https://doi.org/10.1186/s44147-025-00727-4

Journal of Engineering and Applied Science

RESEARCH Open Access

Interpretable machine learning for predicting compression index of clays using SHAP and gradient boosting models

Khaled Hamdaoui[1], Ali Benzaamia[1], Billal Sari Ahmed[1], Mohamed Elhebib Guellil[1] and Mohamed Ghrici[1*]

*Correspondence: m_ghrici@yahoo.fr

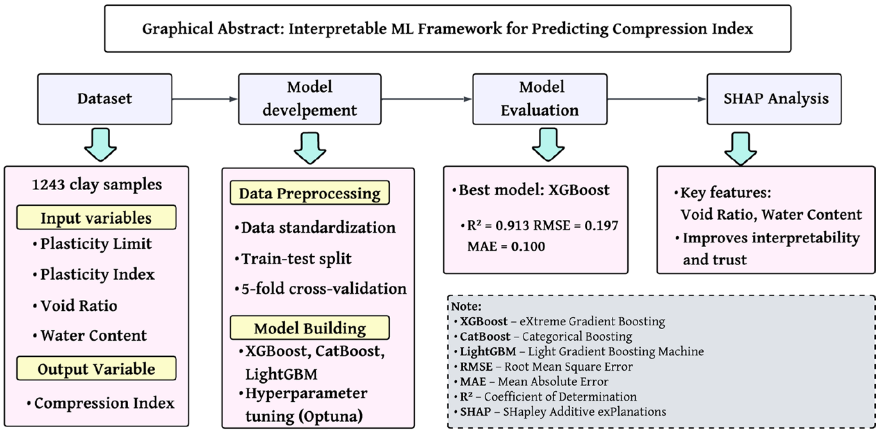
[1] Geomaterials Laboratory, Civil Engineering Department, Hassiba Benbouali University, Chlef, Algeria

Abstract

This study introduces a novel, interpretable machine learning framework for predicting the compression index (Cc) of clay soils by integrating three advanced gradient boosting algorithms—XGBoost, CatBoost, and LightGBM—with SHapley Additive exPlanations (SHAP). A comprehensive dataset of 1,243 clay samples, compiled from peer-reviewed literature, includes four geotechnical input variables: plastic limit (PL), plasticity index (PI), initial void ratio (e_0) and water content (w). Data were standardized and partitioned into training (70%) and testing (30%) subsets. Model development employed fivefold cross-validation and Optuna-based hyperparameter optimization. Among the models, XGBoost demonstrated the highest generalization capability, achieving an R^2 of 0.913, RMSE of 0.197, and MAE of 0.100 on the test set. SHAP analysis revealed that initial void ratio (e_0) and water content (w) were the most influential features, with mean SHAP values of 0.20 and 0.10, respectively, aligning with established geotechnical principles. The proposed framework enhances transparency in machine learning predictions by making the model's decision process understandable, thereby addressing the limitations of traditional "black-box" AI. It offers a reliable and efficient alternative to conventional oedometer testing, particularly beneficial for preliminary geotechnical design where timely and interpretable predictions are essential.

Keywords: Machine learning, Gradient boosting, Compression index, SHAP analysis, Geotechnical engineering, Interpretable AI

Graphical Abstract


Graphical Abstract: Interpretable ML Framework for Predicting Compression Index
Dataset
Model develepement
Model Evaluation
SHAP Analysis
1243 clay samples
Input variables
• Plasticity Limit
• Plasticity Index
• Void Ratio
• Water Content
Output Variable
• Compression Index
Data Preprocessing
• Data standardization
• Train-test split
• 5-fold cross-validation
Model Building
• XGBoost, CatBoost, LightGBM
• Hyperparameter tuning (Optuna)
• Best model: XGBoost
• R² = 0.913 RMSE = 0.197 MAE = 0.100
• Key features: Void Ratio, Water Content
• Improves interpretability and trust
Note:
• XGBoost – eXtreme Gradient Boosting
• CatBoost – Categorical Boosting
• LightGBM – Light Gradient Boosting Machine
• RMSE – Root Mean Square Error
• MAE – Mean Absolute Error
• R² – Coefficient of Determination
• SHAP – SHapley Additive exPlanations

Introduction

Soil subjected to structural loading experiences volumetric change, manifesting as settlement. This process is fundamentally a function of the soil's stress state, with the settlement magnitude influenced by multiple factors including the compression index. The compression index is a crucial parameter in the geotechnical design phase, involving the evaluation of the settlements of clayey soils. It is typically determined from oedometer test data, denoting the slope of the void ratio against the logarithm of effective pressure. However, obtaining oedometer test data requires a significant investment of time and specialized expertise. In response to this challenge, several authors have focused on numerical modeling to establish relationships between the compression index and the physical properties of soil, utilizing various techniques such as machine learning.

Machine learning constitutes a methodological approach that facilitates the construction of models derived from empirical data, enabling the identification and analysis of patterns across multidimensional variable spaces [1, 2]. This approach demonstrates versatility in its application, encompassing regression analysis, classification tasks, and clustering algorithms. The utility of machine learning in predicting soil properties has been extensively demonstrated across various geotechnical applications, offering significant advantages over traditional empirical methods. These applications include the prediction of unconfined compressive strength, maximum dry density, optimum moisture content, and compression index [3, 4]. These advantages include the ability to capture complex non-linear relationships, handle large datasets efficiently, and provide accurate predictions even when dealing with heterogeneous soil conditions. Ozer et al. [5] and Park and Lee [6] were among the first to use an Artificial Neural Network (ANN) model to predict the compression index (Cc), Several recent studies have explored the use of machine learning (ML) techniques for predicting geotechnical parameters such as the compression index (Cc).Recent advances in artificial intelligence and machine learning have significantly transformed the approach to predicting soil compression characteristics. Pham et al. [7] initially demonstrated the potential of AI models, including Artificial Neural Networks (ANN), Adaptive Network-based Fuzzy Inference System (ANFIS), and Support Vector Machine (SVM), establishing SVM's effectiveness in handling input uncertainties. Building upon this foundation, Zhang et al. [8] further advanced the field by evaluating Random Forest and Back-Propagation Neural Networks for predicting compression index of reconstituted clays. Bardhan et al. [9] introduced a hybrid approach combining Extreme Learning Machine with a Modified Equilibrium Optimizer, enhanced by Principal Component Analysis to address multicollinearity issues., while Asteris et al. [10] introduced an innovative approach using Manta Ray Foraging Optimization combined with Extreme Learning Machine. Subsequent research by Bui et al. [11] explored Particle Swarm Optimization with Multi-Layer Perceptron Neural Networks, emphasizing the critical role of parameter tuning. Recent studies have shown increasing sophistication, with Long et al. [12] conducting comparative analyses of tree-based models, and Díaz and Spagnoli [13] introducing an ensemble model approach using a super-learner algorithm. Contemporary research by Qiu Jiadong et al. [14] and Uzer [15] has further refined these approaches, focusing on hybrid models and artificial neural networks respectively, while Kim et al. [16] demonstrated the effectiveness of Random Forest Regressor algorithms in analyzing fine-grained soil compression index.

The application of machine learning in soil properties prediction has proven particularly valuable in addressing the complexity and variability inherent in geotechnical systems. Recent studies have demonstrated the effectiveness of various ML approaches in predicting soil behavior under different conditions [17, 18]. These investigations have shown that machine learning models can effectively capture the intricate relationships between soil index properties and engineering parameters, often outperforming traditional empirical correlations. The ability to process large datasets and identify subtle patterns has made ML techniques increasingly attractive for geotechnical applications, where accurate predictions are crucial for safe and economical design.

Throughout this evolution, certain parameters have consistently emerged as crucial predictors, including initial void ratio, water content, liquid limit, and plasticity index, highlighting their fundamental importance in soil compression behavior prediction [19]. However, despite their predictive accuracy, machine learning models often lack trust, explainability, and transparency [20]. These models are frequently described as "black box" systems [21], where the internal decision-making processes remain unclear, posing challenges for their application in geotechnical engineering. This opacity is particularly problematic in geotechnical applications where design decisions directly impact public safety and require rigorous justification.

To enhance the practical utility of machine learning in this field, eXplainable AI (XAI) introduces two key approaches: first, the development of models that achieve both high accuracy and improved interpretability; and second, the enhancement of user trust, understanding, and usability of advanced AI techniques [22, 23]. Various XAI methods have been proposed to provide explanations for model predictions across different data structures and machine learning architectures. Among these, SHapley Additive exPlanations (SHAP) [24] has emerged as a widely used framework for model interpretability due to its mathematical rigor and ability to provide both global and instance-level insights [25].

The limitations of previous studies in this domain are multifaceted. While numerous researchers have achieved high predictive accuracy using various machine learning algorithms, the majority of these studies have focused primarily on performance metrics without adequately addressing the interpretability and transparency of their models. This creates a significant barrier to practical implementation in geotechnical engineering, where understanding the reasoning behind predictions is as important as the predictions themselves. Furthermore, many existing studies have not sufficiently explored the integration of advanced gradient boosting methods with comprehensive interpretability frameworks, leaving a gap in our understanding of how these powerful algorithms make decisions in the context of soil compression behavior.

