Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physics-guided explainable machine learning and uncertainty quantification for predicting unconfined compressive strength of nano shale–quicklime stabilized kaolin clay

Arash Aminaee
Shahid Rajaee Teacher Training University
Abolfazl Soltani
asoltani@sru.ac.ir

Shahid Rajaee Teacher Training University
Abolfazl Baghbani
La Trobe University
Parisa Salehan
Isfahan University of Technology
Hossam Abuel-Naga
La Trobe University
Pijush Samui
National Institute of Technology Patna

Article

Keywords: Physics-guided machine learning, Explainable artificial intelligence, Uncertainty quantification, Group-aware cross-validation, Unconfined compressive strength, Soil stabilization

Posted Date: September 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10784282/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Physics-guided explainable machine learning and uncertainty quantification for predicting unconfined compressive strength of nano shale–quicklime stabilized kaolin clay

Arash Aminaee [a], Abolfazl Soltani [*a 0000-0001-9490-2686], Abolfazl Baghbani [b], Parisa Salehan [c], Hossam Abuel-Naga [b], Pijush Samui [d]

[a] Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Lavizan, 1678815811, Tehran, Iran
[b] Department of Engineering, La Trobe University, Bundoora, Melbourne, VIC 3086, Australia
[c] Department of Chemistry, Isfahan University of Technology, Isfahan, 84156-83111, Iran
[d] National Institute of Technology Patna, Patna, India

ABSTRACT

Data-driven prediction of stabilized-soil strength is often constrained by small experimental datasets, leakage between related specimens, and limited physical interpretability. This study develops an explainable, uncertainty-aware framework to predict the unconfined compressive strength (UCS) of nano shale–quicklime stabilized kaolin. A specimen-level database of 180 measurements from 36 mixture–curing conditions was analyzed using multiple linear regression (MLR), support vector regression (SVR), random forest (RF), XGBoost, CatBoost, and backpropagation artificial neural network (BP-ANN), with nonlinear models optimized by particle swarm optimization and evaluated through nested GroupKFold cross-validation. Mechanism-inspired features represented binder dosage, binder–curing interactions, nonlinear curing behavior, relative plasticity, and compaction state. Sequential feature ablation identified a compact CatBoost model with six physics-guided descriptors, achieving $R^2 = 0.841 \pm 0.123$ and RMSE = 151.11 ± 34.94 kPa. SHAP analysis, sensitivity assessment, and two-dimensional partial dependence mapping identified plasticity, curing duration, and binder–curing interactions as dominant controls and revealed nonlinear stabilizer effects and favorable binder combinations. Monte Carlo input perturbation and group-bootstrap resampling quantified input-driven and model-related uncertainty, with the highest 95% prediction-interval coverage reaching 0.894. The framework provides a physically interpretable, reliability-aware approach for preliminary stabilized-soil screening while retaining confirmatory laboratory testing.

Keywords: Physics-guided machine learning; Explainable artificial intelligence; Uncertainty quantification; Group-aware cross-validation; Unconfined compressive strength; Soil stabilization

[*] **Corresponding Author:** Abolfazl Soltani, Orcid No. 0000-0001-9490-2686, Email: asoltani@sru.ac.ir

Authors' Email Addresses:
(Abolfazl Soltani, [*]Corresponding Author, asoltani@sru.ac.ir); (Arash Aminaee, arash.aminaee@gmail.com); (Abolfazl Baghbani, a.baghbani@latrobe.edu.au); (Parisa Salehan, p.salehan@ch.iut.ac.ir); Hossam Abuel-Naga H.AboEl-Naga@latrobe.edu.au); (Pijush Samui, pijush@nitp.ac.in).

1. Introduction

Soil stabilization is widely used in geotechnical engineering to improve the mechanical performance of weak fine-grained soils for applications such as pavement subgrades, embankments, foundations, and earthworks [1–4]. Among the available performance indicators, unconfined compressive strength (UCS) remains one of the most practical and widely used parameters for assessing the effectiveness of stabilization and supporting preliminary design decisions [5–7]. The strength development of stabilized soils is governed by several interacting factors, including stabilizer type and dosage, curing days, soil plasticity, moisture condition, and compaction characteristics [8,9]. These factors influence the formation of cementitious and pozzolanic products, interparticle bonding, and the evolution of soil structure [2,4]. As a result, UCS response is commonly nonlinear and interaction-dependent, making it difficult to represent accurately using simple empirical correlations alone [10,11].

Traditional laboratory-based UCS evaluation provides direct and reliable measurements. However, it becomes time-consuming and resource-intensive when many mix designs, binder dosages, and curing conditions must be examined [11]. This limitation has motivated the development of empirical, statistical, and machine learning (ML)-based approaches for rapid strength estimation. Several studies have shown that ML models can capture complex nonlinear relationships in stabilized soils more effectively than conventional regression models. For example, Onyelowe et al. [12] used multiple ensemble-based ML methods to estimate the UCS of cohesive soils stabilized with cement and lime, showing that models such as gradient boosting and K-nearest neighbors could provide high prediction accuracy while identifying the roles of MDD, consistency limits, cement content, and lime content in UCS development. Similarly, Sari-Ahmed et al. [13] developed an XGBoost-based model for lime-stabilized fiber-reinforced clay soils using 341 test data collected from 14 experimental studies, demonstrating the usefulness of boosting algorithms for capturing the effects of lime content, fiber properties, and curing duration on UCS. These studies confirm the value of ML as a practical alternative for reducing experimental effort, but most of them still focus primarily on point prediction accuracy.

Recent studies have increasingly moved toward more interpretable ML frameworks. Mohammed et al. [14] compiled a 200-sample database for GGBS-stabilized soils and compared several ML models, reporting XGBoost as the best-performing model and using SHAP and LIME to interpret the effects of curing period, OMC, and MDD on UCS. Chen et al. [5] examined one-part geopolymer-stabilized soils using hybrid ensemble models optimized by particle swarm optimization (PSO), and further used SHAP and one- and two-dimensional partial dependence plots to interpret the effects of input parameters on UCS. These studies show that explainable ML methods can provide useful insight into model behavior and support engineering interpretation. Nevertheless, interpretability is often added after model development, while the input feature space itself usually remains mainly data-driven and does not explicitly reflect the underlying stabilization mechanisms [15,16].

In parallel, physics-informed and domain-informed ML have gained attention in geotechnical engineering because they can improve model transparency and physical consistency. Recent studies have emphasized that incorporating soil science or geotechnical knowledge into ML models can help address data scarcity, improve interpretability, and reduce physically unrealistic predictions [17,18]. In soil stabilization, Yang et al. [19] recently proposed a physics-informed and SHAP-enhanced framework for cement-stabilized soils, where a modified water–cement ratio model was integrated through hybrid-

corrector and PINN strategies; the best physics-informed model achieved $R^2 = 0.941$ on the test set, and SHAP analysis quantified the contributions of cement, water content, and clay content. This represents an important step toward mechanistically consistent UCS prediction. However, such approaches are often developed for cement-stabilized soils and rely on physical formulations specific to cement-water-clay systems. For alternative stabilizers, especially natural pozzolanic additives, explicit governing equations are often unavailable or insufficiently established. In such cases, physics-guided feature engineering offers a practical intermediate approach by embedding mechanism-inspired variables, such as binder dosage, binder-curing interactions, curing nonlinearity, compaction state, and relative plasticity, directly into the model input space [20]. In this study, the term physics-guided is used in this feature-level sense, where geotechnical mechanisms guide the construction and selection of input descriptors rather than imposing explicit governing-equation constraints on the learning algorithm.

