Geomechanics and Engineering, *Vol. 42, No. 5 (2025) 321-332*
https://doi.org/10.12989/gae.2025.42.5.321

Application of automated machine learning and clustering algorithm for data-driven site characterization: Predicting the soil-rock interface

Dongwoo Lim[1a], Mijin Goo[2b], Han-Saem Kim[3c] and Taeseo Ku*[2]

[1]*Department of Civil, Environmental and Plant Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea, 05029*
[2]*Department of Civil and Environmental Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea, 05029*
[3]*Department of Civil and Environmental Engineering, Dongguk University, 30, Pildong-ro 1-gil, Jung-gu, Seoul, Republic of Korea, 04620*

(*Received June 29, 2025, Revised August 14, 2025, Accepted August 25, 2025*)

Abstract. The development of underground spaces requires detailed insight into subsurface conditions, particularly the soil–rock interfaces, as this information is crucial for the effective design and safe construction of underground infrastructures. Traditional geotechnical site investigations rely mainly on direct drilling and sampling; however, these methods yield data only at specific investigation points, thus posing limitations in comprehensively capturing ground conditions across an entire area. To address this limitation, various studies have aimed to predict unknown subsurface sections using existing borehole data. Conventional methods use geospatial interpolation, while machine learning has emerged as a strong alternative. The selection and proper tuning of an appropriate model are critical to achieving optimal performance. This study applies automated machine learning, focusing on predicting soil-rock interfaces in unsampled regions using borehole data. AutoGluon is used as the machine learning framework to automate data preprocessing, model selection, hyperparameter tuning, and model ensemble. For this study, approximately 20,000 boreholes from the Seoul metropolitan area were collected and employed. Additionally, various digital maps were used to extract input variables. To capture non-linearity among input variables, Uniform Manifold Approximation and Projection were employed to reduce the dimensionality of the dataset, while Hierarchical Density-Based Spatial Clustering of Applications and Noise was implemented as the clustering algorithm. When compared to a model tuned using Bayesian optimization, AutoGluon exhibited superior predictive performance and reduced errors. Furthermore, although the focus of this study is on predicting the soil-rock interface, the methodology can be extended to the prediction of other geotechnical parameters.

Keywords: automated ML; clustering; data-driven; soil-rock interface; spatial prediction

1. Introduction

The development of underground spaces, especially in urban areas, has become increasingly important in recent years. For instance, Singapore integrates underground space development into its two-level urban master plan and special control planning systems to guide 3D geotechnical data-informed practices. China systematically transforms its civil defense-rooted approach to underground space development into comprehensive master and regulatory detailed planning, enhanced by data-driven techniques and smart models. Japan, constrained by strict private land ownership, utilizes legally embedded urban planning systems to focus urban underground space development on public lands through structured master, detailed, and special planning for infrastructure resilience (Peng *et al.* 2023).

*Corresponding author, Associate Professor
E-mail: tsku@konkuk.ac.kr
[a]Graduate Student
[b]Undergraduate Student
[c]Assistant Professor

There is also a growing need for underground development in Seoul, the capital of South Korea. In Korea, the severe rainfall experienced in 2022 led to urban flooding, consequently sparking the commencement of a deep drainage tunnel project. As outlined by the Seoul Urban Master Plan, the city plans to relocate existing transportation systems to subterranean areas. For an effective and safe underground development project, comprehensive understanding of the geotechnical subsurface condition is essential. Inadequate geotechnical investigations can result in over-designs that are un-necessarily costly or under-designs that can lead to potential problems (Zumrawi 2014). Also, Jaksa (2000) discussed a case encompassing several projects where inadequate geotechnical investigation in a highly variable soil profile led to a foundation failure, causing substantial cost overruns and a one-month delay.

In traditional site investigation, boring and penetration tests are extensively performed. However, these tests exhibit limitations as they offer information only at specific investigation points, thereby making it challenging to assess ground conditions comprehensively across a broader area. To address this problem, various site investigation approaches have been explored, such as geophysical

ISSN: 2005-307X (Print), 2092-6219 (Online)

methods (Kim *et al*. 2013, Moon *et al*. 2019) for estimating the mechanical properties of soils or the depth of bedrock, and remote sensing techniques for assessing settlement (Park *et al.* 2024, Nur *et al*. 2025) or landslide map (Lim *et al.* 2024). The recent increase in geotechnical survey data has elicited an amplified interest in the application of data-driven methodologies for site characterization techniques (Phoon *et al.* 2022). In this study, we focus on data driven site characterization methods. There are two primary data-driven approaches for characterizing subsurface conditions. The first approach is generating a prediction model based on geospatial methods (Lee *et al*. 2022, Ijaz *et al*. 2023) and/or statistical methods such as Markov chain, random field theory, and Bayesian inference (Elfeki and Dekking 2001, Li *et al*. 2016, Qi *et al.* 2016, Wang *et al.* 2018, Qi *et al.* 2020, Deng *et al.* 2020, Wang *et al.* 2020, Gong *et al.* 2021). The second approach is machine learning (ML)-based methods. Recent studies have explored various ML approaches – including multi-layer perceptron (Kim and Ji 2022), meta-learning (Wang *et al.* 2023), Gaussian process (Deng *et al.* 2023) and Bayesian inference (Yang *et al.* 2023) – and have shown their applicability in site characterization. Also, there are many studies exploring the application of ML approaches to geotechnical engineering, such as the prediction of soilcrete strength (Al-Shamasneh *et al.* 2025), the analysis of interfacial behavior between frozen soil and structures (Park *et al.* 2025), and the interpretation of geophysical survey results (Vantassel *et al.* 2022, Ryu *et al.* 2025).

Nevertheless, a considerable number of problems encountered in the field of geotechnical engineering are non-linear in nature, making it impractical for a single ML algorithm to handle every dataset effectively. As a result, various ML algorithms are generally employed to develop an appropriate model. The modeling process is knowledge-based and time-consuming (Zhang *et al.* 2022). Model selection and hyperparameter optimization continue to pose challenges for ML-based modeling (Ma *et al.* 2022). This process requires expertise in ML algorithms, however, the need for such expertise can be mitigated by utilizing automated ML platforms (Jiang *et al.* 2022). Automated ML has seen limited but growing application in geotechnical engineering, including the evaluation of landslide susceptibility, tunnel displacement, ground settlement, slope stability classification, and soil liquefaction potential (Bruzón *et al*. 2021, Zhang *et al.* 2022, Hussaine and Mu 2022, Ma *et al.* 2022, Sahin and Demir 2023). However, it is worth noting that its application in addressing geotechnical challenges related to site investigation has rarely been reported.

As aforementioned, proper data processing and feature selection are essential to improve ML performance and ensure reliable predictions. To address this problem, various techniques can be utilized, such as dimensionality reduction and clustering algorithms. Dimensionality reduction transforms high-dimensional data into a lower-dimensional format to prevent the issues associated with the curse of dimensionality. Clustering is a form of unsupervised learning that organizes data according to their similarities, enabling the discovery of patterns and relationships within the dataset. In geotechnical engineering, clustering has been effectively utilized in selected areas, such as soil classification based on in-situ test results (Hudson *et al.* 2023) and the evaluation of ground hazards (Shen *et al.* 2023, Mwakapesa *et al.* 2023). Furthermore, the combined use of dimensionality reduction and clustering for rock classification and data-driven site characterization has been reported elsewhere (Wang and Zhang 2023, Hansen and Aarset 2024). However, in the above applications, clustering itself was typically the primary goal, or it was used to divide the data into groups for building group-specific models with limited applicability. More recently, Gutman *et al.* (2025) demonstrated enhanced prediction accuracy by incorporating cluster IDs as supplementary features while preserving the original data attributes. Nevertheless, this approach remains relatively underexplored within geotechnical engineering.