The research gap addressed by this study centers on the critical need for interpretable machine learning models in geotechnical engineering that can provide both high predictive accuracy and transparent insights into the decision-making process. While existing literature has demonstrated the effectiveness of various machine learning approaches for predicting compression index, there remains a significant lack of studies that systematically integrate advanced gradient boosting algorithms with comprehensive interpretability analysis. This gap is particularly pronounced in the context of SHAP-based interpretability, which has not been extensively applied to compression index prediction using state-of-the-art gradient boosting methods. Our study directly addresses this gap by implementing a comprehensive framework that combines the predictive power of XGBoost, CatBoost, and LightGBM with the interpretability insights provided by SHAP analysis, thereby advancing the state-of-the-art in transparent and reliable geotechnical machine learning applications.

Models and techniques

This section presents the application of three advanced gradient boosting algorithms—XGBoost, CatBoost, and LightGBM—used in this study for predicting the compression index of clays. It outlines the core architectures of each model and explains how SHapley Additive exPlanations (SHAP) were integrated to interpret their outputs. These algorithms were selected based on their proven effectiveness in handling structured tabular data, where they often outperform deep neural networks in terms of accuracy, computational efficiency, and interpretability [26].

Evolution of gradient boosting methods

The foundation of gradient boosting was established by Friedman [27], who introduced it as an ensemble learning method combining weak learners into a strong predictor. Traditional gradient boosting faced challenges with computational efficiency and scalability [28], leading to the development of optimized implementations.

Modern gradient boosting frameworks

XGBoost

XGBoost (eXtreme Gradient Boosting), introduced by Chen and Guestrin [28], represents a significant advancement in machine learning algorithms by enhancing traditional gradient boosting methods. The algorithm builds decision trees sequentially, with each new tree specifically designed to correct the errors made by the existing tree ensemble, while simultaneously employing sophisticated regularization techniques to prevent overfitting [29].

The algorithm's power lies in its optimization approach, which uses second-order approximations to make better decisions about tree construction. This method considers both the direction and magnitude of potential improvements, allowing for more precise model adjustments. Additionally, XGBoost incorporates practical innovations in handling sparse data and missing values, making it particularly robust for real-world applications.

One of XGBoost's most significant contributions is its efficient computing framework, which enables parallel processing through feature-based splitting and column block storage [27, 28]. This design, combined with its sparsity-aware split finding technique, has made XGBoost a preferred choice in both academic research and practical applications, particularly in situations involving large-scale data analysis and prediction tasks.

LightGBM

LightGBM (Light Gradient Boosting Machine), introduced by Ke et al. [30], represents a significant advancement in gradient boosting frameworks, particularly notable for its innovative approaches to handling large-scale datasets. The algorithm introduces two key techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which together dramatically improve computational efficiency while maintaining high prediction accuracy.

The algorithm's GOSS technique revolutionizes instance sampling by focusing on instances with larger gradients while randomly sampling those with smaller gradients, effectively reducing computational overhead without sacrificing model accuracy. Meanwhile, the EFB mechanism addresses feature dimensionality by intelligently bundling mutually exclusive features, treating the process as a graph coloring problem to optimize feature representation.

LightGBM's leaf-wise tree growth strategy, combined with its sophisticated handling of categorical features and missing values, sets it apart from traditional gradient boosting [26–30]. The algorithm employs histogram-based feature discretization and optimized split finding approaches, making it particularly effective for large-scale machine learning applications where both computational efficiency and model performance are crucial.

CatBoost

CatBoost (Categorical Boosting), introduced by Prokhorenkova et al. [31], represents an innovative advancement in gradient boosting algorithms, particularly distinguished by its sophisticated handling of categorical features and its solution to the prediction shift problem. The algorithm introduces ordered boosting, a novel approach that calculates gradients using models trained on carefully ordered subsets of the data, effectively addressing bias issues common in traditional gradient boosting methods.

A key innovation of CatBoost lies in its categorical feature handling mechanism, which employs ordered target statistics for encoding. This approach prevents target leakage while maintaining strong predictive power, making it particularly effective for datasets with categorical variables. The algorithm further enhances performance through its implementation of oblivious decision trees, where identical splitting criteria are applied across entire tree levels, enabling efficient computation while reducing overfitting risks.

The algorithm incorporates several optimization techniques, including an adaptive learning rate schedule and a symmetric tree structure utilizing multiple permutations to reduce prediction variance [28, 31]. These enhancements, combined with its innovative categorical feature processing, make CatBoost particularly effective for real-world applications where categorical data handling and prediction accuracy are crucial considerations.

Model interpretability

Identifying key input variables that significantly impact model predictions through sensitivity analysis is fundamental to developing robust machine learning systems [32]. SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee [24], represents a groundbreaking approach to machine learning model interpretability by applying concepts from cooperative game theory. This framework provides a mathematically rigorous method for understanding how each feature contributes to individual predictions, while maintaining important theoretical properties such as local accuracy and consistency. At its core, SHAP calculates feature importance through Shapley values, which can be expressed as:

$$\phi_i = \sum\nolimits_{S \subseteq F \setminus i} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup i) - f_x(S)] \tag{1}$$

where F represents all features, S is a subset of features excluding feature i, and f_x represents the model's prediction. This formulation ensures fair attribution of feature contributions by considering all possible feature combinations. The framework guarantees that predictions can be decomposed additively:

$$f(x) = \phi_0 + \sum\nolimits_{i=1}^{M} \phi_i \tag{2}$$

where ϕ_0 represents the base value (average prediction) and ϕ_i represents individual feature contributions. This decomposition satisfies the local accuracy property, ensuring that feature attributions sum to the difference between the model prediction and expected value [33].

For practical implementation in different model types, SHAP employs specific optimization techniques. For tree-based models, the TreeSHAP algorithm computes exact SHAP values with complexity $O(TLD^2)$, where T is the number of trees, L is the maximum number of leaves, and D is the maximum depth:

$$E[f(x)|x_S] = \sum\nolimits_{v \in V_L} p(v|x_S) f(v) \tag{3}$$

where V_L represents the set of leaf nodes and $p(v|x_S)$ is the conditional probability of reaching leaf v given feature subset S [24].

Table 1 Summary of Data Employed for Model Development and Corresponding Value Ranges

Reference	Data	PL range (%)	PI range (%)	e_0 range	w range (%)	Cc range
Kalantary and Kordnaeij [35]	391	10.00—34.00	3.00—50.00	0.36—1.88	10.20—70.00	0.05—0.63
Benbouras et al. [36]	353	0.00—35.50	8.20—69.00	0.28—1.14	8.00—42.10	0.01—0.46
Rekonen and Lojander [37]	167	17.70—43.20	4.70—128.00	0.81—3.88	28.00—155.00	0.10—4.22
Pajunen [38]	105	12.00—55.00	1.00—102.00	0.52—4.00	23.00—125.00	0.05—3.86
Díaz and Spagnoli [13]	95	8.00—39.00	2.00—50.00	0.38—1.09	14.90—36.60	0.05—0.23
Alhaji et al. [39]	38	17.70—41.60	8.40—33.60	0.51—1.24	10.90—37.00	0.12—0.32
LCPC [40]	24	16.00—56.00	7.50—82.00	0.50—3.00	20.00—131.00	0.13—1.64
Widodo and Ibrahim [41]	20	14.60—38.40	2.50—29.70	0.88—7.11	34.90—91.00	0.28—1.66
Gardemeister [42]	18	22.40—33.00	1.20—50.00	0.91—2.15	33.00—79.00	0.06—1.33
Zaman et al. [43]	14	11.70—30.90	12.10—44.30	0.64—1.20	25.20—45.90	0.24—0.53
Mitachi and Ono [44]	12	20.00—44.00	11.00—40.00	0.95—1.28	34.60—47.20	0.29—0.54
Koskinen [45]	3	22.30—29.80	52.80—74.20	2.08—3.58	76.70—129.00	1.03—2.99
Pätsi [46]	3	23.00—50.40	57.00—153.70	2.17—4.57	76.68—178.20	1.81—2.21
Total	**1243**	**0.00—56.00**	**1.00—153.70**	**0.28—7.11**	**8.00—178.20**	**0.01—4.22**

Table 2 Description of a compiled dataset

Variables	Symbol	Units	Category	Statistics						
				Mean	Std	Min	25%	50%	75%	Max
Plasticity limit	PL	(%)	Input	24.03	5.77	0.00	20.00	23.00	27.00	56.00
Plasticity Index	PI	(%)	Input	26.65	16.63	1.00	16.00	24.00	31.00	153.70
Void Ratio	e_0	–	Input	1.06	0.73	0.28	0.62	0.76	1.19	7.11
Water Content	w	(%)	Input	38.37	26.18	8.00	22.05	28.00	44.45	178.20
Compression index	**Cc**	**–**	**Output**	**0.44**	**0.61**	**0.01**	**0.15**	**0.20**	**0.35**	**4.22**

Materials and methodology

The Materials and Methodology section describes the sequential process implemented in this study. The methodology comprises clay data collection, data preparation, validation methodology, and predictive modeling workflow. These components establish the framework for developing reliable clay property predictions through systematic data analysis and model development.

Clay data collection

The dataset utilized in this study has been previously employed in research works by Löfman and Korkiala-Tanttu [34] and Díaz and Spagnoli [13]. Table 1 presents the compilation of these data points from various authors, including their respective input and output ranges collected through literature review.