Another important issue is validation reliability. Many UCS prediction studies still rely on simple random train-test splits or standard cross-validation [21–23], which may overestimate model generalization when repeated specimens, similar mix designs, or related curing conditions appear in both training and testing subsets [5]. This concern is particularly relevant for small laboratory datasets, where multiple records may share the same mixture composition or experimental origin. More robust validation strategies, such as nested cross-validation, have recently been adopted in geotechnical ML to reduce selection bias and improve performance estimation. For example, Teodoru et al. [20] used nested cross-validation and Bayesian hyperparameter optimization for cement-treated soils, while also developing an uncertainty quantification system to provide both central UCS estimates and confidence intervals. This reflects a broader shift from reporting only high accuracy toward assessing model robustness and reliability.

Uncertainty quantification is also becoming increasingly important in geotechnical ML because design decisions are rarely based on deterministic values alone [24,25]. Material variability, testing scatter, curing conditions, and compaction-related uncertainties can all affect the reliability of predicted UCS [20,24,26]. Some recent studies have begun to incorporate uncertainty-aware prediction systems for cement-treated soils [20], but uncertainty analysis is still less common in ML studies on stabilized soils compared with deterministic model comparison and feature-importance analysis. In particular, limited attention has been given to distinguishing uncertainty arising from input variability from uncertainty associated with model instability.

Despite these advances, several gaps remain. First, most available ML-based UCS prediction studies are centered on cement-, lime-, GGBS-, fly ash-, geopolymer-, or fiber-reinforced stabilized soils [25,27–30], while NS-QL stabilized kaolin clay has received little attention from an interpretable ML perspective [26]. Second, although XAI methods such as SHAP and PDP are increasingly used, fewer studies combine them with mechanism-inspired feature engineering. Third, physics-informed frameworks for stabilized soils are still mostly developed around cement-specific physical equations, whereas many emerging sustainable stabilizers lack sufficiently established governing formulations. Fourth, robust validation and uncertainty-aware assessment are not always integrated into a single workflow, especially for small laboratory datasets with repeated mixture and curing structures. In addition, limited attention has been given to how the availability of input information affects the practical use of ML models, particularly the distinction between early-stage predictions based only on binder dosage and curing days and more reliable predictions that also require geotechnical index and compaction properties.

To address these gaps, this study develops an interpretable and uncertainty-aware ML framework supported by mechanism-inspired, physics-guided feature engineering for UCS prediction of kaolin clay stabilized with NS and QL. The experimental database used in this work was obtained from a previously published laboratory study and includes 180 UCS measurements covering different mix designs and curing days [26]. The present contribution focuses on predictive modeling, physics-guided feature construction, robust validation, interpretability, and uncertainty quantification rather than new laboratory testing. The proposed workflow, shown in Figure 1, includes exploratory analysis, construction of mechanism-inspired features, PSO-based model optimization, nested GroupKFold validation, ablation-based feature selection, SHAP/sensitivity/2D-PDP interpretation, and Monte Carlo plus group-bootstrap uncertainty analysis. Using this proposed model, the study aims to provide not only accurate UCS prediction, but also physically interpretable and reliability-aware decision support for preliminary mix design and more efficient use of stabilizing agents in geotechnical soil stabilization. Accordingly, the novelty of this study does not lie in introducing each ML component independently, but in integrating mechanism-inspired feature engineering, group-aware validation, ablation-based feature refinement, explain ability, reduced-input assessment, and uncertainty quantification into a single workflow for NS-QL stabilized kaolin clay.

The main contributions of this study are as follows:

1. Development of an interpretable ML framework supported by mechanism-inspired, physics-guided feature engineering for UCS prediction of NS-QL stabilized kaolin clay using both raw geotechnical variables and engineered descriptors.
2. Evaluation of multiple optimized ML models under a nested GroupKFold validation framework to improve the reliability of generalization assessment for a small laboratory dataset.
3. Assessment of reduced-input modeling scenarios to quantify the predictive value of binder dosage and curing days alone for preliminary mix design screening.
4. Use of ablation analysis to identify a compact and physically meaningful engineered feature subset.
5. Interpretation of the selected best-performing model using SHAP, sensitivity analysis, and 2D partial dependence plots to examine whether the learned relationships are consistent with expected stabilization behavior.
6. Quantification of predictive uncertainty using Monte Carlo input perturbation and group-bootstrap resampling to compare input-driven uncertainty with model-instability uncertainty.