Accordingly, this study aims to explore the application of automated machine learning to data-driven site characterization, focusing on predicting the soil-rock interface in unsampled areas by leveraging existing boring survey data. AutoGluon, an open-source automated machine learning (AutoML) framework developed by Amazon Web Services, is used as a robust platform for automating data preprocessing, model selection, hyperparameter tuning, and model ensembling. Furthermore, to discern interrelations among input variables, dimensionality reduction and clustering techniques are employed concurrently. This methodology not only improves predictive performance but also simplifies the modeling process, thereby rendering it more approachable for professionals within the field of geotechnical engineering. The results of this study have the potential to advance geotechnical practice by offering a comprehensive framework for data-driven site characterization and enhancing the understanding of subsurface conditions.

2. Automated machine learning

Selecting an appropriate machine learning algorithm depends on the characteristics of the dataset and the purpose of data analysis. Therefore, various machine learning algorithms are generally employed to create a suitable model. However, the modeling process requires hyperparameter tuning and training for each algorithm, which is time-consuming. In recent years, attempts to automate the machine learning process have led to the development of AutoML as a solution to this issue.

AutoML automates machine learning processes such as hyperparameter tuning and model selection. A variety of Python libraries provide AutoML capabilities. Ferreira *et al.* (2021) conducted a comparative analysis of AutoML libraries for tasks including regression, binary classification, and multi-class classification, under three distinct machine learning scenarios: general machine learning, deep learning, and XGBoost. In the context of regression tasks within the general machine learning framework, AutoGluon, H2O, and TransmogrifAI exhibited the fastest performance while achieving comparable outcomes. In this study, AutoGluon is

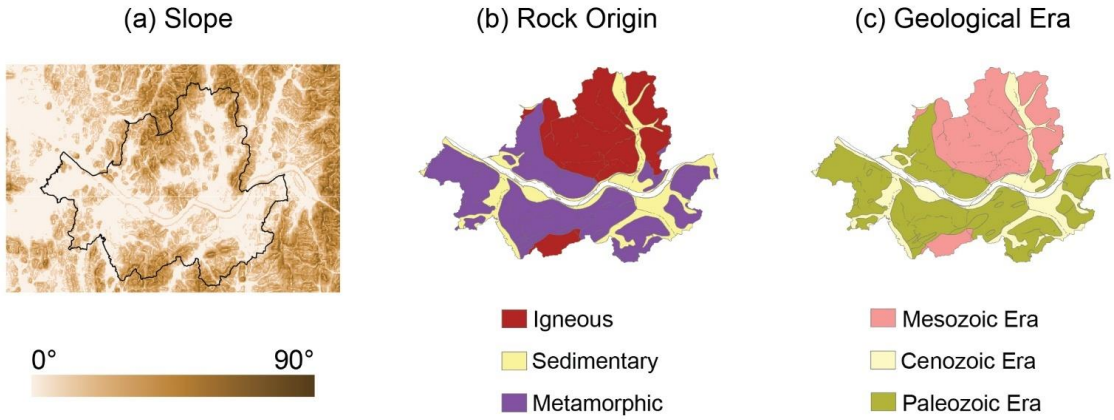

Fig. 1 Topographic and geological maps of Seoul area: (a) Slope, (b) Rock origin and (c) Geological era

utilized due to its computational efficiency and capability in multi-layer stacking.

AutoGluon-Tabular (Erickson *et al.* 2020) differs significantly from other AutoML frameworks that primarily automate the Combined Algorithm Selection and Hyperparameter (CASH) tuning procedure, as it provides a fully end-to-end ML pipeline. In addition, AutoGluon offers strong capabilities in data preprocessing and modern neural network architecture and employs advanced ensemble methods that leverage multi-layer stacking combined with repeated k-fold bagging.

Table 1 Digital maps used in this study and their sources

Data Type	Source
1:250,000 Digital Geological Map	Geo Big Data Open Platform[a)]
Digital Elevation Model (90 m)	V-world[b)]
Transportation Facilities Distribution	V-world[b)]
Borehole Data	Geotechnical Information DB System[c)]

a) Open data platform operated by Korea Institute of Geoscience and Mineral Resources
b) Open spatial data platform operated by Korea Ministry of Land, Infrastructure and Transport
c) Database operated by Korea Institute of Civil Engineering and Building Technology

3. Spatial properties and borehole dataset

3.1 Geological and topographic features

The topographic and geological features of Seoul are analyzed using publicly accessible digital maps with Geographic Information System. The digital maps employed in the analysis, along with their sources, are summarized in Table 1.

Seoul is a basin surrounded by mountains, with the Han River running through the center of the city. Fig. 1(a) presents an analysis of the topographical slope in Seoul. It reveals steeper slopes in the north-south part due to mountainous terrain, and gentler slopes around the central region where the Han River is located. Fig. 1(b) shows the spatial distribution of rocks classified by their formation origin. Seoul's foundation consists of metamorphic rocks (45.7%), igneous rocks (32.27%), and sedimentary rocks (17.94%). The northern areas predominantly feature igneous rocks, the southern parts mainly host metamorphic rocks, and sedimentary rocks are chiefly located around the Han River. Fig. 1(c) illustrates the distribution of geological rock types, demonstrating a strong correlation between geological periods and the origins of formation.

3.2 Borehole dataset

The dataset used in this study contains a total of 20,393 borehole data points. The results of hotspot analysis for borehole data are presented in Fig. 2(a), along with the distribution of transportation infrastructure. Blue dots indicate cold spots, representing boreholes with relatively shallow depth, while red dots indicate hot spots, corresponding to boreholes with relatively greater depths. Cold spots are primarily distributed along the north-south axis of Seoul, whereas hot spots are mainly concentrated near the Han River. It also reveals that borehole data are predominantly distributed along transportation networks. Fig. 2(b) illustrates the depth distribution categorized by geological formation origins. It indicates that the igneous rock region possesses the shallowest soil-rock interface, succeeded by the metamorphic rock region, with the sedimentary rock region exhibiting greater depths.