The dataset utilized in this study comprises 1243 clay samples collected from various geotechnical projects. Each sample includes measurements of five input variables: initial void ratio (e_0), water content (w), liquid limit (LL), plastic limit (PL), and plasticity index (PI), with compression index (Cc) as the output variable. Table 2 presents the statistical summary of these parameters.

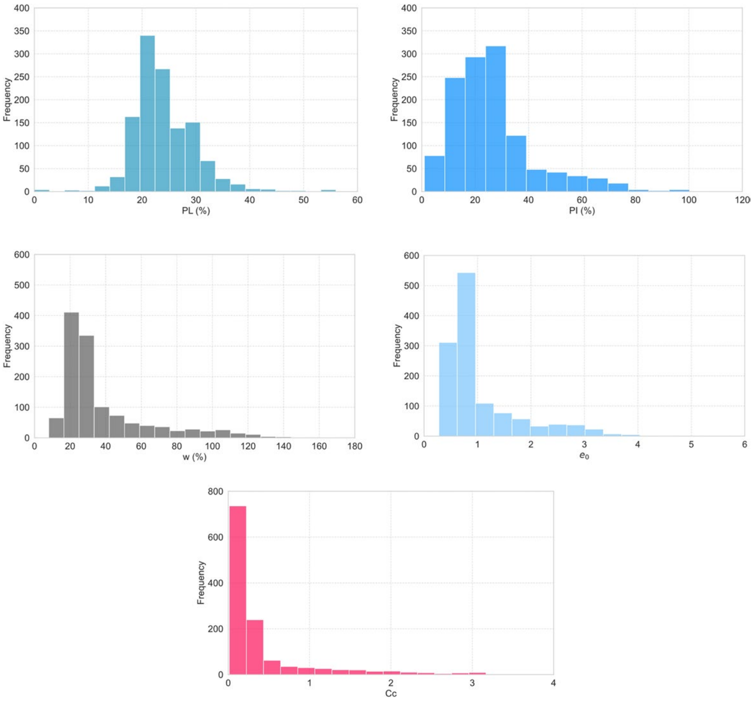


Fig. 1 The distribution of the compiled data

The compression index (Cc), which is the target variable for prediction, ranges from 0.01 to 4.22 with a mean of 0.44 and standard deviation of 0.61. The distribution shows most values concentrated in the lower range (25th-75th percentile: 0.15–0.35), with some high-value outliers increasing the maximum to 4.22. Figure 1 illustrates this distribution pattern, highlighting the right-skewed nature of the data with a concentration of values in the lower range.

The correlation matrix presented in Fig. 2 demonstrates strong relationships between several variables, notably between e_0 and Cc (r=0.89), and between water content and both e_0 (r=0.97) and Cc (r=0.89). These statistical relationships align with established soil mechanics principles and provide a solid foundation for developing predictive models.

Data preparation

Data standardization

Data standardization is a preprocessing technique that transforms features to a comparable scale by centering them around zero and scaling to unit variance. This transformation is achieved by subtracting the mean (μ) from each value and dividing by the standard deviation (σ). For a feature x, the standardized value z is calculated as:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

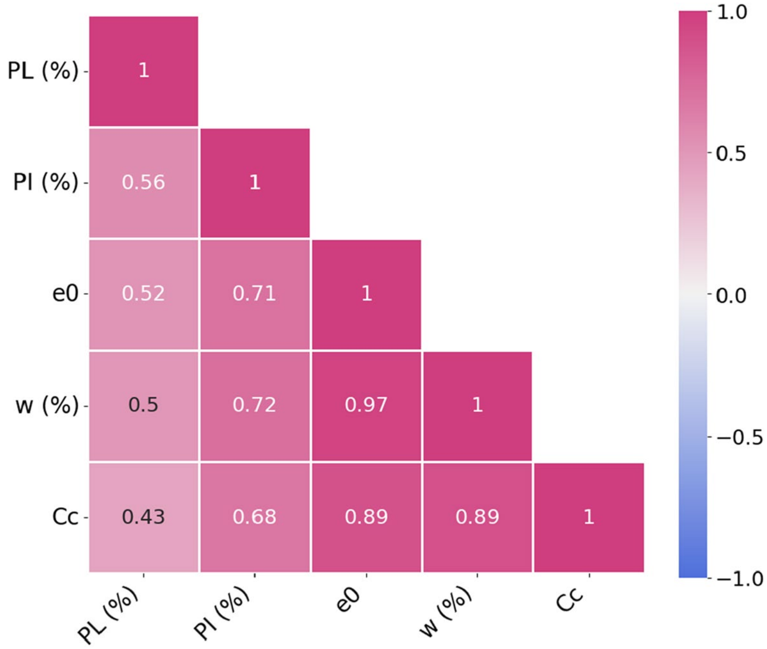


Fig. 2 Pearson's correlation

where:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2} \tag{6}$$

This process results in standardized features with a mean of 0 and a standard deviation of 1, which helps improve the performance and convergence of many machine learning algorithms.

Data preparation and validation methodology

Effective data preparation and validation are critical for ensuring the reliability and generalizability of machine learning models, particularly in the context of predicting the compression index (Cc). In this study, the dataset was initially partitioned into an 70% training set and a 30% test set, a two-way split designed to allocate sufficient data for model training while reserving an independent subset for final performance assessment. This approach proves advantageous for gradient boosting algorithms such as XGBoost, CatBoost, and LightGBM, which benefit from robust training data to capture the complex, non-linear relationships inherent in geotechnical datasets, often characterized by noise and variability.

To enhance model tuning and mitigate overfitting, a fivefold cross-validation strategy was implemented on the training set. This method divides the 70% training data into five equal subsets, where each iteration utilizes four subsets for training and the remaining subset (20% of the training data) for validation. The process is repeated five times, ensuring that each data point serves as validation data exactly once across all folds. This rigorous validation framework, recommended by Kohavi [47] and Hastie et al. [48], balances computational efficiency with result reliability, making it particularly suitable for the moderately sized geotechnical dataset employed in this study. Cross-validation provides a more robust estimate of model performance compared to a single train-test split, especially when dealing with limited or imbalanced data, as it leverages the entire training set to assess generalization across multiple configurations.

The combination of an 80%/20% split with fivefold cross-validation ensures a comprehensive evaluation pipeline. The initial split isolates the test set for unbiased performance assessment, while cross-validation within the training set facilitates hyperparameter optimization (e.g., via Optuna, as detailed in Results and discussion section) and model selection. This methodology addresses the challenges posed by geotechnical data, such as potential outliers or spatial heterogeneity, thereby enhancing the reliability of the predictive models for Cc in engineering applications.

Performance evaluation metrics for machine learning models

To evaluate the performance of machine learning models, various statistical metrics are used. Below are the definitions of commonly used metrics:

Root Mean Square Error (RMSE) measures the standard deviation of residuals, indicating the model's prediction accuracy. A lower RMSE value suggests better model performance.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2} \tag{7}$$

Mean Absolute Error (MAE) calculates the average absolute difference between predicted and actual values. It provides a direct interpretation of prediction errors in the same units as the target variable.

$$\mathrm{MAE} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|y_i - \widehat{y}_i\right| \tag{8}$$

It is worth noting that RMSE and MAE are typically referred to as "dimensional" metrics, as they express the average prediction error in the units of the target variable, which makes them easily interpretable in practical contexts.

Mean Absolute Percentage Error (MAPE) expresses the prediction error as a percentage, making it useful for comparing models across different datasets.

$$\mathrm{MAPE} = \frac{100}{n}\sum\nolimits_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{9}$$

Nash–Sutcliffe Efficiency (NSE) evaluates the predictive power of models, comparing the model's performance to a simple mean-based predictor. A value close to 1 indicates high predictive accuracy.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{10}$$

The Coefficient of Determination (R^2) measures the proportion of variance in the observed data explained by the model. A value near 1 indicates a good fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{11}$$