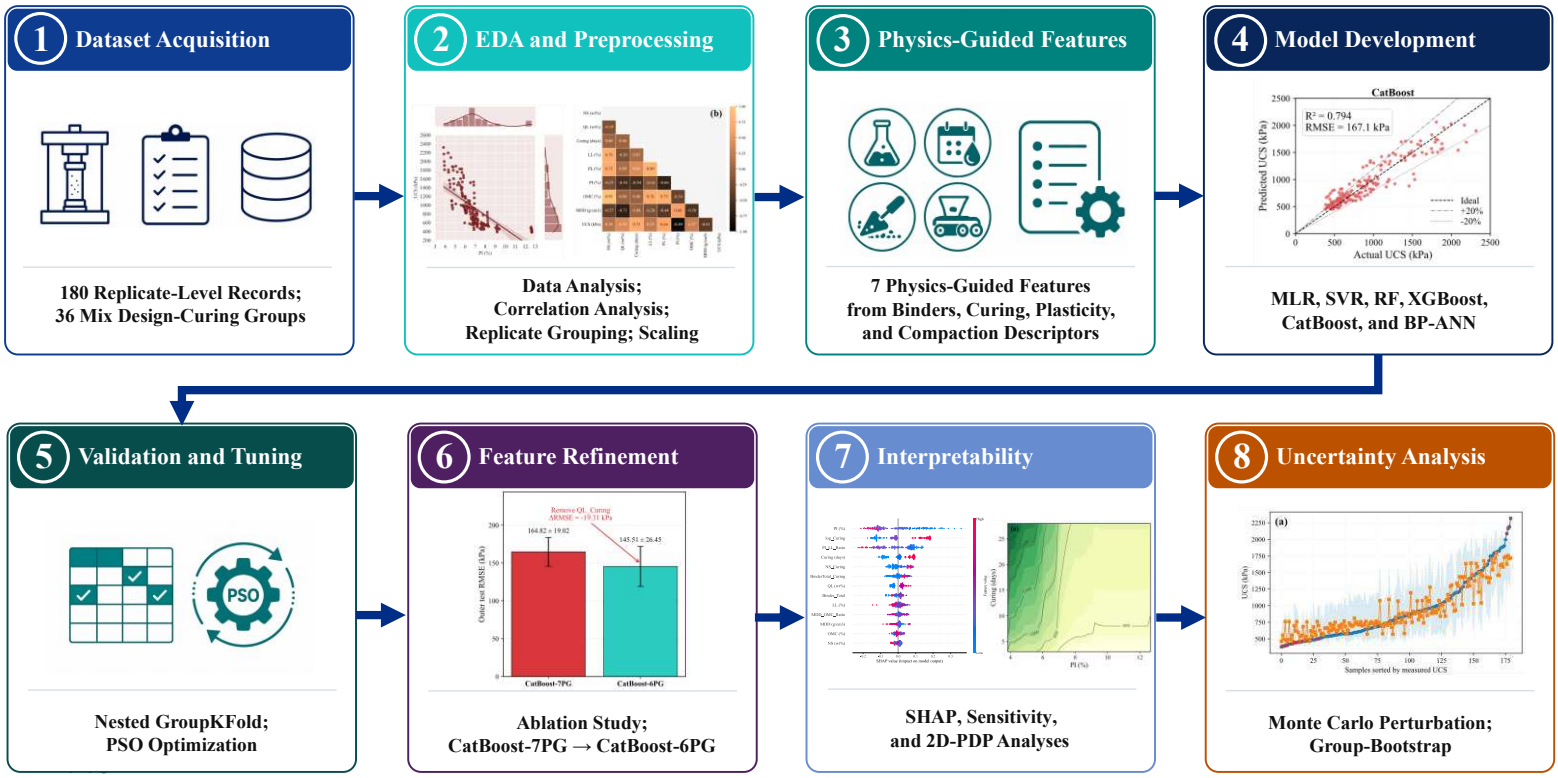


Figure 1. Workflow of this study

2. Materials, Dataset, and Methodology

2.1. Experimental Database

The experimental dataset used in this study was collected from a previously published investigation on the stabilization of kaolin clay using nano hydrocarbonated shale and quicklime [26]. The base soil was kaolin clay, which is characterized by weak mechanical properties and a relatively low plasticity index. The present study does not involve new laboratory testing; instead, it uses the reported experimental results as the basis for developing, validating, interpreting, and quantifying the uncertainty of ML models.

The nano-shale (NS) samples were collected from the Tazareh Coal Mine, located approximately 45 km east of Damghan, Iran. The NS particles had undergone natural thermal metamorphic processes and were subsequently produced from hydrocarbon-bearing shale through mechanical crushing and grinding without any additional thermal treatment. This preparation process may reduce energy demand compared with thermally activated binders. Its chemical composition, with total ($SiO_2 + Al_2O_3 + Fe_2O_3$) > 70 wt.% and loss on ignition below 10%, satisfies the pozzolanic requirements of ASTM C618-22 [31]. Quicklime (QL) was primarily used as an alkaline activator to increase pH of mixtures and promote pozzolanic reactions. To evaluate the individual effect of lime, QL portions of 1, 3, and 5 wt.% were first considered. Based on the initial consumption of lime (ICL) test, 3 wt.% was identified as the optimum value to provide sufficient alkalinity (pH = 12.37 > 12) [32]. Therefore, this value was selected for the combined NS-QL mix designs. Using a constant QL content also allows a clearer evaluation of the effect of NS, while also supporting more efficient use of QL in the investigated mix designs.

A total of 180 individual laboratory records were used for ML model development. These records were obtained from 36 mix design-curing combinations, with five independent replicate specimens tested for each combination. The NS content varied from 0 to 20 wt.% at five levels: 0, 5, 10, 15, and 20 wt.%. QL was used either individually at 1, 3, and 5 wt.% or in combination with NS at a constant content of 3 wt.%. The curing periods were 3, 7, and 28 days, representing early, intermediate, and longer-term strength development. All statistical analyses, ML model training, cross-validation, explainability analysis, and uncertainty quantification were performed using the full replicate-level dataset of 180 individual records. The 180 records were not generated by duplicating the averaged values in Table 1; rather, they preserve the specimen-level variability in UCS, Atterberg limits, compaction properties, and mixture response.

Table 1. Summary of mix design-curing combinations and mean experimental results for stabilized kaolin clay

Mix Design	NS	QL	Curing	Atterberg Limits			Compaction Properties		UCS
	(wt.%)	(wt.%)	(days)	PL (%)	LL (%)	PI (%)	OMC (%)	MDD (g/cm^3)	(kPa)
CS	0	0	3	13.38	26.00	12.62	15.89	1.83	427.7
	0	0	7	13.38	26.00	12.62	15.89	1.83	431.5
	0	0	28	13.38	26.00	12.62	15.89	1.83	446.9
NS5	5	0	3	20.62	29.80	9.18	16.72	1.79	459.3
	5	0	7	21.18	29.42	8.24	16.72	1.79	708.8
	5	0	28	22.02	28.90	6.88	16.72	1.79	1293.0
NS10	10	0	3	22.50	30.13	7.63	17.22	1.75	515.0
	10	0	7	22.80	30.00	7.20	17.22	1.75	698.8
	10	0	28	23.67	29.37	5.70	17.22	1.75	1550.8
NS15	15	0	3	23.64	31.45	7.81	17.45	1.70	521.8
	15	0	7	23.89	30.87	6.98	17.45	1.70	803.7
	15	0	28	24.80	29.65	4.85	17.45	1.70	1621.7
NS20	20	0	3	21.54	30.42	8.88	18.13	1.68	488.1
	20	0	7	22.35	30.00	7.65	18.13	1.68	612.9
	20	0	28	23.20	28.77	5.57	18.13	1.68	1365.0
QL1	0	1	3	20.03	26.86	6.83	15.97	1.72	525.7
	0	1	7	20.11	26.97	6.86	15.97	1.72	633.5
	0	1	28	20.34	27.08	6.74	15.97	1.72	825.2
QL3	0	3	3	20.00	27.00	7.00	16.35	1.67	604.5
	0	3	7	20.72	27.18	6.46	16.35	1.67	932.0
	0	3	28	21.09	27.38	6.29	16.35	1.67	1056.3
QL5	0	5	3	20.84	27.11	6.27	16.54	1.63	899.0
	0	5	7	21.38	27.43	6.05	16.54	1.63	1068.6
	0	5	28	22.20	27.93	5.73	16.54	1.63	1304.8
QL3NS5	5	3	3	20.50	28.11	7.61	16.90	1.71	573.8
	5	3	7	21.80	28.77	6.97	16.90	1.71	857.6
	5	3	28	24.05	29.67	5.62	16.90	1.71	1700.4
QL3NS10	10	3	3	21.03	28.13	7.10	17.39	1.65	616.0
	10	3	7	23.82	29.54	5.72	17.39	1.65	943.5
	10	3	28	25.63	30.63	5.00	17.39	1.65	1910.2
QL3NS15	15	3	3	21.78	28.63	6.85	17.84	1.62	683.3
	15	3	7	24.50	29.20	4.70	17.84	1.62	1014.2
	15	3	28	26.38	30.28	3.90	17.84	1.62	2038.2
QL3NS20	20	3	3	22.30	29.05	6.75	18.22	1.60	539.0
	20	3	7	23.80	29.76	5.96	18.22	1.60	907.2
	20	3	28	25.89	30.31	4.42	18.22	1.60	1603.2

Note: Values represent the average of five independent replicate specimens for each condition. The ML models were trained and evaluated using the full replicate-level dataset containing 180 individual records.