4. Data preprocessing

To develop a machine learning model for predicting soil-rock interface, this study employed not only the spatial coordinates (x, y) but also surface-accessible variables as input features. Recent studies (Zhu *et al*. 2021, Wang *et al.* 2023) utilized topographic features such as elevation, slope, aspect and curvature to predict unknown region's soil-rock interface. Predicting the soil-rock interface equates to predicting soil layer thickness, driven by soil erosion, transport, and decomposition. Topographic features

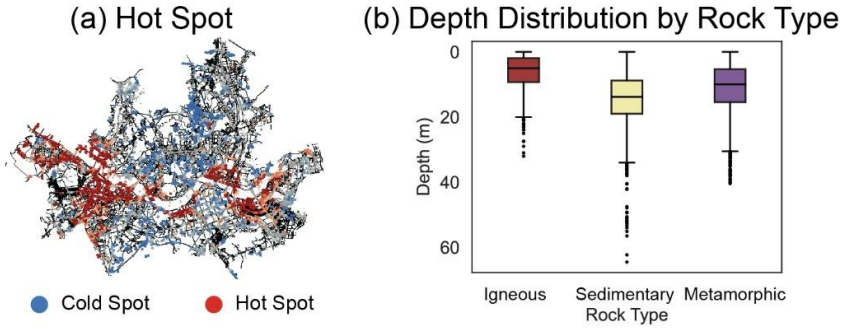

Fig. 2 Spatial and statistical distribution of soil-rock interface depth: (a) Spatial clustering of soil-rock interface depth (hotspots: deeper areas, cold spots: shallower areas) overlaid with transportation infrastructure and (b) Box plot of rock depth categorized by rock origin

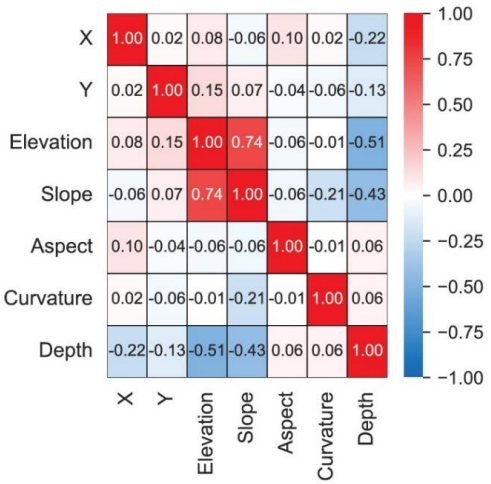

Fig. 3 Matrix of Pearson correlation coefficients between variables

considered in this study effectively represent these processes and serve as sensible input parameters, as they can be easily obtained from processing DEM data for specific locations. However, since the recent studies were conducted in the geological context of Singapore, their findings may not be directly transferable to the geological and topographical conditions of Korea. Therefore, to determine appropriate input features, Pearson's correlation analysis was conducted.

The variables used in the analysis include elevation, slope, aspect and curvature, and the heatmap of the covariance matrix is presented in Fig. 3. As shown in Fig. 3, elevation and slope exhibit a moderate negative correlation with depth, while curvature and aspect show little to no significant correlation. Therefore, elevation and slope were readily selected as input variables. Furthermore, since the relationship between the origin of geological formations and depth was confirmed in an earlier section, it was also included as one of the input variables. Although curvature and aspect did not demonstrate strong linear correlations in the Pearson analysis, they were nonetheless incorporated into the model to allow for sub-sequent evaluation of their effect through feature importance analysis, considering that machine learning algorithms may capture non-linear or interaction effects.

In the training of ML models, a high-dimensional input dataset can cause the curse of dimensionality, which raises computational costs and the risk of overfitting. To mitigate this problem, principal component analysis was conducted for dimensionality reduction. This approach enhances not only computational efficiency but also the overall performance of the model. In this study, the focus is on enhancing prediction accuracy by incorporating a new feature that captures the non-linear relationships among data attributes, achieved through the integration of dimensionality reduction and clustering algorithms. The algorithms utilized in this research will be introduced in the following section.

4.1 Uniform manifold approximation and projection

Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al.* 2018) is a nonlinear dimensional reduction algorithm that preserves the underlying structure of data and efficiently represents relationships in high-dimensional spaces based on the concepts of Riemann geometry and algebraic topology. UMAP approximates a low-dimensional manifold representing the data, and this manifold can be understood as a curved surface within a high-dimensional space. The algorithm constructs a fuzzy phase structure based on the assumption that data points are uniformly distributed in this manifold, and this is implemented in a way that includes ambiguity or uncertainty in connectivity between points. UMAP offers superior preservation of the overall structure compared to existing techniques such as t-Distributed Stochastic Neighbor Embedding, which is advantageous for understanding the overall shape of the dataset. Also, due to its high reproducibility, there are few changes in the embedding structure even when new samples are added, and stable application is possible when expanding the model. Thus, UMAP was adopted as a dimensional reduction technique in this study.

Throughout the study, Bayesian optimization was consistently employed for all parameter search tasks, including the identification of optimal UMAP parameters. This technique is based on Bayesian inference, using a surrogate model and an acquisition function to efficiently find the next point to explore. It outperforms conventional

Table 2 Bayesian optimization results for UMAP

		Axis = 2	Axis = 3
Metric	Trustworthiness	0.9804	0.9931
	KNN preservation	0.6406	0.6978
Parameter	Number of neighbors	14	12
	Minimum distance	0.208	0.279
	Distance metric	Euclidean	Euclidean

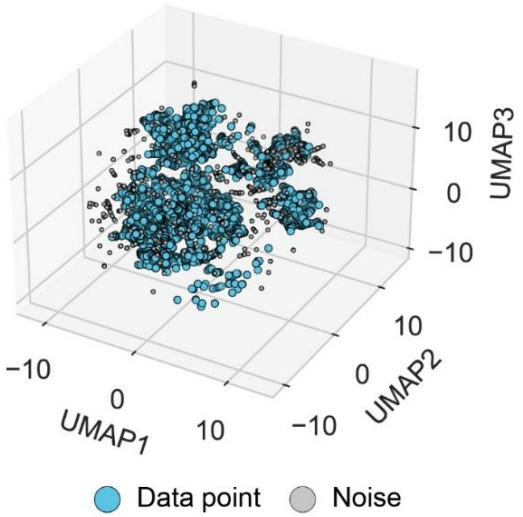

Fig. 4 Optimal UMAP embedding obtained through Bayesian optimization

methods such as grid search or random search by incorporating prior knowledge into the hyperparameter selection (Choi *et al*. 2020). The objective function was designed to optimize a combination of metrics assessing the structural integrity of the embedding. These metrics include trustworthiness, which penalizes incorrect neighbors in the lower-dimensional space, and KNN preservation, which evaluates the retention of neighbor information from high-dimensional spaces in lower-dimensional embeddings (Huang *et al.* 2022). The parameters optimized were the number of neighbors and the minimum distance. The results for the indicators and parameters with respect to the number of axes are presented in Table 2, indicating that 3-axis UMAP demonstrates superior performance. Consequently, this research employs 3D UMAP. The resulting UMAP is presented in Fig. 4. There is clear separation between data points, which indicates that clustering was performed properly.

4.2 Hierarchical density-based spatial clustering of applications with noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester *et al.* 1996) is a data clustering method that, unlike K-means clustering which requires a predetermined number (K) of clusters, utilizes a density-based nonparametric strategy. This method organizes closely grouped data points into clusters, while points found in low-density areas are categorized as outliers (noise). Nevertheless, because DBSCAN relies on a single density threshold to define clusters, its detection capabilities are restricted in datasets that exhibit varying densities or overlapping cluster formations. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello *et al.* 2013) addresses this shortcoming by allowing for the effective identification of clusters with varying densities through the establishment of a cluster hierarchy at multiple density levels. This facilitates a more advanced analysis of data structures characterized by overlapping clusters or gradual changes in density. In this study, given that the spatial arrangement of boreholes is irregular and displays density variations, HDBSCAN's adaptable density-based clustering method was deemed suitable and implemented.