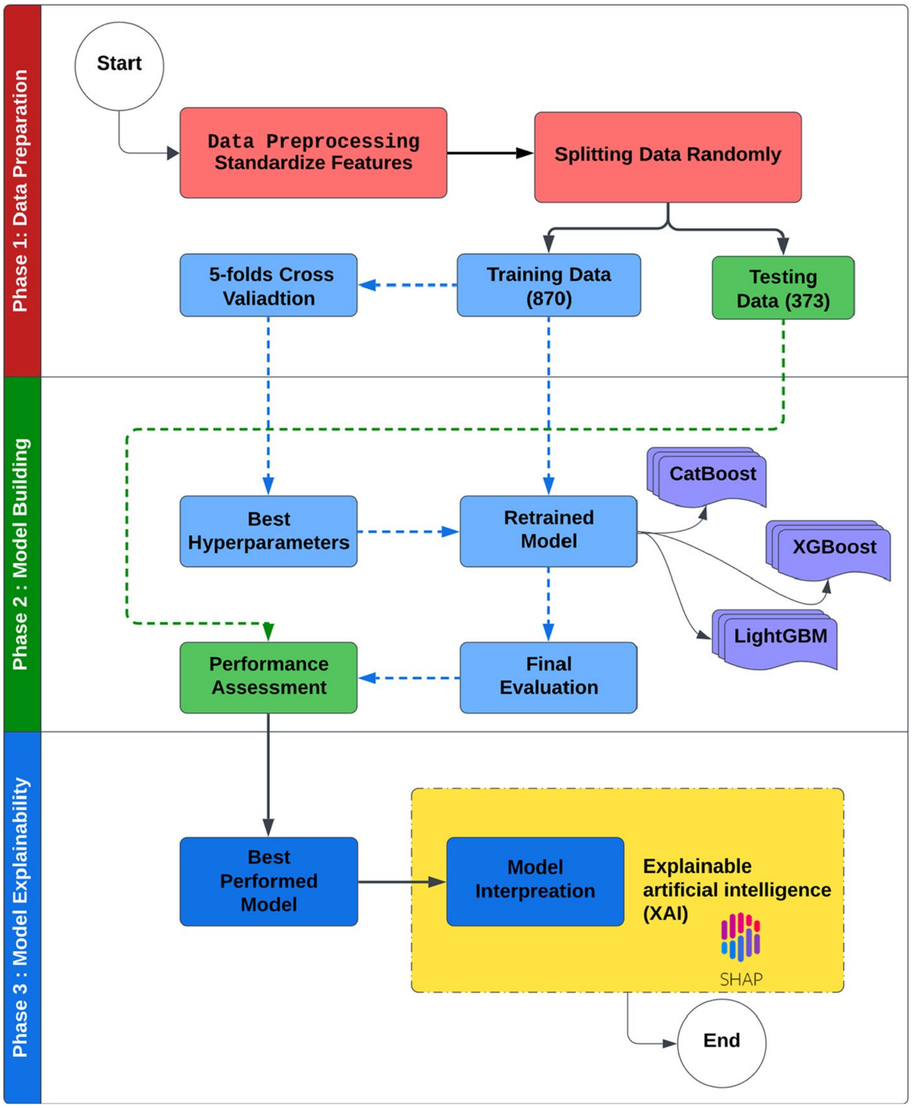


Fig. 3 Model interpretation Flowchart for the present study

Predictive modeling workflow

The workflow presented in Fig. 3 illustrates a comprehensive machine learning pipeline divided into three essential phases, beginning with data preparation. In this initial phase, the data undergoes preprocessing and feature standardization, followed by a random split into training (870 samples) and testing (373 samples) datasets. The training data is then subjected to fivefold cross-validation, which feeds into the model building phase. During this second phase, three gradient boosting algorithms—CatBoost, XGBoost, and LightGBM—are employed, with hyperparameters optimized through the cross-validation process.

The model building phase continues with retraining using the optimized parameters, followed by a final evaluation and performance assessment using the test data. In the third and final phase, focused on model explainability, the best-performing model is selected based on the performance assessment. This model then undergoes interpretation using SHAP (SHapley Additive exPlanations) to provide explainable artificial intelligence (XAI) insights, making the model's decisions transparent and interpretable. The workflow employs both dashed lines to indicate data flow and validation processes, and solid lines to show the main process flow, ensuring a balance between model performance and interpretability.

Results and discussion

This section presents the results of our machine learning models (XGBoost, CatBoost, and LightGBM) in predicting the compression index, along with SHAP-based interpretations of these models. We first discuss the performance metrics of each model, followed by an in-depth analysis of feature importance and interactions revealed by SHAP values.

Model performance comparison

Hyperparameter optimization results

To achieve optimal performance in predicting the compression index (Cc), the hyperparameters of the XGBoost, CatBoost, and LightGBM models were fine-tuned using Optuna, a next-generation hyperparameter optimization framework [49]. Unlike traditional methods such as grid search or random search, Optuna employs a Tree-structured Parzen Estimator (TPE) as its default sampler, which models the hyperparameter space probabilistically to efficiently identify promising configurations [50, 51]. Furthermore, Optuna supports trial pruning, enabling the early termination of unpromising trials based on intermediate performance, thereby reducing computational overhead [52]. This efficiency makes Optuna particularly well-suited for optimizing complex gradient boosting models, where hyperparameter tuning significantly influences predictive accuracy.

The optimization process was conducted over 1000 iterations, with the objective function defined as the mean RMSE of fivefold cross-validation scores on the validation dataset. This rigorous approach ensures robust model performance across different data subsets, mitigating the risk of overfitting and enhancing generalization

Table 3 Optimized Hyperparameters for XGBoost, CatBoost, and LightGBM Models

Parameter	XGBoost	CatBoost	LightGBM
Learning Rate	0.037253	0.094426	0.099849
Tree Depth	6	6	-
Number of Leaves	-	-	16
Number of Estimators	191	70	-
L2 Regularization	6.398019	-	0.003845
L1 Regularization	0.12936	-	0.001862
Subsampling Rate	0.628644	0.66197	-
colsample_bytree	0.830197	-	-
Column Sampling by Tree	-	0.835529	-
Column Sampling by Level	-	-	0.467955
Feature Fraction	9	-	-
Minimum Data in Leaf	-	2	-
Minimum Child Samples	-	-	13
Gamma (Regularization)	0.005704	-	-
L2 Leaf Regularization	-	1.19215	-
Maximum Bins	-	36	-
Random Strength	-	3.150102	-
Bagging Temperature	-	0.055311	0.471099
Bagging Fraction	-	-	4
Bagging Frequency	-	-	-

Table 4 Evaluation of Performance Indices for XGBoost, LightGBM, and CatBoost on Trainingand Testing Data

Stage	Models	RMSE	MAE	MAPE	NSE	R^2	R
Training	MLR	0.259	0.143	50.92	0.797	0.798	0.893
	LightGBM	0.170	0.073	22.647	0.912	0.912	0.955
	CatBoost	0.160	0.079	26.270	0.922	0.922	0.960
	XGBoost	**0.167**	**0.074**	**26.114**	**0.915**	**0.915**	**0.957**
Testing	MLR	0.278	0.160	45.75	0.829	0.829	0.915
	LightGBM	0.198	0.107	28.030	0.913	0.913	0.956
	CatBoost	0.207	0.104	25.673	0.905	0.905	0.951
	XGBoost	**0.197**	**0.100**	**25.474**	**0.913**	**0.913**	**0.956**

to unseen data. Table 3 summarizes the optimal hyperparameters identified for each model, revealing distinct configurations tailored to their respective architectures: XGBoost benefited from a relatively low learning rate (0.037253) and a high number of estimators (191), while CatBoost and LightGBM favored higher learning rates (0.094426 and 0.099849, respectively) with fewer estimators for CatBoost (70). Notably, the tree-based structure varies across models, with XGBoost and CatBoost sharing a tree depth of 6, while LightGBM optimized for 16 leaves. The regularization parameters (L1, L2) and sampling strategies also differ significantly among the models, highlighting the importance of model-specific hyperparameter tuning in achieving optimal performance for soil compression index prediction.

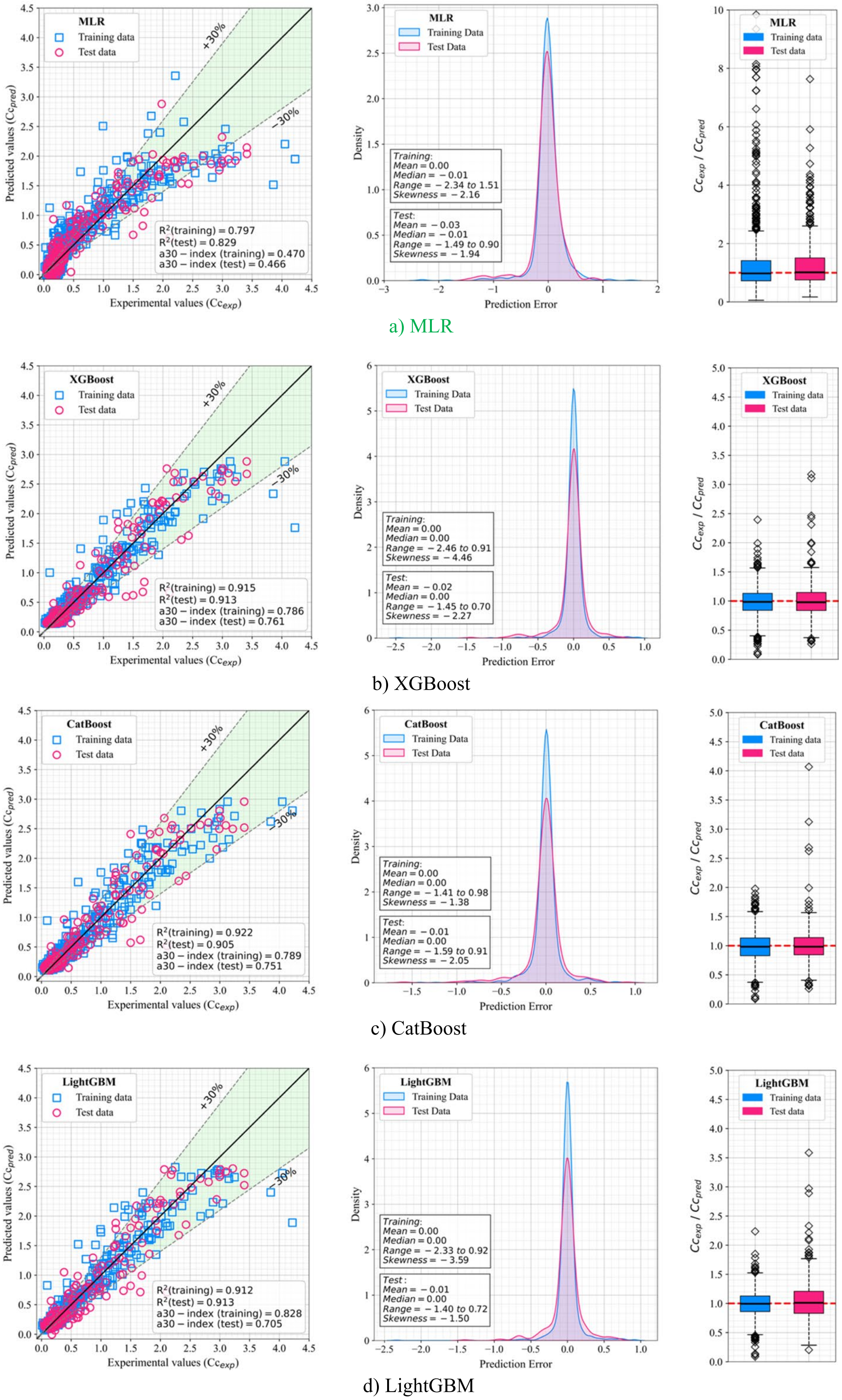
Evaluation of models performance using statistical metrics and visual analysis