As shown in Table 1, the addition of NS and QL improved the strength and modified the geotechnical properties of the treated kaolin clay. Plasticity decreased with increasing stabilizer content, while OMC increased and MDD decreased due to changes in soil structure. The UCS increased with binder content and curing days up to an optimum level. The highest UCS was obtained for QL3NS15 mix design at 28 days, with an increase of approximately 4.8 times higher than the untreated control sample. The values shown in Table 1 are mean values of five replicate specimens and are presented only to summarize the experimental trends. The replicate-level measurements were retained in the modeling dataset.

2.2. Exploratory Data Characteristics and Preprocessing

2.2.1. Definition of Input and Output Variables

The input variables were selected from the experimental dataset to represent the main factors controlling the behavior of stabilized clay. Nano shale (NS) content and quicklime (QL) content were included as the primary stabilizing materials, governing the binder dosage and chemical reactions. Curing days were considered to capture the time-dependent development of strength due to chemical or pozzolanic reactions. The Atterberg limit parameters, including liquid limit (LL), plastic limit (PL), and plasticity index (PI), were initially considered to describe the consistency and plasticity characteristics of the untreated and stabilized soils. These parameters are relevant because changes in plasticity reflect modifications in soil fabric, water affinity, and particle interaction caused by stabilization. Compaction properties, namely optimum moisture content (OMC) and maximum dry density (MDD), were included to represent the physical state of the soil and its compaction characteristics. The output variable of the study is the unconfined compressive strength (UCS), which is a key indicator of soil strength and stabilization performance. At this exploratory stage, all available variables were retained to examine their distributions, relationships with UCS, and possible redundancy before defining the final input set for ML model development.

2.2.2. Exploratory Data Analysis

The relationships between the experimental variables and UCS were first examined using joint distribution plots, as shown in Figure 2. A clear positive trend is observed between UCS and curing days, indicating the strong influence of time-dependent lime activation, pozzolanic reactions, and progressive soil-binder bonding. In contrast, UCS shows a clear decreasing tendency with increasing plasticity index (PI), indicating that lower plasticity is generally associated with higher strength in the stabilized mixtures. The relationships between UCS and binder contents (NS and QL) are distinctly nonlinear and mix-dependent. In particular, UCS increases with NS content up to an optimum level (around 15 wt.%) and then decreases, suggesting the existence of an optimum NS range rather than a monotonic response. This behavior is consistent with the experimental observation that the QL3NS15 mixture provided the highest UCS after 28 days of curing. The UCS-QL relationship also shows a nonlinear pattern because QL was used both as an individual stabilizer and as a fixed alkaline activator in the combined NS-QL mixtures. These trends, together with the considerable scatter observed in the joint plots, indicate that the strength response cannot be adequately described by simple linear relationships alone.

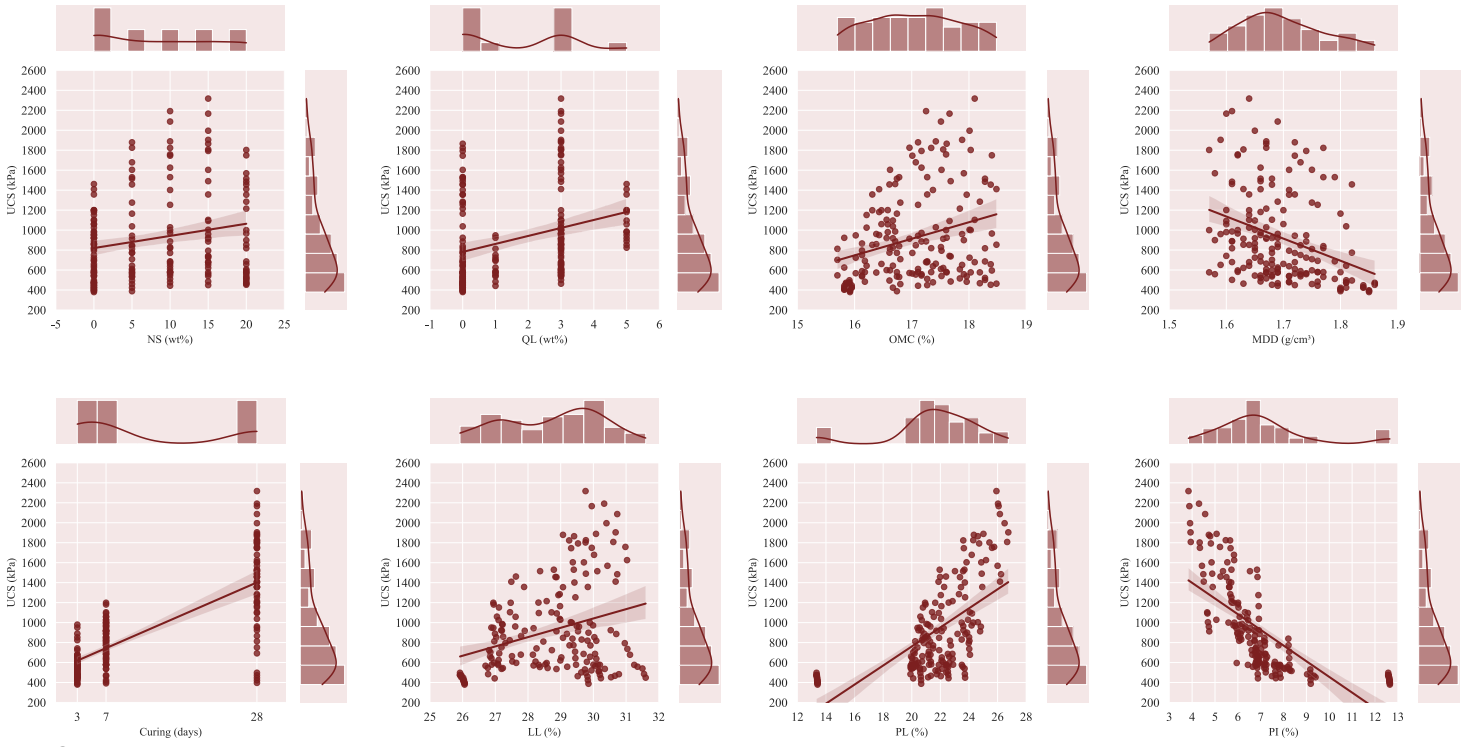


Figure 2. Joint distributions of UCS versus input variables for the full experimental dataset (n = 180)

Compaction-related variables also show distinct patterns. UCS tends to increase with OMC but decrease with MDD. This apparent negative association between UCS and MDD should not be interpreted as a direct detrimental effect of dry density on strength. Instead, it reflects the coupled changes caused by stabilization, where the addition of NS and QL increases water demand and reduces compacted density while simultaneously promoting flocculation, pozzolanic bonding, and cementitious structure formation. Therefore, the observed UCS-MDD relationship is mainly an indirect consequence of mixture modification and binder-induced structural changes.