Table 3 Bayesian optimization results for HDBSCAN

Metric	Probabilities	0.835
	Persistence	0.256
Parameter	Minimum cluster size	116
	Minimum samples	15

To search for the optimal HDBSCAN parameters, Bayesian optimization was again employed to ensure consistency and efficiency in parameter selection. HDBSACN's indexes are probabilities and persistence. Probabilities represent the likelihood of each data point belonging to a cluster and serve as indicators of the point's stability within that cluster. Persistence indicates how long each cluster remains stable across varying density levels and is used to measure the relative coherence among clusters. The objective function was defined as the average of these two metrics. This aims to find clusters that are both structurally stable and have clear member assignments. While this may not be a theoretically perfect solution, it represents a practical choice that aligns with the objectives of the study. The adjusted parameters include the minimum cluster size, which refers to the least number of data points required in a cluster, and the minimum samples, which indicate the number of neighboring points required to identify a key point. The parameters and metrics obtained through Bayesian optimization are presented in Table 3.

Fig. 5 shows the results of applying HDBSCAN in the space where 3D UMAP was performed. A total of 17 groups were formed, and it can be observed that the clusters are properly separated. Additionally, to further validate the clustering, the properties of each cluster were analyzed, with the results shown in Fig. 6. Fig. 6(a) illustrates the distribution of each cluster concerning soil-rock interface depth. It can be observed that each cluster possesses a distinctly different average. Furthermore, since cluster 13 exhibits the lowest average depth, it can be inferred that it corresponds to a mountainous region, while cluster 5, having the deepest average depth, is anticipated to represent a sedimentary rock area. Fig. 6(b) depicts the distribution of each cluster in relation to elevation. Considering the elevated positions of clusters 13 and 14, it is likely that mountainous areas are included, aligning with the earlier

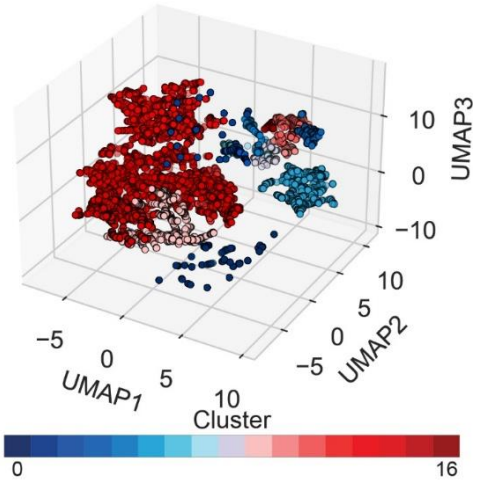

Fig. 5 Optimal HDBSCAN clustering obtained through Bayesian optimization

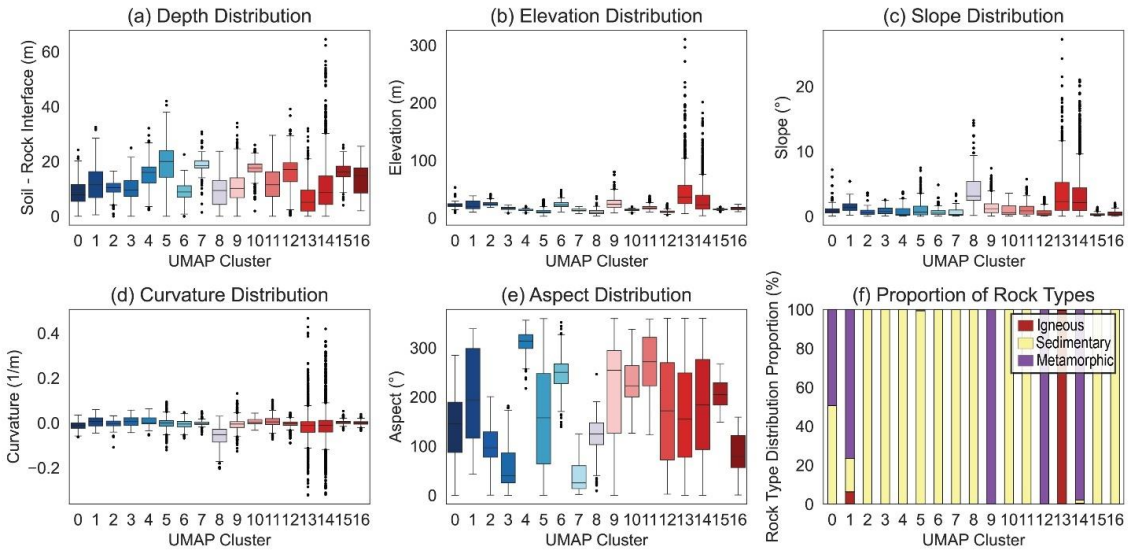

Fig. 6 UMAP cluster vs Properties: (a) Soil-rock interface depth, (b) Elevation, (c) Slope, (d) Curvature, (e) Aspect and (f) Rock type

analysis. Fig. 6(c) presents the distribution of clusters based on slopes. Similar to the analysis in Fig. 6(b), the slopes of clusters 13 and 14 are significant, suggesting the inclusion of mountainous regions. In contrast, cluster 8, despite its lower altitude, is expected to be a sedimentary terrain due to its steep slope. Fig. 6(d) provides an analysis of curvature. A positive curvature indicates a convex shape, whereas a negative curvature denotes concavity. As noted in the previous analysis, clusters 13 and 14 are characterized by a combination of convex and concave terrain, suggesting a complex mountainous area. In the case of cluster 8, it is assessed to be concave, suggesting favorable conditions for sedimentation. Fig. 6(e) represents an aspect, but it poses challenges for analysis, making it difficult to interpret its orientation or derive definitive insights from it. Fig. 6(f) illustrates the distribution of rocks within each cluster. This analysis confirms the prior conclusions, namely, that clusters 5 and 8 indicate a sedimentary area, while clusters 13 and 14 indicate a mountainous area. In particular, the considerable average depth in cluster 14 is attributed to the depth characteristics of the mixed sedimentary rock area. To validate this estimation, the spatial distributions of clusters 5, 8, 13, and 14 were examined, as shown in Figure 7.

Figs. 7(a)-7(d) show the spatial distributions of clusters 5, 8, 13, and 14, respectively. From these distributions, it was also confirmed that cluster 5 is located in a flat sedimentary area, while cluster 8 corresponds to a sedimentary area of the Han River around the mountainous regions, where significant sedimentation occurs. In addition, cluster 13 is exclusively distributed in igneous rock and mountainous areas, and cluster 14 is mainly located in metamorphic rock mountainous areas. However, some sedimentary rocks were also found to be distributed around the Han River. This confirms that the clustering is effective in capturing relevant property characteristics.