The performance of the three gradient boosting models—XGBoost, CatBoost, and LightGBM—in predicting the compression index (Cc) is evaluated through a combination of statistical metrics and graphical visualizations. A comprehensive assessment is provided by first examining the numerical performance indices, followed by a detailed visual analysis to elucidate the models' predictive capabilities and generalization performance across training and test datasets.

Table 4 presents a quantitative evaluation of the models' performance across both training and testing phases. For baseline comparison, a Multiple Linear Regression (MLR) model is also evaluated. During the training phase, CatBoost demonstrates the highest predictive accuracy, achieving the lowest RMSE (0.160) and the highest NSE and R^2 values (0.922 for both), suggesting a strong capability to capture complex relationships within the training data. XGBoost and LightGBM exhibit comparable training performance, with RMSE values of 0.167 and 0.170, respectively, and R^2 values of 0.915 and 0.912. Notably, LightGBM records the lowest MAPE (22.647%), indicating smaller relative errors in training predictions. In contrast, MLR shows significantly higher RMSE (0.259) and MAPE (50.92%), with lower R^2 (0.798), highlighting its limited ability to model the nonlinear relationships present in the data.

In the testing phase, however, XGBoost and LightGBM outperform CatBoost, both achieving an R^2 of 0.913 compared to CatBoost's 0.905. XGBoost, in particular, records the lowest RMSE (0.197) and MAE (0.100), indicating superior generalization to unseen data, while CatBoost's higher test RMSE (0.207) suggests a decline in performance on the test set. XGBoost also achieves the lowest MAPE (25.474%) among the three models on the test data, further confirming its robustness and stability in prediction accuracy. MLR again shows relatively poor test performance, with the highest RMSE (0.2773), MAE (0.1598), and MAPE (45.75%), further emphasizing the advantages of advanced ensemble methods over traditional linear modeling**.** These numerical results highlight the importance of evaluating both training and testing performance to assess model robustness, with XGBoost emerging as the most balanced performer across both phases.

Figure 4 complements the numerical analysis by providing a detailed visual evaluation of the models' predictive accuracy through scatter plots, error density distributions, and ratio box plots for each model. The figure comprises three rows, each corresponding to a model (MLR, XGBoost, CatBoost, and LightGBM), with three subplots per row. The first column presents scatter plots of predicted versus experimental Cc values, with training data (blue squares) and test data (red circles) plotted against the 1:1 line, supplemented by ± 30% error bounds. For XGBoost, the scatter plot demonstrates strong alignment between predicted and experimental values, with R^2 values of 0.915 for the training set and 0.913 for the test set, and a30-index values (the proportion of predictions within ± 30% of the 1:1 line) of 0.786 and 0.761, respectively, aligning with its superior test performance in Table 4. CatBoost exhibits slightly higher training accuracy ($R^2 = 0.922$, a30-index = 0.789) but shows a decline in test performance ($R^2 = 0.905$, a30-index = 0.751), consistent with its higher test RMSE and suggesting potential overfitting. LightGBM yields R^2 values of 0.912 (training) and 0.913 (test), with a30-index values of 0.828 and 0.705, respectively, indicating robust training performance but a lower proportion of test predictions within the ± 30% bounds compared to XGBoost. In contrast,



a) MLR

b) XGBoost

c) CatBoost

d) LightGBM

Fig. 4 Performance evaluation of a) MLR, b) XGBoost, c) CatBoost, and d) LightGBM models: (left) scatter plots of predicted versus experimental Cc values, with ± 30% error bounds; (middle) kernel density plots of prediction errors ($Cc_{pred} - Cc_{exp}$); (right) box plots of the ratio Cc_{exp}/Cc_{pred}

the MLR model records considerably lower R^2 values of 0.797 (training) and 0.829(testing), along with a30-index values of 0.470 and 0.466, respectively, reinforcing its limited predictive capability and reduced reliability in capturing complex nonlinear relationships inherent in the dataset.

The second column of Fig. 4 displays kernel density plots of prediction errors, defined as ($Cc_{pred} - Cc_{exp}$), for both training and test datasets, where positive errors indicate overestimation and negative errors reflect underestimation. Annotated statistical values on the plot reveal the distribution characteristics. For XGBoost, the training and test error distributions peak at approximately 0, with a minor tendency to underestimate Cc in the test set, as indicated by a negative skew.

CatBoost's training errors are tightly clustered, while the test errors exhibit a wider spread, suggesting underestimation and potential overfitting. LightGBM's test errors show a slight negative shift, also indicating minor underestimation. Across all models, the test distributions exhibit negative skewness, with CatBoost displaying the widest error range and largest spread, consistent with its lower test R^2. The MLR model displays a notably different pattern, with pronounced negative skewness in both training and test error distributions, indicating a strong and consistent tendency to underestimate Cc. This extreme skewness, coupled with broader error spread, highlights the model's limitations in accurately capturing the underlying data distribution.

The third column presents box plots of the ratio Cc_{exp}/Cc_{pred}, where a ratio of 1 indicates perfect prediction, values above 1 suggest underestimation, and values below 1 indicate overestimation. XGBoost's training box plot has a median of about 0.99 and an IQR of 0.84 to 1.13, while the test box plot has a median of 0.98 and an IQR of 0.84 to 1.15, with outliers from 0.26 to 3.17, showing a slight underestimation in the test set. CatBoost's training box plot has a median of 0.98 and an IQR of 0.83 to 1.13, but the test box plot's IQR widens to 0.84 to 1.14 with outliers from 0.26 to 4.07, reinforcing underestimation and overfitting. LightGBM's training box plot has a median of 0.99 and an IQR of 0.86 to 1.13, while the test box plot's median is 1.01 with an IQR of 0.83 to 1.21 and outliers from 0.19 to 3.59, indicating slight underestimation consistent with its lower a30-index (0.705).In contrast, MLR's training box plot has a median of 0.90 and an IQR of 0.70 to 1.40, while the test box plot has a median of 0.95 and an IQR of 0.75 to 1.50, with ratios ranging from 0 to 10, showing a wide spread in prediction results. The low a30-index values of 0.470 for training and 0.466 for testing indicate that many predictions fall outside the acceptable ± 30% range, highlighting the model's poor accuracy and its limitations in capturing the complex behavior of the data.

The combined insights from Table 4 and Fig. 4 highlight the importance of assessing both metrics and visuals for model robustness. XGBoost's consistent performance—low test RMSE (0.197), high R2 (0.913), narrow error and ratio distributions, and high a30-index (0.761)—makes it the most reliable model for predicting the compression index in geotechnical engineering.

While XGBoost achieves the lowest MAPE (25.474%) among the tested models, this value remains relatively high. This can be attributed to the heterogeneous nature of the literature-derived dataset, which includes soils from diverse sources and conditions, and to the limited number of input parameters considered. Such variability inherently increases relative error, even when overall predictive accuracy remains strong.