The correlation structure among the variables was further examined using Pearson and Spearman correlation matrices, as shown in Figure 3. Pearson correlation measures linear association, whereas Spearman correlation evaluates monotonic association based on rank order. Therefore, comparing both matrices provides useful insight into whether the observed relationships are mainly linear or nonlinear/monotonic.

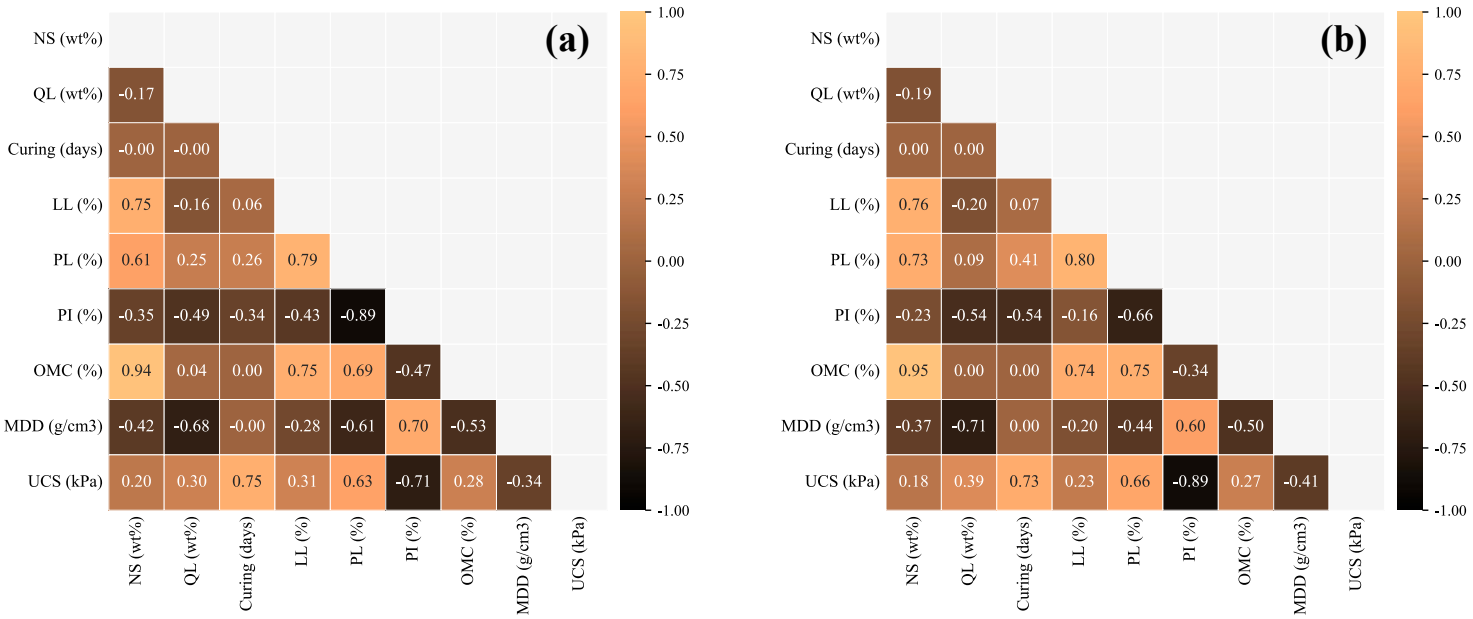


Figure 3. Correlation matrices of variables using: **(a)** Pearson; and **(b)** Spearman methods

A strong positive correlation is observed between UCS and curing days, with Pearson and Spearman coefficients of r = +0.75 and r_s = 0.73, respectively. This confirms the dominant role of curing days in strength development. In contrast, UCS exhibits a strong negative correlation with PI (Pearson r = −0.71, Spearman r_s = −0.89), indicating that plasticity reduction is strongly linked to strength improvement. The stronger Spearman correlation suggests that the UCS-PI relationship is more monotonic than strictly linear. Among the Atterberg limit parameters, LL, PL, and PI show substantial interdependence. LL and PL are strongly positively correlated (Pearson r = +0.79, Spearman r_s = +0.80), while PI is inherently related to both parameters through PI = LL − PL. Therefore, including all three Atterberg limit variables may introduce redundant information into the feature space. To reduce unnecessary model complexity and avoid redundancy, PL was excluded from the final input variables, while LL and PI were retained to represent soil consistency and plasticity.

In addition, NS exhibits a strong positive correlation with OMC (Pearson r = +0.94, Spearman r_s = +0.95), reflecting the increase in water demand with higher stabilizer content. A moderate negative correlation is observed between UCS and MDD (Pearson r = −0.34, Spearman r_s = −0.41), indicating that, in this system, a reduction in dry density is associated with an increase in strength. As discussed above, this trend reflects the structural and compaction changes induced by stabilization rather than a simple direct relationship between density and strength.

As a result, the exploratory analysis reveals that UCS is controlled by nonlinear, interdependent, and mixture-specific relationships among curing days, stabilizer contents, plasticity, and compaction characteristics. These observations justify the use of ML models capable of capturing nonlinear interactions and also motivate the development of physics-guided engineered features in the subsequent sections.

2.2.3. Data Cleaning and Outlier Control

Before model development, the dataset was examined for consistency, completeness, and potential anomalous observations. No missing values were found in the compiled database, and all records were obtained from controlled laboratory tests reported in the source experimental study. As described in Section 2.1, Table 1 reports averaged values only for compact presentation of the experimental program. However, all model development, preprocessing, cross-validation, explainability analysis, and uncertainty analysis were performed using the full replicate-level dataset containing 180 individual records. Therefore, the data cleaning and outlier screening procedures were applied to the individual replicate-level observations rather than to the averaged values reported in Table 1.

Potential outliers were examined using descriptive statistics, joint distribution plots, and correlation analysis. A few relatively high UCS values were observed, particularly for mixtures with longer curing days and higher binder effectiveness. However, these observations were consistent with the expected strength gain of stabilized soils and corresponded to physically meaningful experimental responses rather than evident measurement errors. Therefore, no data points were removed solely based on statistical distance or extreme magnitude.

Retaining these observations is important because the objective of the model is to learn the full range of behavior exhibited by the stabilized kaolin clay, including high-strength responses associated with favorable binder combinations and curing conditions. Removing such points could artificially narrow the response domain and reduce the ability of the model to capture nonlinear strength development. Accordingly, all experimentally valid records were retained for subsequent preprocessing, feature engineering, and ML model development.

2.2.4. Validation Strategy and Leakage Control

The statistical characteristics of the full experimental dataset are summarized in Table 2. The dataset contains 180 individual replicate-level records covering a broad range of mixture compositions, curing days, Atterberg limits, compaction properties, and UCS values. NS content ranges from 0 to 20 wt.% with a mean value of 8.33 wt.%, while QL content varies from 0 to 5 wt.% with a mean of 1.75 wt.%. The curing period includes 3, 7, and 28 days, resulting in a mean value of 12.67 days. These ranges indicate that the dataset includes untreated samples, individually stabilized mixtures, and combined NS-QL mixtures over different curing stages.