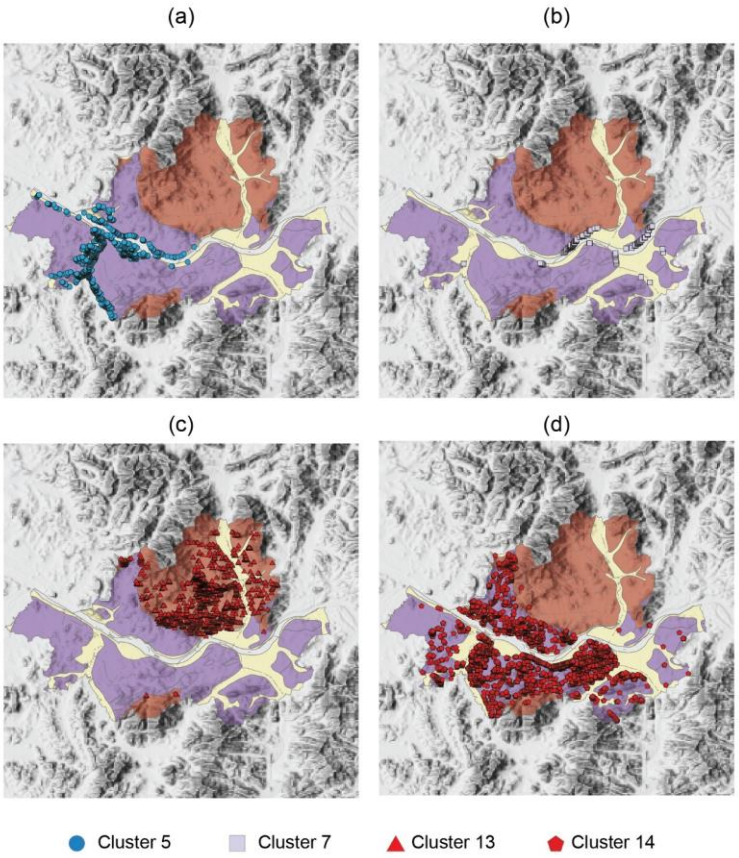

Fig. 7 Distribution of chosen clusters across spaces: (a) Cluster 5 is mostly found in low-altitude sedimentary areas, (b) Cluster 8 is mainly situated along riverine settings, (c) Cluster 13 is chiefly centralized in igneous highland regions and (d) Cluster 14 covers metamorphic landscapes and stretches into nearby sedimentary regions

4.3 Outlier detection

According to Phoon and Kulhawy (1999), a failure to properly address outliers can lead to a significant misinterpretation of data. To obtain robust prediction model, it is crucial to detect and eliminate outliers (Kim and Ji 2022). In this study, outliers are detected by jointly considering Local Moran's I and clustering results used in the previous chapter. Local Moran's I is a geographical autocorrelation statistic method, which classifies clusters and outliers by comparing each point to its neighbors. Additionally, z-score and p-value are used to assess the statistical significance test. This study used a local distance of 1,798 m to ensure that each data point included at least three neighbors. The clustering results obtained using HDBSCAN not only group similar elements together but also identify data points that do not belong to any cluster as noise. Subsequently, data points identified as high-low and low-high clusters with a p-value below 0.05 by Local Moran's I, or as noise by HDBSCAN, were classified as outliers. A total of 5,586 such data points were removed.

4.4 Data split

Earlier research (Gosciewski 2013, Amoroso *et al.* 2023) indicated that the density of spatial data influences the effectiveness of spatial interpolation. Consequently, to mitigate this issue, it is advisable to split data randomly rather than assigning the density of each point based on the number of neighboring points within a specified location distance 1,798 m, as described in the previous section. Utilizing these density measurements, the dataset was segmented into 30 distinct density bins. Each bin was subsequently partitioned into subsets for training (70%), validation (10%), and testing (20%). Using 30 bins offers maximum granularity for better data representation while ensuring each bin has enough data for the specified training, validation, and testing ratios

5. Results and discussion

This study aims to address two main objectives: First, it seeks to improve predictive performance in data-driven site characterization using machine learning by explicitly encoding spatial or structural relationships through the inclusions of cluster IDs—derived via nonlinear dimensionality reduction and density-based clustering—as additional input features without information loss. Second, the study explores the ability of automated machine learning algorithms to overcome the challenges posed by the CASH process, which often hinders the effective

Table 4 List of input variables

Feature	Description	Type
X	Borehole Latitude Coordinate (EPSG 5186)	Numerical
Y	Borehole Longitude Coordinate (EPSG 5186)	Numerical
DEM	Digital Elevation Model, elevation at the borehole location	Numerical
Slope	Surface slope at borehole location, computed from DEM using the Slope tool in ArcGIS pro	Numerical
Aspect	Surface aspect at borehole location, computed from DEM using the Aspect tool in ArcGIS pro	Numerical
Curvature	Surface curvature at borehole location, computed from DEM using the Curvature tool in ArcGIS pro	Numerical
Rock type	Rock type based on geological map	Categorical
Cluster ID	Cluster ID derived from HDBSCAN with UMAP	Categorical

application of machine learning. Accordingly, the result is separated into two domains: the validation of clustering-based feature augmentation, and the evaluation of AutoML algorithm applicability. For the manual tuning, Bayesian optimization was employed. The characteristics of the input variables are summarized in Table 4.

5.1 Validation of clustering-based feature augmentation

To evaluate the influence of cluster granularity on predictive performance, the initial 17 cluster IDs derived from a HDBSCAN were used as categorical input features in regression models. Given that a large number of categorical values may increase feature dimensionality and risk overfitting, the original clusters were hierarchically grouped into 5 and 10 broader categories based on inter-cluster similarity. Figure 8 illustrates the hierarchical structure of the initial clusters and the basis for forming simplified groupings. This shows that the clusters were merged in a systematic manner. These simplified cluster IDs were encoded and included as additional input features. For comparative experiments, four input combinations were constructed. Combination A included all input parameters listed in Table 4, excluding the cluster ID. Combination B extended Combination A by incorporating the original 17 cluster IDs as categorical features. Combination C added the 5 group merged cluster IDs to Combination A, while Combination D added the 10 group merged IDs. To assess how well regression models perform, two metrics were used: Root Mean Square Error (RMSE) to assess the prediction error, and the coefficient of determination (R^2) to indicate the fit quality of models. RMSE is the square root of the average of the squared differences between predicted and actual values; lower RMSE values indicate less prediction error. R^2 shows the percentage of variation in the observed data that is explained by the model's predictions; higher values indicate better predictive accuracy. Then, for each combination, five regression models were tuned using Bayesian optimization, and the learning results were presented in Table 5.

Table 5 Prediction accuracy of manually tuned machine learning models with and without clustering ID

ML models		A	B	C	D
CatBoost	RMSE	3.56	3.11	3.39	3.33
	R^2	0.78	0.83	0.80	0.81
ExtraTrees	RMSE	2.83	2.88	2.87	2.87
	R^2	0.86	0.86	0.86	0.86
LightGBM	RMSE	3.80	3.92	3.85	3.76
	R^2	0.75	0.73	0.74	0.76
Random Forest	RMSE	3.91	3.90	3.66	4.13
	R^2	0.74	0.74	0.77	0.71
XGBoost	RMSE	3.48	3.79	3.52	3.39
	R^2	0.79	0.75	0.79	0.80

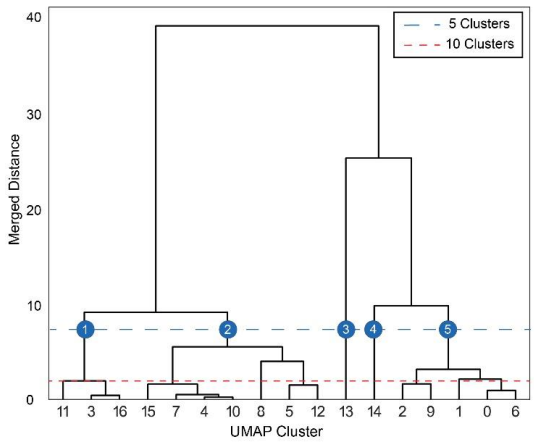

Fig. 8 Hierarchical structure of the 17 clusters obtained from HDBSCAN

According to Table 5, selected combinations with cluster IDs provided improved results, except for the ExtraTrees model. For the RandomForest model, combination C achieved the highest performance. In the case of XGBoost and LightGBM, combination D led to the best performance, while for CatBoost, combination B provided the top results. On average, combination C showed promising performance across all models.