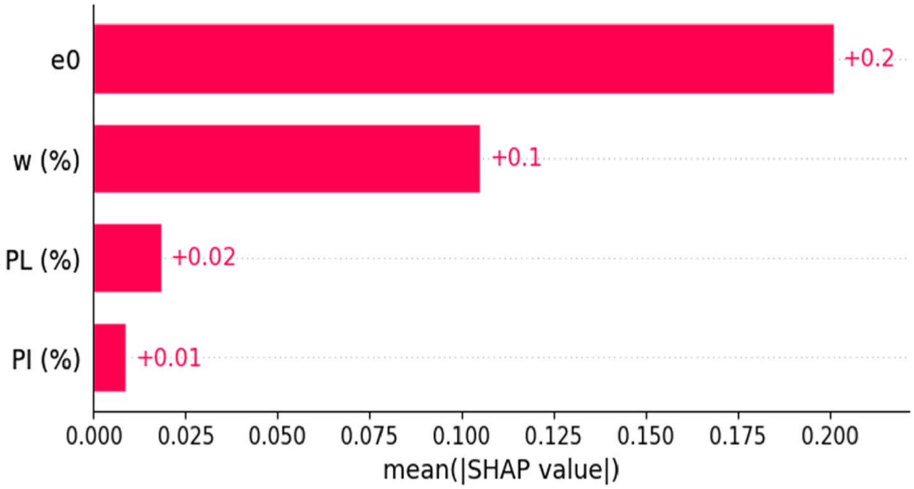


Fig. 5 Feature importance based on mean absolute SHAP values for the XGBoost model

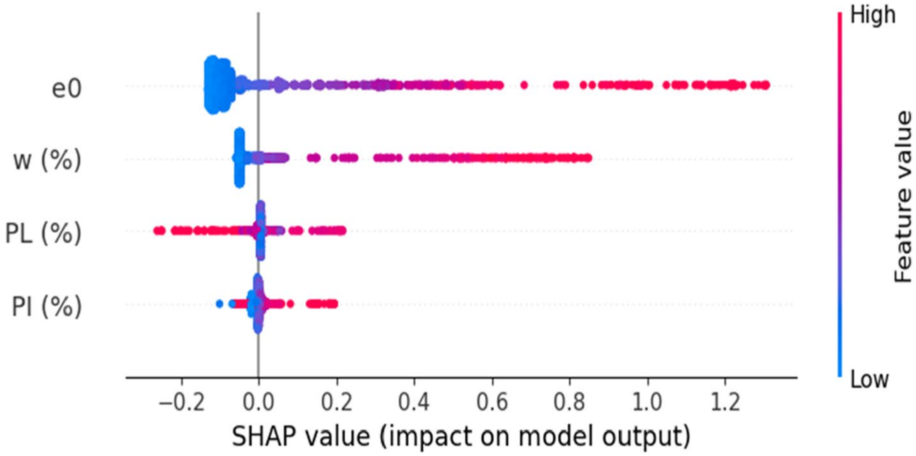


Fig. 6 SHAP summary plot illustrating feature importance in the XGBoost model

Nevertheless, the predictive capability of XGBoost can still provide substantial practical value in geotechnical engineering. By enabling rapid screening of multiple soil scenarios and identifying critical cases during early project stages, it can reduce the need for repetitive and time-consuming laboratory tests. Moreover, it can guide testing campaigns by pinpointing where and when detailed investigations are most needed, ultimately lowering costs and improving efficiency. However, laboratory tests remain indispensable for accurate, project-specific calculations—particularly in final design stages or for atypical soil conditions where empirical data is scarce.

SHAP interpretations

To clarify the decision-making processes of the XGBoost model, identified as the best-performing model in this study, and to enhance the interpretability of its predictions, SHapley Additive exPlanations (SHAP) analysis was applied. SHAP values provide a consistent framework for evaluating feature importance, quantifying each

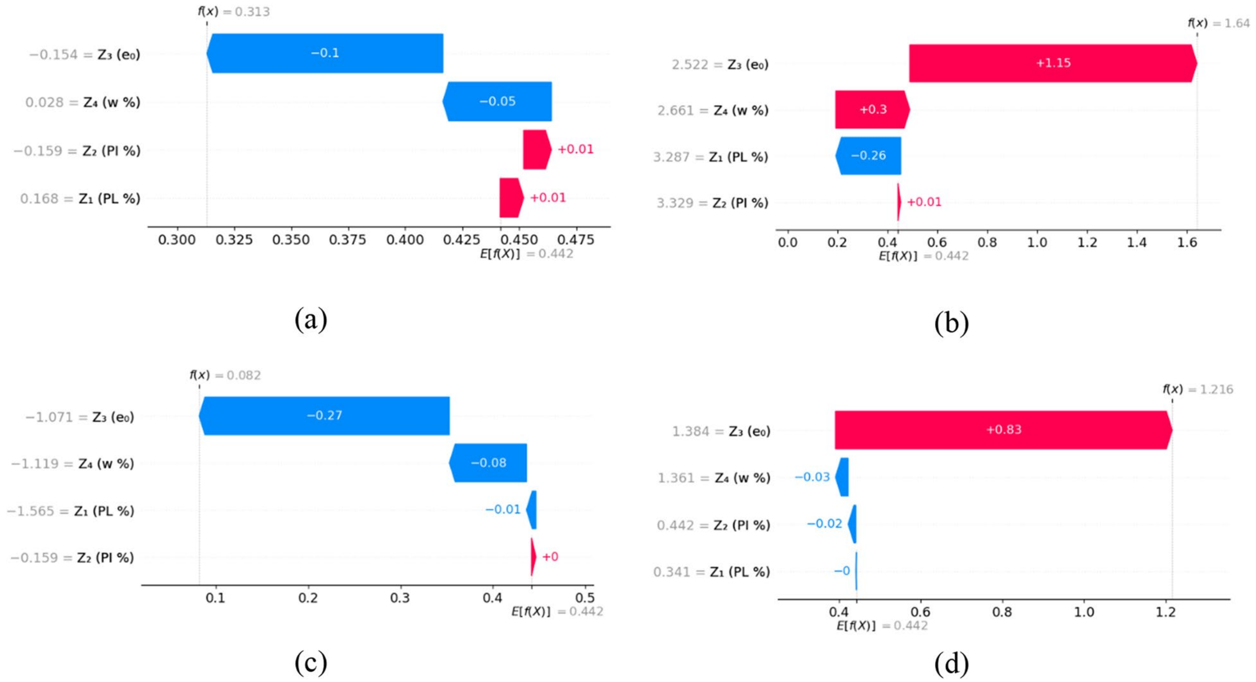


Fig. 7 SHAP force plot illustrating feature contributions to a single compression index prediction by the XGBoostmodel. (typical and atypical samples)

feature's contribution to individual predictions in terms of magnitude and direction. This approach holds particular relevance for compression index prediction, where understanding the influence of soil properties is paramount for geotechnical engineering applications.

Figure 5 presents the mean absolute SHAP values, facilitating the identification of features exerting the greatest influence on the XGBoost model's output across all instances. Furthermore, Fig. 6 offers a global explanation of the XGBoost model, illustrating how variations in feature values affect predictions and providing a comprehensive depiction of its decision-making framework. These interpretability tools, as shown in Figs. 6 and 7, enable a clear delineation of the primary factors shaping the model's predictions and their relative significance in determining outcomes. The analysis reveals void ratio (e_0) as the most influential parameter, with a mean SHAP value of 0.20, followed by initial water content (w) with a mean SHAP value of 0.10. By contrast, plasticity limit (PL) and plasticity index (PI) exhibit lesser impacts, with mean SHAP values of 0.02 and 0.01, respectively. These results align with established geotechnical principles, which emphasize the critical roles of void ratio and water content in soil compressibility, as supported by studies such as Zhang et al. [8], Pham et al. [7], Nhu et al. [53], Ramachandiran and Murugaiyan [54], and Long et al. [12].

The SHAP summary plot in Fig. 6 further clarifies the relative importance of factors influencing the compression index within the XGBoost model. Void ratio (e0) stands out as the predominant variable, with low values substantially decreasing the predicted compression index and high values significantly increasing it. Initial water content (w) follows as the second most impactful parameter, displaying a comparable trend: low values reduce the compression index, whereas high values elevate it markedly. Plasticity limit (PL) and plasticity index (PI) demonstrate moderate effects, with higher values generally increasing the compression index, while lower values exert minimal influence. These findings correspond closely with relationships observed during the data preprocessing phase. The figure effectively depicts how feature values, ranging from low (denoted in blue) to high (denoted in red), correspond to negative or positive impacts on the model's output across the range of SHAP values, offering a thorough perspective on feature contributions to compression index prediction.

To better understand how XGBoost predicts the compression index (Cc) of soil samples, SHAP (SHapley Additive exPlanations) analysis was applied to four representative cases.The SHAP force plots (Figs. 7-a to 7-d) break down the model's predictions into contributions from each input feature, starting from the model's expected value $E[f(X)] = 0.442$,hich reflects the average prediction across the dataset. All feature values used in the SHAP analysis were standardized using z-score normalization. Specifically, each input feature is represented as Z_i (Eq. 4) This allows for consistent comparison of the relative impact of different features across varying scales.