Table 2. Descriptive statistics of the full experimental dataset (n = 180)

Variable	Count	Mean	SD	Min.	Q1	Median	Q3	Max.	IQR	Skewness	Kurtosis
NS (wt.%)	180	8.33	7.47	0.00	0.00	7.50	15.00	20.00	15.00	+0.282	-1.37
QL (wt.%)	180	1.75	1.69	0.00	0.00	2.00	3.00	5.00	3.00	+0.294	-1.25
Curing (days)	180	12.67	11.00	3.00	3.00	7.00	28.00	28.00	25.00	+0.643	-1.51
PL (%)	180	21.68	2.54	17.06	20.25	22.05	23.75	25.55	3.50	-0.358	-1.03
LL (%)	180	28.72	1.51	25.92	27.22	28.99	29.89	31.60	2.67	-0.242	-1.02
PI (%)	180	7.03	2.06	3.84	5.83	6.81	7.62	12.65	1.80	+1.431	+2.11
OMC (%)	180	17.05	0.78	15.70	16.46	17.10	17.62	18.48	1.16	+0.060	-1.05
MDD (g/cm^3)	180	1.70	0.07	1.57	1.65	1.69	1.74	1.86	0.10	+0.432	-0.54
UCS (kPa)	180	921.70	458.04	379.30	567.10	799.85	1168.98	2317.30	601.88	+1.027	+0.19

The Atterberg limit parameters show moderate variability. LL varies from 25.92% to 31.60%, with a mean of 28.72%, while PL ranges from 17.06% to 25.55%, with a mean of 21.68%. PI has a mean value of 7.03% and ranges from 3.84% to 12.65%, indicating that the stabilized mixtures mostly remain within a low-plasticity range. Compared with LL and PL, PI exhibits stronger positive skewness (+1.431) and higher kurtosis (+2.11), reflecting the presence of a limited number of mixtures with relatively higher plasticity. The compaction variables also show systematic variation, with OMC ranging from 15.70% to 18.48% and MDD ranging from 1.57 to 1.86 g/cm^3. The relatively narrow but meaningful ranges of OMC and MDD reflect the controlled laboratory preparation of the mixtures while still capturing the compaction changes induced by stabilization.

The UCS values range from 379.30 to 2317.30 kPa, with a mean of 921.70 kPa and a standard deviation of 458.04 kPa. This relatively large spread confirms that the dataset captures substantial strength variation caused by changes in stabilizer content and curing days. The UCS distribution is positively skewed (+1.027), indicating that most samples are concentrated in the low-to-moderate strength range, while a smaller number of high-strength mixtures occur at favorable NS-QL combinations and longer curing days. The interquartile range of UCS is 601.88 kPa, further confirming the broad variability of strength response within the experimental program.

For model evaluation, a nested GroupKFold cross-validation strategy was adopted instead of relying on a single random train-test split. This choice is consistent with recent geotechnical ML studies that used nested cross-validation to obtain more reliable performance estimates for relatively small UCS datasets [20] and with the broader emphasis on cross-validation and robustness assessment in stabilized-soil strength prediction [33]. In the present study, the outer loop was used for model performance evaluation, while the inner loop was used for hyperparameter optimization.

To reduce replicate-level information leakage, each record was assigned to a group based on the combination of mix design and curing days. During cross-validation, samples belonging to the same group were kept within the same fold and were not split between training and testing subsets. In practical terms, all five replicate specimens corresponding to the same mix design-curing condition were assigned to the same group. Therefore, no replicate from a given mixture-curing condition was allowed to appear in the training set while another replicate from the same condition appeared in the test set. This grouping strategy was used because records from the same mixture and curing condition are experimentally related and may share similar material characteristics. Therefore, the adopted validation procedure provides a more realistic assessment of model generalization for unseen mixture-curing combinations.

All preprocessing steps, feature scaling where required, physics-guided feature construction, hyperparameter optimization, and model training were performed within the cross-validation workflow. This ensured that information from the outer test folds was not used during model tuning or preprocessing. The final reported performance metrics therefore represent the mean and standard deviation across the outer folds rather than the result of a single data partition.

2.2.5. Cross-Validation Workflow and Feature Scaling

All preprocessing and model-training steps were implemented within the nested cross-validation workflow described in Section 2.2.4. The outer loop consisted of a 5-fold GroupKFold cross-validation and

was used exclusively for final model evaluation. The inner loop consisted of a 3-fold GroupKFold cross-validation and was applied only to the outer-training portion for PSO-based hyperparameter optimization. In each outer fold, PSO searched the predefined hyperparameter space by minimizing the mean validation RMSE across the inner folds. The selected hyperparameters were then used to retrain the model on the full outer-training data, and the trained model was evaluated once on the held-out outer-test groups. The outer-test groups were not used during hyperparameter tuning, model selection, or performance optimization. This structure follows the same rationale adopted in recent UCS prediction studies, where nested cross-validation has been used to obtain more reliable performance estimates for relatively small geotechnical datasets and to reduce model-selection bias [20]. Similar cross-validation procedures have also been recommended in stabilized-soil strength prediction studies to reduce overfitting and improve robustness assessment [33].

Within each fold, preprocessing parameters were fitted only on the training portion and then applied to the corresponding validation or test portion. This fold-wise implementation prevents information from the held-out data from influencing feature scaling, feature transformation, or model fitting. The same principle was applied to both raw input variables and physics-guided engineered features, ensuring that all transformations were learned exclusively from the training data within each cross-validation iteration.

Feature scaling was performed using the Standard Scaler for models sensitive to the magnitude of input variables, including SVR, MLR, and BP-ANN. Standardization was performed using the training-fold mean and standard deviation as:

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

where X_i is the original value of the feature, and μ and σ represent the mean and standard deviation of the feature computed from the training fold. This transformation standardizes the features to zero mean and unit variance, improving numerical stability and convergence of the models.

Tree-based models, including RF, XGBoost, and CatBoost, are generally less sensitive to feature scaling because their splitting rules depend on feature ordering rather than absolute magnitude. Therefore, scaling was not required for these models from a methodological perspective. However, all models were implemented using a consistent pipeline structure to ensure reproducibility and to avoid accidental information leakage across folds. Because UCS values showed a positively skewed distribution, a logarithmic transformation was applied to the target variable during model training to improve numerical stability and reduce the influence of high-strength samples. Model predictions were then back-transformed to the original UCS scale before calculating the final evaluation metrics. RMSE was used as the primary optimization objective during hyperparameter tuning because it penalizes large prediction errors and is directly expressed in UCS units after inverse transformation.

The final model performance was summarized using the mean and standard deviation of the evaluation metrics across the outer folds. This reporting format provides information on both predictive accuracy and fold-to-fold variability, allowing the stability of each model to be assessed in addition to its average performance. The relatively small and grouped nature of the dataset motivated the use of this structured methodology. Nested GroupKFold validation was used to obtain conservative performance estimates,

ablation analysis was used to avoid unnecessary engineered-feature complexity, explainability methods were used to check whether the model learned physically meaningful trends, and uncertainty analysis was used to avoid relying only on deterministic predictions. Therefore, each methodological component was included to address a specific reliability, interpretability, or practical-use issue associated with small laboratory datasets.