These results indicate that incorporating the cluster ID as input features can enhance model performance. This suggests that the insights gained from clustering reveal important structural relationships that the original input features do not capture on their own. Notably, merged cluster combinations (combination C, D) frequently outperformed the original cluster (combination B). This suggests that decreasing cluster granularity can mitigate model complexity and enhance generalization. Nonetheless, in certain instances, this simplification resulted in the elimination of crucial detailed information, leading to degraded performance. This highlights that the ideal level of cluster granularity is dependent on the specific model,

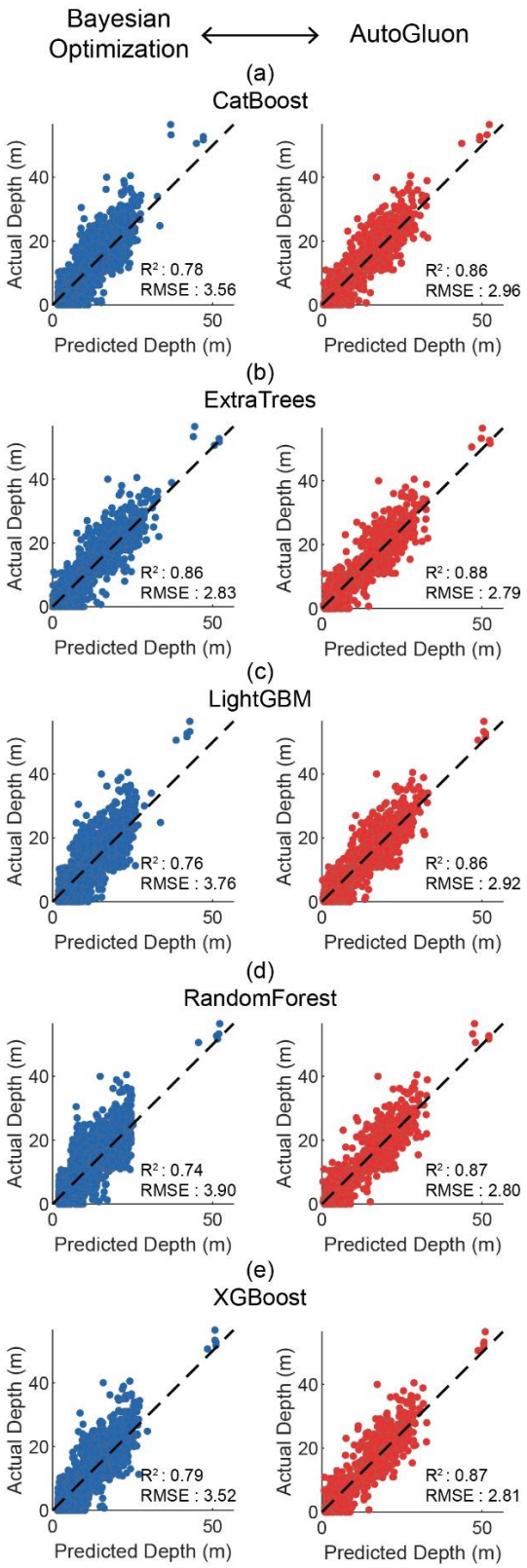

Fig. 9 Comparison of prediction performance between Bayesian optimization (manual tuning) and AutoGluon. For each model, a 1:1 plot is presented, using the best-performing input feature combination selected by AutoGluon: (a) CatBoost (combination A), (b) ExtraTrees (combination A), (c) LightGBM (combination D), (d) RandomForest (combination B), and (e) XGBoost (combination C)

Table 6 Prediction accuracy of AutoGluon models with and without clustering ID

		A	B	C	D
CatBoost	RMSE	2.96	3.17	3.17	3.17
	R^2	0.86	0.84	0.84	0.84
ExtraTrees	RMSE	2.79	2.80	2.82	2.81
	R^2	0.88	0.87	0.87	0.87
LightGBM	RMSE	2.93	2.93	2.92	2.92
	R^2	0.86	0.86	0.86	0.86
NeuralNet FastAI	RMSE	3.52	3.64	2.79	3.91
	R^2	0.80	0.79	0.88	0.75
NeuralNet Torch	RMSE	2.79	2.79	2.79	2.78
	R^2	0.88	0.88	0.88	0.88
Random Forest	RMSE	2.99	2.80	2.82	2.82
	R^2	0.86	0.87	0.87	0.87
Weighted Ensemble	RMSE	2.77	2.78	2.79	2.78
	R^2	0.88	0.88	0.88	0.88
XGBoost	RMSE	2.91	2.96	2.81	3.01
	R^2	0.86	0.86	0.87	0.85

and that well-constructed clustering-based features can significantly enhance the input space.

5.2 Evaluation of automated machine learning applicability

To evaluate the applicability of Automated ML, only AutoGluon was applied under the same experimental conditions as previously used for comparison. For this study, the training time was set to 24 hours, and the computing environment for the experiment was an Intel Core i7-14700KF CPU, an NVIDIA RTX 4070 Ti GPU, and 32GB of RAM. Fig. 9 compares manually tuned models with those tuned by AutoGluon. For each model, the comparison is based on the performing input configuration selected by AutoGluon, and results are visualized using a 1:1 plot. The plot demonstrates that AutoGluon achieves superior predictive accuracy with minimal variance, notably showing fewer outliers and delivering consistent performance across different test algorithms. Additionally, Table 6 provides an extensive summary of the performance outcomes from various learning models and combinations of input features. Compared to Table 5, AutoGluon clearly improves the predictive performance for most models and input combinations. These findings imply that AutoGluon's training approach, especially for tree-based models, effectively captures the inherent nonlinearity of the data without requiring explicit prior clustering. Notably, the nonlinear structures introduced by UMAP and HDBSCAN might have added redundancy or noise, which degraded performance in certain models.

Nonetheless, when evaluating both tree-based and neural network models, combination C exhibits the best overall performance. Notably, when applying the input

configuration that comprises 5 merged cluster IDs (Combination C), neural network-based models (NeuralNetFastAI and Neural-NetTorch) exhibited particularly strong performance; however, the performance diminished when the number of cluster IDs increased to 10 and 17 (Combinations D and B, respectively). This implies that an excessive quantity of categorical cluster features may result in overfitting neural network models, which are more susceptible to high-dimensional or sparse categorical inputs. These results suggest that, although cluster-based features can improve the model performance, their effectiveness is dependent on the model architecture and the degree of granularity. Specifically, neural networks may benefit from simplified clustering strategies to reduce the risk of overfitting. While models manually tuned using Bayesian optimization displayed significant performance fluctuations depending on the inclusion and configuration of cluster IDs, the models developed by AutoGluon remained relatively stable.

These findings emphasize AutoGluon's capability to capture inherent spatial and structural relationships through automated CASH and ensembling. As a result, the extra structural information encoded in cluster IDs may become redundant by AutoGluon's integrated modeling capabilities.