Figure 7-a illustrates a typical sample, where all input features have standardized values close to zero. The model adjusts the baseline prediction downward from 0.442 to a final output of $f(x) = 0.313$ the most influential feature is$Z_3(e_0) = -0.154$, which contributes $-$ 0.10 to the prediction. It is followed by$Z_4(w) = 0.028$, which contributes $-$ 0.05. In contrast, $Z_2(PI) = -0.159$ and $Z_1(PL) = 0.168$ slightly raise the prediction by $+$ 0.01 each. This case demonstrates a moderate prediction shift, primarily driven by slightly low values of void ratio and water content. Figure 7-b shows an atypical sample with a very high predicted compression index off(x) = 1.640. The standardized values of the dominant features are exceptionally high: $Z_3(e_0) = 2.522$ and$Z_4(w) = 2.661$, contributing + 1.15 and + 0.30, respectively. Although $Z_1(PL) = 3.287$ reduces the prediction by − 0.26, and $Z_2(PI) = 3.329$ adds a marginal + 0.01, the net effect remains strongly positive. This case clearly illustrates how high values of void ratio and water content dominate the model's decision, resulting in a significantly elevated compression index. Figure 7-c reflects an atypical sample with a very low predicted compression index off(x) = 0.082. The prediction is driven by low standardized feature values: $Z_3(e_0) = -1.071$ and$Z_4(w) = -1.119$, which contribute − 0.27 and − 0.08, respectively. The effects of $Z_1(PL) = -1.565$ and $Z_2(PI) = -0.159$ are minimal, contributing − 0.01 and ~ 0.00, respectively. This example emphasizes the influence of low void ratio and water content on reducing the compressibility of the soil sample, in line with geotechnical expectations. Finally, Fig. 7-d highlights an anomalous case with the largest prediction error among the four. Despite its input features having only moderately low standardized values—specifically$Z_3(e_0) = 1.384$,$Z_4(w) = 1.361$,$Z_2(PI) = 0.442$, and $Z_1(PL) = 0.341$ .The model predicts an output of f(x) = 1.216, which deviates significantly from the base value$E[f(X)] = 0.442$. The dominant contribution to this prediction comes from$Z_3(e_0)$, with a SHAP value of + 0.83, indicating a strong positive influence on the output. In contrast, the other features contribute minimally or negatively, with SHAP values of − 0.03, − 0.02, and ~ 0.00 for$Z_4(w)$,$Z_2(PI)$, and$Z_1(PL)$, respectively. Although the features individually seem moderate, their collective effect—dominated by e_0—results in an unexpectedly high prediction. This suggests that the model may have overfitted to patterns in regions of the feature space with limited representation, highlighting the importance of data distribution and model generalizability in interpreting extreme cases.

These observations regarding the dominant influence of void ratio e_0 and water content (w) on the compression index (Cc) are in strong agreement with previous findings reported by other researchers [33]. A plausible explanation is that higher e_0 values indicate looser soil structures with more void space, which are more susceptible to compression, while higher w values often correlate with weaker interparticle bonding and reduced effective stress, further enhancing compressibility. Conversely, low e_0 and w values generally correspond to denser, drier soils that exhibit limited deformation under loading.

Overall, this SHAP-based analysis demonstrates how XGBoost integrates and balances the contributions of standardized soil properties to produce reliable predictions. By distinguishing between typical, extreme, and anomalous responses, this interpretability framework strengthens confidence in the model and enhances its applicability in geotechnical engineering practices. It is worth noting that SHAP, particularly in its TreeSHAP implementation, assumes feature independence, an assumption that may be challenged in correlated geotechnical datasets. Additionally, SHAP values can sometimes misrepresent feature interactions when the model complexity is high.

Conclusions

This study introduced an interpretable machine learning framework for predicting the compression index (Cc) of clays using three gradient boosting algorithms—XGBoost, CatBoost, and LightGBM—combined with SHapley Additive exPlanations (SHAP) to enhance model transparency. Using a dataset of 1,243 clay samples characterized by five standardized geotechnical parameters, the models were developed through Optuna-based hyperparameter tuning and fivefold cross-validation. The study bridges the gap between high-performing machine learning models and interpretable engineering tools, offering a practical alternative to conventional lab-based methods in early geotechnical design. The following findings summarize the model's performance, interpretability outcomes, and contributions to engineering practice.

1. The XGBoost model showed superior predictive accuracy, generalizing well across data splits. Its test set metrics ($R^2 = 0.913$; RMSE = 0.197; MAE = 0.100) indicate strong model reliability for estimating Cc, making it suitable for practical engineering use.
2. SHAP analysis confirmed e_0 and w as the dominant features, with mean SHAP values of 0.20 and 0.1. This interpretability aligns with geotechnical theory, enhancing the credibility and applicability of the model's predictions.
3. The novelty of the study lies in integrating SHAP with ensemble methods for both performance and interpretability. This approach supports transparent decision-making, particularly in engineering fields that require explainable outputs.
4. The model offers a cost-effective, time-saving alternative to oedometer testing, enabling quick estimates of soil compressibility in preliminary design phases. Its transparent outputs are beneficial for both academic research and professional practice.

Research limitations

1. The dataset is literature-derived and lacks region-specific, mineralogical, or spatial attributes that might influence Cc. This may affect generalizability in site-specific applications.
2. Only four geotechnical parameters were considered. Incorporating additional soil descriptors could capture more complex behaviors and improve accuracy across diverse conditions.
3. SHAP may be limited by feature correlation, a common issue in geotechnical data. Interpretations should be cross-validated with engineering judgment to avoid misrepresentation of variable interactions.
4. The developed models are applicable primarily to fine-grained soils classified as CH, OH, CL, OL, and ML, according to the Unified Soil Classification System (ASTM D 2487–06).
5. The applicable ranges for both input and output variables are clearly summarized in Table 2, which defines the models' validity domain.

Recommendations for future research

1. Expand the feature set to include mineralogy, soil classification, or spatial/geographic variables to improve model robustness and contextual relevance.
2. Explore hybrid models combining gradient boosting with neural networks or metaheuristics to capture nonlinear interactions and enhance performance.
3. Introduce uncertainty quantification methods to improve reliability and support risk-informed decision-making.
4. Validate the model using experimental or region-specific field data to confirm its applicability in real-world geotechnical scenarios and promote industry adoption.

Abbreviation

Cc	Compression Index
Cc_{exp}	Experimental Compression Index
Cc_{pred}	Predicted Compression Index
e_0	Initial Void Ratio
w	Water Content
LL	Liquid Limit
PL	Plastic Limit
PI	Plasticity Index
ANN	Artificial Neural Network
ANFIS	Adaptive Network Fuzzy Inference System
SVM	Support Vector Machine
RF	Random Forest
BPNN	Backpropagation Neural Network
ELM	Extreme Learning Machine
MEO	Modified Equilibrium Optimizer
PCA	Principal Component Analysis
MRFO	Manta Ray Foraging Optimization
PSO	Particle Swarm Optimization
MLP	Multi-Layer Perceptron
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
XGBoost	EXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
CatBoost	Categorical Boosting
GOSS	Gradient-based One-Side Sampling
EFB	Exclusive Feature Bundling
Optuna	Hyperparameter Optimization Framework
TPE	Tree-structured Parzen Estimator
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NSE	Nash-Sutcliffe Efficiency
R^2	Coefficient of Det

Acknowledgements
This research was sponsored by the General Directorate for Scientific Research and Technological Development (DGRSDT) of the Algerian Minister of Higher Education and Scientific Research.

Authors' contributions

K. Hamdaoui: conceptualization, data curation, software, methodology, visualization, writing—original draft preparation. A. Benzaamia: data curation, software, methodology, writing—original draft preparation. B. Sari-Ahmed: conceptualization, investigation, methodology validation, writing—reviewing and editing. M. E. Guellil: methodology, writing—reviewing and editing. M. Ghrici: methodology, writing—reviewing and editing, supervision. All authors have read and approved the final manuscript.

Funding
No external funding was used.

Data availability

Data will be made available on reasonable request.

Declarations

Ethics approval and consent to participate

This article does not include any studies with human participants performed by any of the authors.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 21 May 2025 Accepted: 22 August 2025
Published online: 09 September 2025