2.2.6. Computational Environment

All models and analyses in this study were implemented in Python (version 3.13.13) using a range of well-established libraries. The computational environment included tools for machine learning, data processing, numerical computation, visualization, and model interpretability. Table 3 summarizes the complete details of the computational environment, including library versions and corresponding tools.

Table 3. Computational environment and Python libraries used in this study

Component / Model	Library / Software	Version	Class / Tool	Purpose
MLR	Scikit-learn	1.6.1	LinearRegression	Linear benchmark model
SVR	Scikit-learn	1.6.1	SVR	Kernel-based regression
RF	Scikit-learn	1.6.1	RandomForestRegressor	Ensemble regression
XGBoost	Xgboost	3.1.3	XGBRegressor	Gradient boosting model
CatBoost	Catboost	1.2.8	CatBoostRegressor	Gradient boosting model
BP-ANN	Scikit-learn	1.6.1	MLPRegressor	Neural network model
Feature scaling	Scikit-learn	1.6.1	StandardScaler	Scaling for MLR, SVR, and BP-ANN
PSO	Custom	-	PSOOptimizer	Hyperparameter optimization
Numerical computation	NumPy	2.0.2	-	Matrix and array operations
Data handling	Pandas	2.2.2	-	Data processing
Visualization	Matplotlib	3.10.0	-	Plotting
SHAP	SHAP	0.50.0	-	Model explainability
Python	Python	3.13.13	-	Programming environment
Notebook	Visual Studio Code	7.5.3	-	Model development

2.3. Physics-Guided Feature Engineering

To incorporate geotechnical domain knowledge into the modeling framework, a set of physics-guided engineered features was constructed based on the main mechanisms controlling the strength development of stabilized clay. These features were designed to represent interactions among stabilizer content, curing days, plasticity, and compaction state that are not explicitly captured by the original input variables alone. In this study, the term physics-guided refers specifically to mechanism-inspired feature engineering based on geotechnical understanding of stabilization processes. The physical guidance is introduced at the input-feature level through descriptors related to binder dosage, curing-dependent reactions, compaction state, and relative plasticity. Therefore, the proposed framework should not be interpreted as a physics-informed neural network, a physics-constrained loss formulation, or a model that directly enforces governing equations.

Initially, seven engineered features were defined, as summarized in Table 4. The total binder content, expressed as NS + QL, was introduced to represent the overall dosage of reactive stabilizing materials. Interaction terms between stabilizers and curing days were included to account for the time-dependent evolution of pozzolanic reactions, alkaline activation, and soil-binder bonding. In particular, NS × Curing, QL × Curing, and (NS + QL) × Curing were used to describe the combined effects of binder availability and curing duration on strength development.

Table 4. Physics-guided engineered features

Feature	Expression	Physical meaning
Binder_Total	NS + QL	Total binder content
Binder Total_Curing	(NS + QL) × Curing	Time-dependent binder effect
NS_Curing	NS × Curing	Pozzolanic contribution of NS over time
QL_Curing	QL × Curing	Quicklime activation over time
log_Curing	log(Curing)	Nonlinear strength development
MDD_OMC_Ratio	MDD / OMC	Compaction state indicator
PI_LL_Ratio	PI / LL	Relative plasticity measure

Because the strength gain of stabilized soils is typically nonlinear with curing days, a logarithmic transformation of curing days was also included. In addition, MDD/OMC was introduced as a compaction-state indicator, reflecting the combined influence of dry density and water demand. The PI/LL ratio was used as a relative plasticity indicator to describe the contribution of plasticity with respect to the overall liquid limit. These ratio-based features were included to provide normalized descriptors of soil physical state rather than relying only on absolute index values.

Following the initial construction of these seven engineered features, their usefulness was further examined through the ablation analysis presented in Section 3. The ablation results showed that QL_Curing introduced redundant information and did not improve the generalization performance of the model. Therefore, this feature was removed from the final physics-guided feature set. The final CatBoost-6PG model retained six engineered features: Binder_Total, NS_Curing, BinderTotal_Curing, log_Curing, MDD_OMC_Ratio, and PI_LL_Ratio. These features were used together with the original input variables in the final modeling stage.

2.4. Machine Learning Model Development

A set of ML models with different mathematical structures and learning mechanisms was developed to predict the UCS of stabilized kaolin clay. The selected models include a linear benchmark model, a kernel-based nonlinear model, ensemble tree models, gradient boosting models, and a neural-network model. This selection enables a balanced comparison between simple interpretable approaches and more flexible nonlinear algorithms. All models were trained and evaluated within the nested GroupKFold framework described in Section 2.2.4, and hyperparameters were optimized using PSO in the inner cross-validation loop.

2.4.1. Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) was used as a baseline model to establish a linear relationship between input variables and UCS. The model can be expressed as:

$$UCS = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \quad (2)$$

where β_0 is the intercept, β_i are regression coefficients, and x_i are the input features. Despite its simplicity, MLR provides a useful reference for evaluating the improvement achieved by more advanced models.

2.4.2. Support Vector Regression (SVR)

Support Vector Regression (SVR) is a kernel-based method that maps input data into a higher-dimensional space to capture nonlinear relationships [34]. The SVR function is defined as:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (3)$$

where $K(\cdot)$ is the kernel function, and α_i, α_i^* are Lagrange multipliers.

In this study, the radial basis function (RBF) kernel was used, which is defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (4)$$

where γ is a kernel parameter controlling the influence of individual training samples.

The RBF kernel was selected due to its strong capability to model complex and nonlinear relationships without requiring prior assumptions about the data structure. This is particularly important for the present dataset, where the relationships between input variables and UCS are highly nonlinear, as demonstrated in the exploratory data analysis (Section 2.2.2). By mapping the data into an infinite-dimensional feature space, the RBF kernel allows SVR to effectively capture local variations and interaction effects between variables.

2.4.3. Random Forest (RF)

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees using bootstrapped samples and random feature selection [35]. The final prediction is obtained by averaging the outputs of individual trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} y_t \quad (5)$$

where T is the number of trees. This strategy reduces variance and improves robustness, making RF effective for modeling nonlinear relationships.

2.4.4. Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced gradient boosting algorithm that builds decision trees sequentially by minimizing a regularized objective function [36]:

$$Obj = \sum_{i=1}^{n} L(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

where $L(\cdot)$ is the loss function and $\Omega(f_k)$ is a regularization term that penalizes model complexity.

Unlike traditional boosting methods, XGBoost uses second-order gradient information to improve optimization accuracy and convergence speed. It also incorporates regularization, tree pruning, and shrinkage mechanisms to prevent overfitting. These features allow XGBoost to effectively capture complex nonlinear interactions between input variables while maintaining high generalization performance.