6. Conclusions

In geotechnical engineering, the effective application of machine learning is often hindered by practitioners' limited expertise in model selection and hyperparameter tuning. To address this challenge, this study investigated AutoML's capabilities for data-driven site characterization, with a focus on predicting the soil-rock interface. To tackle the inherent non-linearity of geotechnical problems, our approach utilized non-linear dimensionality reduction (UMAP) and clustering techniques (HDBSCAN), along with Local Moran's I for identifying spatial outliers. The AutoGluon framework demonstrated superior predictive performance compared to a model manually optimized via Bayesian methods, delivering higher accuracy with reduced error. A significant finding was that while adding cluster-based input features enhanced the manually tuned models, they provided little benefit to AutoGluon, suggesting the framework can intrinsically capture complex relational patterns without explicit feature engineering. Although automated process required longer computational time, our findings confirm that AutoML enables geotechnical experts with limited ML knowledge to construct robust predictive models, highlighting its adaptability and potential for broad application across various geotechnical challenges.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (Project Number RS-2023-00210968). We acknowledge the borehole data were obtained from Geotechnical Information DB System (Korea Institute of Civil Engineering and Building Technology).

References

Al-Shamasneh, A.R., Mahmoodzadeh, A., Ghazouani, N. and Ouni, M.H.E. (2025), "Forecasting mechanical properties of soilcrete enhanced with metakaolin employing diverse machine learning algorithms", *Geomech. Eng.*, **40**(2), 123-137. https://doi.org/10.12989/gae.2025.40.2.123.

Amoroso, P.P., Falchi, U., Figliomeni, F.G. and Vallario, A. (2023), "The influence of interpolation methods and point density on the accuracy of a bathymetric model", *Proceedings of the 2023 IEEE international workshop on metrology for the sea; learning to measure sea health parameters (MetroSea)*.

Bruzón, A.G., Arrogante-Funes, P., Arrogante-Funes, F., Martín-González, F., Novillo, C.J., Fernández, R.R., Vázquez-Jiménez, R., Alarcón-Paredes, A., Alonso-Silverio, G.A., Cantu-Ramirez, C.A. and Ramos-Bernal, R.N. (2021), "Landslide susceptibility assessment using an autoML framework", *IJERPH*, **18**(20), 10971. https://doi.org/10.3390/ijerph182010971.

Campello, R.J.G.B., Moulavi, D. and Sander, J. (2013), "Density-based clustering based on hierarchical density estimates", *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, (Eds., Pei, J., Tseng, V.S., Cao, L., Motoda, H. and Xu, G.), 160-172. Berlin, Heidelberg: Springer Berlin Heidelberg.

Choi, Y., Yoon, D., Choi, J. and Byun, J. (2020), "Hyperparameter search for facies classification with bayesian optimization", *Geophys. Geophys. Explor.*, **23**(3), 157-167. https://doi.org/10.7582/GGE.2020.23.3.00157.

Deng, Z.P., Jiang, S.H., Niu, J.T., Pan, M. and Liu, L.L. (2020), "Stratigraphic uncertainty characterization using generalized coupled Markov chain." *Bull Eng Geol Environ.*, **79**(10), 5061-5078. https://doi.org/10.1007/s10064-020-01883-y.

Deng, Z.P., Pan, M., Niu, J.T., Jiang, S.H., Wu, B. and Li, S. (2023), "Spatial prediction of rockhead profile using the Gaussian process regression method", *Can. Geotech. J.*, **60**(12), 1849-1860. NRC Research Press. https://doi.org/10.1139/cgj-2022-0372.

Elfeki, A. and Dekking., M. (2001), "A Markov chain model for subsurface characterization: theory and applications", *Math. Gool.*, **33**, 569-589.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M. and Smola, A. (2020), "Autogluon-tabular: Robust and accurate autoML for structured data", arXiv.

Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996), "A density-based algorithm for discovering clusters in large spatial databases with noise", *kdd*, 226-231.

Ferreira, L., Pilastri, A., Martins, C.M., Pires, P.M. and Cortez, P. (2021), "A comparison of AutoML tools for machine learning, deep learning and xgboost", *Proceedings of the 2021 international joint conference on neural networks (IJCNN)*, 1-8.

Gong, W., Zhao, C., Juang, C.H., Zhang, Y., Tang, H. and Lu, Y. (2021), "Coupled characterization of stratigraphic and geo-properties uncertainties – A conditional random field approach", *Eng. Geol.*, **294**, 106348. https://doi.org/10.1016/j.enggeo.2021.106348.

Gosciewski, D. (2013), "The effect of the distribution of measurement points around the node on the accuracy of interpolation of the digital terrain model", *J. Geograph. Syst.*, **15**(4), 513-535. https://doi.org/10.1007/s10109-012-0176-x.

Gutman, D., Perel, N., Bărbulescu, O. and Koren, O. (2025), "A hybrid dimensionality reduction procedure integrating clustering with KNN-based feature selection for unsupervised data", *Algorithms*, **18**(4), 188. https://doi.org/10.3390/a18040188.

Hansen, T.F. and Aarset, A. (2024), "Unsupervised machine learning for data-driven rock mass classification: Addressing limitations in existing systems using drilling data", *Rock Mech.*

Rock Eng., https://doi.org/10.1007/s00603-024-04280-z.

Huang, H., Wang, Y., Rudin, C. and Browne, E.P. (2022), "Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization", *Commun. Biol.*, **5**(1), 719. https://doi.org/10.1038/s42003-022-03628-x.

Hudson, K.S., Ulmer, K.J., Zimmaro, P., Kramer, S.L., Stewart, J.P. and Brandenberg, S.J. (2023), "Unsupervised machine learning for detecting soil layer boundaries from cone penetration test data", *Earthq. Eng. Struct. D.*, **52**(11), 3201-3215. https://doi.org/10.1002/eqe.3961.

Hussaine, S.M. and Mu, L. (2022), "Intelligent prediction of maximum ground settlement induced by EPB shield tunneling using automated machine learning techniques", *Mathematics*, **10**(24), 4637. https://doi.org/10.3390/math10244637.

Ijaz, Z., Zhao, C., Ijaz, N., Rehman, Z.U. and Ijaz, A. (2023), "Statistical evaluation of multiple interpolation techniques for spatial mapping of highly variable geotechnical facets of soil in natural deposition", *Earth Sci. Inform.*, **16**(1), 105-129. https://doi.org/10.1007/s12145-022-00924-2.

Jaksa, M. (2000), "Geotechnical risk and inadequate site investigations: a case study", *Australian Geomech.*, **35**(2), 39-46.

Jiang, S., Ma, J., Liu, Z. and Guo, H. (2022), "Scientometric analysis of Artificial Intelligence (AI) for geohazard research", *Sensors*, **22**(20), 7814. https://doi.org/10.3390/s22207814.

Kim, D.S., Youn, J.U. and Park, H.J. (2013), "Session report: Applications of shear wave velocity on various geotechnical problems", *Geotechnical and Geophysical Site Characterization : Proceedings of the 4th International Conference on Site Characterization ISC-4*, 661-673. Boca Raton, United Kingdom: Taylor & Francis Books Ltd.

Kim, H.S. and Ji, Y. (2022), "Three-dimensional geotechnical-layer mapping in Seoul using borehole database and deep neural network-based model", *Eng. Geol.*, **297**, 106489. https://doi.org/10.1016/j.enggeo.2021.106489.