References

1. Benzaamia A, Ghrici M, Rebouh R (2024) Machine learning approaches for predicting compressive and shear strength of EB FRP-reinforced concrete elements: A comprehensive review [J]. Stud Syst Decis Control 547:221–249. https://doi.org/10.1007/978-3-031-65976-8_12
2. Sari-Ahmed B, Ghrici M, Benzaamia A, Khatti J (2024) Assessment of unconfined compressive strength of stabilized soil using artificial intelligence tools: A scientometrics review. Studies in Systems, Decision and Control 547:271–288. https://doi.org/10.1007/978-3-031-65976-8_15
3. Mawlood YI, Salih A, Hummadi RA, Hasan AM, Ibrahim HH (2021) Comparison of artificial neural network (ann) and linear regression modeling with residual errors to predict the unconfined compressive strength and compression index for Erbil city soils, Kurdistan-Iraq. Arab J Geosci. https://doi.org/10.1007/s12517-021-06712-4
4. Ali HFH, Omer B, Mohammed AS, Faraj RH (2024) Predicting the maximum dry density and optimum moisture content from soil index properties using efficient soft computing techniques. Neural Comput Appl 36(19):11339–11369. https://doi.org/10.1007/s00521-024-09734-7
5. Ozer M, Isik NS, Orhan M (2008) Statistical and neural network assessment of the compression index of clay-bearing soils. Bull Eng Geol Environ 67(4):537–545. https://doi.org/10.1007/s10064-008-0168-8
6. Park HI, Lee SR (2011) Evaluation of the compression index of soils using an artificial neural network. Comput Geotech 38(4):472–481. https://doi.org/10.1016/j.compgeo.2011.02.011
7. Pham BT, Nguyen MD, Dao DV, Prakash I, Ly HB, Le TT, Ho LS, Nguyen KT, Ngo TQ, Hoang V, Son LH, Ngo HTT, Tran HT, Do NM, Van Le H, Van Khoi D, Nguyen HQ, Bui DT (2019) Development of artificial intelligence models for the prediction of compression coefficient of soil: an application of Monte Carlo sensitivity analysis. Sci Total Environ 679:172–184. https://doi.org/10.1016/j.scitotenv.2019.05.061
8. Zhang P, Yin ZY, Jin YF, Liu XF (2020) Intelligent modelling of clay compressibility using hybrid meta-heuristic and machine learning algorithms. Geosci Front 12(1):441–452. https://doi.org/10.1016/j.gsf.2020.02.014
9. Bardhan A, Asteris PG, Skentou AD, Samui P, Pilikas T (2022) Novel integration of extreme learning machine and improved Harris hawks optimization with particle swarm optimization-based mutation for predicting soil consolidation parameter [J]. J Rock Mech Geotech Eng 14(5):1588–1608. https://doi.org/10.1016/j.jrmge.2021.12.018
10. Asteris PG, Rizal F, Koopialipoor M, Roussis PC, Ferentinou M, Armaghani DJ, Gordan B (2022) Predicting clay compressibility using a novel Manta ray foraging optimization-based extreme learning machine model [J]. Transp Geotech 37:100861. https://doi.org/10.1016/j.trgeo.2022.100861
11. Bui DT, Nhu VH, Hoang ND (2018) Prediction of soil compression coefficient for urban housing project using novel integration machine learning approach of swarm intelligence and multi-layer perceptron neural network. Adv Eng Inform 38:593–604. https://doi.org/10.1016/j.aei.2018.09.005
12. Long T, Tran TT, Jiadong Q (2023) Tree-based techniques for predicting the compression index of clayey soils, www.jsoftcivil.com [Preprint]. https://doi.org/10.22115/scce.2023.377601.1579
13. Díaz E, Spagnoli G (2024) A super-learner machine learning model for a global prediction of compression index in clays. Appl Clay Sci 249:107239. https://doi.org/10.1016/j.clay.2023.107239
14. Jiadong Q, Ohl JP, Tran TT (2023) Predicting clay compressibility for foundation design with high reliability and safety: a geotechnical engineering perspective using artificial neural network and five metaheuristic algorithms. Reliab Eng Syst Saf 243:109827. https://doi.org/10.1016/j.ress.2023.109827
15. Uzer AU (2024) Accurate prediction of compression index of normally consolidated soils using artificial neural networks. Buildings 14(9):2688. https://doi.org/10.3390/buildings14092688
16. Kim M, Senturk MA, Li L (2024) Compression index regression of fine-grained soils with machine learning algorithms. Appl Sci 14(19):8695. https://doi.org/10.3390/app14198695
17. Subramaniam DN, Dassanayake DHHP, Ahilash N, Wijekoon SHB, Sathiparan N (2024) Characterisation of the shape of aggregates using image analysis. Int J Pavement Eng. https://doi.org/10.1080/10298436.2024.2349905
18. Sathiparan N (2024) Prediction model for compressive strength of rice husk ash blended sandcrete blocks using a machine learning models. Asian J Civ Eng 25(6):4745–4758. https://doi.org/10.1007/s42107-024-01077-x
19. Ali HFH, Mohammed AS (2024) Refining compression index estimation for fine soils: insights from large data and sensitivity analysis. Geotech Geol Eng. https://doi.org/10.1007/s10706-024-03017-7
20. Abdollahi A, Pradhan B (2023) Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. Sci Total Environ 879:163004. https://doi.org/10.1016/j.scitotenv.2023.163004
21. Castelvecchi D (2016) Can we open the black box of AI? Nature 538(7623):20–23. https://doi.org/10.1038/538020a

22. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R, Herrera F (2019) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [J]. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
23. Gunning D, Aha DW (2019) DARPA's explainable artificial intelligence (XAI) program. AI Mag 40(2):44–58. https://doi.org/10.1609/aimag.v40i2.2850
24. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9
25. Sathiparan N, Tharuka RANS, Jeyananthan P (2024) Prediction of moisture content of cement-stabilized earth blocks using soil characteristics, cement content, and ultrasonic pulse velocity. J Eng Appl Sci. https://doi.org/10.1186/s44147-024-00527-2
26. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G (2024) Deep neural networks and tabular data: a survey. IEEE Trans Neural Netw Learn Syst 35(6):7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161
27. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
28. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
29. Sari Ahmed B, Benzaamia A, Ghrici M, Ali A, Moghal B (2024) Strength prediction of fiber-reinforced clay soils stabilized with lime using XGBoost machine learning. Civil and Environmental Engineering Reports 34(2):1–18. https://doi.org/10.59440/ceer/190062
30. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154. https://doi.org/10.5555/3295222.3295349
31. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) Catboost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 31:6638–6648. https://doi.org/10.5555/3295222.3295474
32. Ali HFH, Mohammed AS (2024) New approaches to evaluate the impact of chemical oxides on the liquid limit, plasticity index, and unconfined compressive strength of clay soils. Geomech Geoengin 20(3):635–660. https://doi.org/10.1080/17486025.2024.2433627
33. Štrumbelj E, Kononenko I (2013) Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41(3):647–665. https://doi.org/10.1007/s10115-013-0679-x
34. Löfman MS, Korkiala-Tanttu L (2021) Transformation models for the compressibility properties of Finnish clays using a multivariate database. Georisk Assess Manag Risk Eng Syst Geohazards 16(2):330–346. https://doi.org/10.1080/17499518.2020.1864410
35. Kalantary F, Kordnaeij A (2012) Prediction of compression index using artificial neural network. Sci Res Essays 7(31):2835–2848. https://doi.org/10.5897/SRE12.1025
36. Benbouras MA, Mitiche R, Zedira H, Petrişor A, Mezouar N, Debiche F (2018) A new approach to predict the compression index using artificial intelligence methods [J]. Mar GeoresourGeotechnol 37(6):704–720. https://doi.org/10.1080/1064119x.2018.1484533
37. Rekonen R, Lojander M (1999) Painumaparametrien vaiheittainen määrittäminen (RITA-tietokanta). Internal Report 12(3):1999 (unpublished)
38. Pajunen H (1976) Hienorakeisten maalajien geoteknisten ominaisuuksien tutkimus v. 1976. Report M/1/1976, City of Vantaa
39. Alhaji MM, Alhassan M, Tsado TY, Mohammed YA (2017) Compression index prediction models for fine-grained soil deposits in Nigeria. Proceedings of the 2nd International Engineering Conference, Federal University of Technology, Minna, Nigeria
40. LCPC (1973) Remblais sur sols compressibles. Bulletin des Laboratoires des Ponts et Chaussées, Spécial T, 58. Paris (France), 361 p.
41. Widodo S, Ibrahim A (2012) Estimation of primary compression index (Cc) using physical properties of Pontianak soft clay. Int J Eng Res Appl 2:2231–2235
42. Gardemeister R (1973) Hienorakeisten maalajien geologisia ja geoteknisiä tutkimustuloksia. Otaniemi: VTT Technical Research Centre of Finland.
43. Zaman MW, Hossain MR, Shahin H, Alam AA (2016) A study on correlation between consolidation properties of soil with liquid limit, in situ water content, void ratio and plasticity index. In: Geotechnics for Sustainable Infrastructure Development, pp. 899–902.
44. Mitachi T, Ono T (1985) Prediction of undrained shear strength of overconsolidated clay. Tsuchi to Kiso (JSSMFE) 33:21–26
45. Koskinen E (2014) Porausdatan hyödyntäminen tunnelilouhinnassa kalliolaadun ja räjäytystulosten ennakoinnissa. [Predicting rock quality and blasting results in tunnelling with drilling data]. Master's Thesisi, Aalto University, (in Finnish)
46. Pätsi, K. (2009). Suurpellon syvästabiloidun koepenkereen analysointi (Master's thesis). Helsinki University of Technology, (in Finnish)
47. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95) 2:1137–1145. https://doi.org/10.5555/1643031.1643047
48. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York
49. Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: A next-generation hyperparameter optimization framework [C] // Proc 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD '19):2623–2631. https://doi.org/10.1145/3292500.3330701
50. Benzaamia A, Ghrici M, Rebouh R, Pilakoutas K, Asteris PG (2024) Predicting the compressive strength of CFRP-confined concrete using deep learning [J]. Eng Struct 319:118801. https://doi.org/10.1016/j.engstruct.2024.118801

51. Rebouh R, Benzaamia A, Ghrici M (2025) Bayesian-optimized tree-based models for predicting the shear strength of U-shaped externally bonded FRP-strengthened RC beams. Asian J Civ Eng 26:1465–1478. https://doi.org/10.1007/s42107-024-01258-8
52. Benzaamia A, Ghrici M, Rebouh R, Zygouris N, Asteris PG (2024) Predicting the shear strength of rectangular RC beams strengthened with externally-bonded FRP composites using constrained monotonic neural networks [J]. Eng Struct 313:118192. https://doi.org/10.1016/j.engstruct.2024.118192
53. Nhu VH, Samui P, Kumar D, Singh A, Hoang ND, Bui DT (2020) Advanced soft computing techniques for predicting soil compression coefficient in engineering project: a comparative study. Eng Comput 36:1405–1416. https://doi.org/10.1007/s00366-019-00772-7
54. Saisubramanian R, Murugaiyan (2021) Prediction of compression index of marine clay using artificial neural network and multilinear regression models, DOAJ (DOAJ: Directory of Open Access Journals) [Preprint]. https://doi.org/10.22115/scce.2021.287537.1324

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.