets, respectively. These results demonstrate the advantage of integrating data-driven clustering into OCR prediction, particularly in capturing complex patterns not easily modeled by physical variables alone.

Despite its promising results, this study is limited by the specific set of CPT parameters and the scope of the compiled dataset. Future research could expand the database to include a wider range of soil conditions, regional diversity, and additional field or laboratory parameters. Moreover, sequential prediction approaches using reinforcement learning could be explored to capture the vertical dependency of stratigraphic labels along a CPT profile. Such extensions would further improve the model's adaptability and robustness in real-world site characterization scenarios.

7 ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00210968).

8 REFERENCES

Azzouz, A. S., Baligh, M. M. & Ladd, C. C. 1981. Cone penetration and engineering properties of the soft Orinoco clay.

Baligh, M. M., Azzouz, A. S., Wissa, A. E., Martin, R. T. & Morrison, M. J. 1981. The piezocone penetrometer. OAR (Oceanic and Atmospheric Research): MIT (MIT Sea Grant).

Breiman, L. 2001. Random forests. *Machine Learning,* 45**,** 5-32.

Campanella, R. G. & Robertson, P. K. 1981. *Applied cone research*.

Chan, K. & Poon, B. 2015. Assessment of the coefficient of consolidation for staged preloading operations. *In:* RAMSEY, G., ed. *Proc. the 12th Australia New Zealand Conference on Geomechanics*, Wellington, New Zealand. International Society for Soil Mechanics and Geotechnical Engineering.

Chen, B. S. & Mayne, P. W. 1994. *Profiling the overconsolidation ratio of clays by piezocone tests*, School of Civil and Environmental Engineering, Georgia Institute of Technology.

Feng, Z.-Y., Zhou, J.-W., Yang, X.-G., Tan, L.-J. & Liao, H.-M. 2025. Prediction of landslide dam stability and influencing factors analysis. *Engineering Geology,* 350**,** 108021.

Gwak, D. & Ku, T. 2025. Data-driven machine learning approach for stress history evaluation in cohesive soils using cone penetration test data. *Engineering Geology,* 355**,** 108246.

Houlsby, G. T. & Teh, C. I. 1988. Analysis of the piezocone in clay. *Proc. the Penetration Testing*, Rotterdam. 777-783.

Jamiolkowski, M., Ladd, C. C., Germaine, J. T. & Lancellotta, R. 1985. New developments in field and laboratory testing of soils. *Proc. the 11th International Conference on Soil Mechanics and Foundation Engineering*, San Francisco. 12-16.

Konrad, J. M. & Law, K. T. 1987. Preconsolidation pressure from piezocone tests in marine clays. *Géotechnique,* 37**,** 177-190.

Leonards, G. A. & Frost, J. D. 1988. Settlement of shallow foundations on granular soils. *Journal of Geotechnical Engineering,* 114**,** 791-809.

Ma, X., Liu, Z., Wang, W., Wang, J., Lu, L., Zhou, D. & Zhang, H. 2024. Characteristics of physical parameters and predictive modeling of mechanical properties in loess-like silty clay for engineering geology. *Engineering Geology,* 339**,** 107672.

Mayne, P. W. 1991. Determination of OCR in clays by piezocone tests using cavity expansion and critical state concepts. *Soils and Foundations,* 31**,** 65-76.

Mayne, P. W. & Bachus, R. C. 1988. Profiling OCR in clays by piezocone soundings. *International Symposium on penetration testing*; ISOPT-1. 1, 857-864.

Mayne, P. W. & Holtz, R. D. 1988. Profiling stress history from piezocone soundlngs. *Soils and Foundations,* 28**,** 16-28.

Mcinnes, L., Healy, J. & Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Reynolds, D. 2015. Gaussian Mixture Models. *In:* LI, S. Z. & JAIN, A. K. (eds.) *Encyclopedia of Biometrics*. Boston, MA: Springer US.

Schmertmann, J. H. 1978. Guidelines for cone penetration test : performance and design. United States. Federal Highway Administration.

Schofield, A. N. & Wroth, P. 1968. *Critical state soil mechanics*, McGraw-Hill.

Smits, F. P. 1982. Penetration pore pressure measured with piezometer cones. *Proc. the 2nd European Symposium on Penetration Testing (ESOPT II)*, 871-876.