Lee, S.Y., Song, K.I., Kang, K.N., Kim, W. and An, J.S. (2022), "Applicability analysis of measurement data classification and spatial interpolation to improve IUGIM accuracy", *J. Korean Geotech. Soc.*, **38**(10), 17-29. https://doi.org/10.7843/KGS.2022.38.10.17.

Li, Z., Wang, X., Wang, H. and Liang, R.Y. (2016), "Quantifying stratigraphic uncertainties by stochastic simulation techniques based on Markov random field", *Eng. Geol.*, **201**, 106-122. https://doi.org/10.1016/j.enggeo.2015.12.017.

Lim, H.H., Lee, S.M., Cheon, E., Song, E., Jeon, J.S. and Lee, S.R. (2024), "Landslide mapping using a combination of sentinel-2 multi spectral instruments and GIS data at Namwon, Jeollabuk-do, South Korea", *Geomech. Eng.*, **39**(4), 385-396. https://doi.org/10.12989/gae.2024.39.4.385.

Ma, J., Jiang, S., Liu, Z., Ren, Z., Lei, D., Tan, C. and Guo, H. (2022), "Machine learning models for slope stability classification of circular mode failure: An updated database and Automated Machine Learning (AutoML) approach", *Sensors*, **22**(23), 9166. https://doi.org/10.3390/s22239166.

McInnes, L., Healy, J. and Melville, J. (2018), "Umap: Uniform manifold approximation and projection for dimension reduction", *arXiv preprint arXiv:1802.03426*.

Moon, S.W., Subramaniam, P., Zhang, Y., Vinoth,G. and Ku. T. (2019), "Bedrock depth evaluation using microtremor measurement: empirical guidelines at weathered granite formation in Singapore", *J. Appl. Geophys.*, **171**, 103866. https://doi.org/10.1016/j.jappgeo.2019.103866.

Mwakapesa, D.S., Lan, X., Nanehkaran, Y.A. and Mao, Y. (2023), "Landslide susceptibility mapping using O-CURE and PAM clustering algorithms", *Front. Environ. Sci.*, **11**, 1140834. https://doi.org/10.3389/fenvs.2023.1140834.

Nur, A.S., Kim, Y.J. Nam, B.H., Park, K. and An, J. (2025), "Subsidence characterization of karst sinkholes using satellite remote sensing: A Missouri case study", *Geomech. Eng.*, **41**(1), 151-163. https://doi.org/10.12989/gae.2025.41.1.151.

Park, K., Kim, Y.J., Chen, J. and Nam, B.H. (2024), "InSAR-based investigation of ground subsidence due to excavation: a case study of Incheon City, South Korea", *Int. J. Geo-Eng.*, **15**(1), 1-10.

Park, S., Hwang, C., Hwang, B. and Choi, H. (2025), "Data-driven modeling for interfacial behaviors between frozen soil and existing structures for applications of artificial ground freezing", *Geomech. Eng.*, **40**(3), 151-163. https://doi.org/10.12989/gae.2025.40.3.151.

Peng, F.L., Dong, Y.H., Wang, W.X. and Ma, C.X. (2023), "The next frontier: Data-driven urban underground space planning orienting multiple development concepts", *Smart Constr. Sustain. Cities*, **1**(1), 3. https://doi.org/10.1007/s44268-023-00003-5.

Phoon, K.K., Ching, J. and Shuku, T. (2022), "Challenges in data-driven site characterization", *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, **16**(1), 114-126. https://doi.org/10.1080/17499518.2021.1896005.

Phoon, K.K. and Kulhawy, F.H. (1999), "Characterization of geotechnical variability", 36.

Qi, X., Pan, X., Chiam, K., Lim, Y.S. and Lau, S.G. (2020), "Comparative spatial predictions of the locations of soil-rock interface", *Eng. Geol.*, **272**, 105651. https://doi.org/10.1016/j.enggeo.2020.105651.

Qi, X.H., Li, D.Q., Phoon, K.K., Cao, Z.J. and Tang, X.S. (2016), "Simulation of geologic uncertainty using coupled Markov chain", *Eng. Geol.*, **207**, 129-140. https://doi.org/10.1016/j.enggeo.2016.04.017.

Ryu, H.H., Choi, S., Chong, S.H., Kim, T.Y., Lee, J. and Kang, M. (2025), "Machine learning-based prediction of underground utility counts using electrical resistance numerical data", *Geomech. Eng.*, **41**(1), 11-19. https://doi.org/10.12989/gae.2025.41.1.011.

Sahin, E.K. and Demir, S. (2023), "Greedy-autoML: A novel greedy-based stacking ensemble learning framework for assessing soil liquefaction potential", *Eng. Appl. Artif. Intell.*, **119**, 105732. https://doi.org/10.1016/j.engappai.2022.105732.

Shen, Y., Zhang, D., Wang, R., Li, J. and Huang, Z. (2023), "SBD-K-medoids-based long-term settlement analysis of shield tunnel", *Transport. Geotech.*, **42**, 101053. https://doi.org/10.1016/j.trgeo.2023.101053.

Vantassel, J.P., Kumar, K. and Cox, B.R. (2022), "Using convolutional neural networks to develop starting models for near-surface 2-D full waveform inversion", *Geophys. J. Int.*, **231**(1), 72-90. https://doi.org/10.1093/gji/ggac179.

Wang, R. and Zhang, L. (2023), "K-means-based heterogeneous tunneling data analysis method for evaluating rock mass parameters along a TBM tunnel", *Sci, Rep,*, **13**(1), 21564. https://doi.org/10.1038/s41598-023-49033-0.

Wang, X., Wang, H., Liang, R.Y., Zhu, H. and Di, H. (2018), "A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data", *Struct. Saf.*, **70**, 128-138. https://doi.org/10.1016/j.strusafe.2017.10.011.

Wang, Y., Hu, Y. and Zhao, T. (2020), "Cone penetration test (CPT)-based subsurface soil classification and zonation in two-dimensional vertical cross section using Bayesian compressive sampling", *Can. Geotech. J.*, **57**(7), 947-958. https://doi.org/10.1139/cgj-2019-0131.

Wang, Z.Z., Hu, Y., Guo, X., He, X., Kek, H.Y., Ku, T., Goh, S.H. and Leung., C.F. (2023), "Predicting geological interfaces using stacking ensemble learning with multi-scale features", *Can. Geotech. J.*, **60**(7), 1036-1054. NRC Research Press.

https://doi.org/10.1139/cgj-2022-0365.
Yang, H.Q., Chu, J., Qi, X., Wu, S. and Chiam, K. (2023), "Bayesian evidential learning of soil-rock interface identification using boreholes", *Comput. Geotech.*, **162**, 105638. https://doi.org/10.1016/j.compgeo.2023.105638.
Zhang, D., Shen, Y., Huang, Z. and Xie, X. (2022), "Auto machine learning-based modelling and prediction of excavation-induced tunnel displacement", *J. Rock Mech. Geotech. Eng.*, **14**(4), 1100-1114. https://doi.org/10.1016/j.jrmge.2022.03.005.
Zhu, X., Chu, J., Wang, K., Wu, S., Yan, W. and Chiam, K. (2021), "Prediction of rockhead using a hybrid N-XGBoost machine learning framework", *J. Rock Mech. Geotech. Eng.*, **13**(6), 1231-1245. https://doi.org/10.1016/j.jrmge.2021.06.012.
Zumrawi, M.M. (2014), "Effects of inadequate geotechnical investigations on civil engineering projects".

IC