2.4.5. Categorical Boosting (CatBoost)

CatBoost is a gradient boosting algorithm specifically designed to reduce prediction bias and overfitting through an ordered boosting strategy [37]. Unlike conventional boosting methods, CatBoost constructs trees using permutations of the dataset, which helps avoid target leakage during training. The model iteratively minimizes a loss function by adding decision trees:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \quad (7)$$

where $F_m(x)$ is the model at iteration m, $h_m(x)$ is the new tree, and η is the learning rate.

CatBoost is particularly effective for small and structured datasets due to its ability to handle feature interactions, reduce overfitting, and provide stable performance without extensive preprocessing. These characteristics make it well-suited for the present dataset.

2.4.6. Backpropagation Artificial Neural Network (BP-ANN)

The BP-ANN is a feedforward neural network capable of modeling complex nonlinear relationships between input variables and UCS. The network consists of an input layer, one or more hidden layers, and an output layer, where each neuron performs a weighted summation followed by a nonlinear activation function [38]. The output of a neuron can be expressed as:

$$y = f\left(\sum_{j=1}^{n} \omega_j x_j + b\right) \quad (8)$$

where ω_j are the connection weights, b is the bias term, and $f(\cdot)$ is the activation function.

During training, the network parameters are updated using the backpropagation algorithm, which minimizes the prediction error by propagating gradients backward through the network. The weights are iteratively updated using gradient-based optimization:

$$\omega^{(t+1)} = \omega^{(t)} - \eta \frac{\partial L}{\partial \omega} \quad (9)$$

where η is the learning rate and L is the loss function.

BP-ANN is particularly suitable for problems with complex and highly nonlinear interactions, as it can approximate arbitrary nonlinear functions. However, its performance depends on network architecture and hyperparameter tuning, and it may be prone to overfitting in relatively small datasets [23]. In this study, BP-ANN was used to evaluate the capability of neural networks in capturing the nonlinear behavior of stabilized clay, especially in comparison with tree-based and boosting models.

2.5. Hyperparameter Optimization

To improve model performance and ensure fair comparison, hyperparameter optimization was performed using Particle Swarm Optimization (PSO). PSO is a population-based metaheuristic algorithm inspired by the collective behavior of social organisms such as bird flocks [39,40]. In this study, PSO was implemented as a custom Python routine and embedded within the nested GroupKFold workflow. Each particle represented a candidate hyperparameter configuration, and the optimization objective was to

minimize the validation RMSE obtained from the inner GroupKFold loop. The velocity and position of each particle are updated as follows:

$$v_i^{(t+1)} = \omega v_i^{(t)} + c_1 r_1 \left(p_i - x_i^{(t)}\right) + c_2 r_2 \left(g - x_i^{(t)}\right) \quad (10)$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \quad (11)$$

where v_i and x_i are the velocity and position of particle i, p_i is the personal best position, g is the global best position, ω is the inertia weight, and c_1, c_2 are cognitive and social coefficients, respectively.

In each outer fold, PSO was applied only to the corresponding outer-training data. The inner GroupKFold loop was used to evaluate candidate hyperparameter sets, and the configuration with the lowest mean validation RMSE was selected. The optimized model was then retrained on the full outer-training fold and evaluated on the held-out outer-test fold. This procedure separates hyperparameter selection from final model evaluation and reduces model-selection bias.

The search ranges and PSO-selected hyperparameters are summarized in Table 5. Values in the last column represent the mean and standard deviation of the selected hyperparameters across the five outer folds. For categorical or conditional parameters, the most frequently selected option is reported.

Table 5. PSO hyperparameter search ranges and selected values across the outer folds

Model	Hyperparameter	Lower bound	Upper bound	Selected value across outer folds
SVR	C	1	5000	2858.46 ± 2064.14
	epsilon	0.001	0.2	0.0860 ± 0.0803
	gamma	0.01	5	0.0301 ± 0.0234
RF	n_estimators	300	900	658.8 ± 168.9
	max_depth	6	30	15.6 ± 9.7
	max_features	0.4	0.9	0.6039 ± 0.2109
	min_samples_split	2	12	2.8 ± 1.8
	min_samples_leaf	1	5	1.8 ± 1.8
	max_leaf_nodes	30	200/None	129.5 ± 29.1; None in 1/5 folds
	bootstrap	fixed	fixed	True in 5/5 folds
XGBoost	n_estimators	400	900	695.0 ± 244.9
	learning_rate	0.08	0.18	0.1326 ± 0.0478
	max_depth	4	6	5.4 ± 0.9
	subsample	0.65	0.85	0.7509 ± 0.0560
	colsample_bytree	0.75	1	0.8846 ± 0.1155
	min_child_weight	1	5	3.5235 ± 1.9567
	reg_lambda	8	20	13.2601 ± 5.2149
	reg_alpha	0	3	0.0476 ± 0.1065
	gamma	0	1.5	0.0000 ± 0.0000
CatBoost	iterations	300	1200	831.8 ± 254.4
	learning_rate	0.01	0.15	0.1142 ± 0.0335
	depth	4	10	4.0 ± 0.0
	l2_leaf_reg	3	20	6.6221 ± 4.9227
	random_strength	0	5	2.4025 ± 1.8094
	bagging_temperature	0	5	1.3329 ± 1.8793
BP-ANN	hidden_layer_sizes	8	64	25.6 ± 9.8; two-layer in 1/5 folds

	activation	relu / tanh	relu / tanh	tanh in 4/5 folds
	alpha	0.001	1	0.1404 ± 0.0892
	learning_rate_init	0.0001	0.01	0.0065 ± 0.0021
	batch_size	8	32	22.2 ± 6.3
CatBoost-7PG	iterations	300	1000	585.2 ± 264.9
	learning_rate	0.01	0.12	0.1152 ± 0.0107
	depth	3	6	3.2 ± 0.4
	l2_leaf_reg	5	40	8.4049 ± 6.1621
	random_strength	1	10	6.5523 ± 3.4858
	bagging_temperature	0	8	4.4186 ± 3.0952
CatBoost-6PG	iterations	300	1000	752.4 ± 230.3
	learning_rate	0.01	0.12	0.1130 ± 0.0102
	depth	3	6	3.8 ± 0.8
	l2_leaf_reg	5	40	15.3381 ± 13.1745
	random_strength	1	10	5.7927 ± 3.8172
	bagging_temperature	0	8	1.5567 ± 3.3898

Note: Values in the last column represent the mean ± standard deviation of the PSO-selected hyperparameters across the five outer folds of nested GroupKFold cross-validation. For categorical parameters, the most frequent selected value is reported. For RF, max_leaf_nodes = None was selected in one outer fold, indicating that no explicit maximum leaf-node constraint was applied in that fold.

2.6. Model Evaluation Metrics

The performance of the developed models was evaluated using several statistical metrics to provide a comprehensive assessment of prediction accuracy and error characteristics. These metrics include the coefficient of determination (R^2), mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and variance accounted for (VAF).

The coefficient of determination (R^2) is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{12}$$

and represents the proportion of variance in the observed data explained by the model. Higher R^2 values indicate better performance, with a value of 1 corresponding to a perfect fit.

The mean squared error (MSE) is expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{13}$$

which measures the average squared difference between predicted and observed values. Lower MSE values indicate better performance, with values closer to zero representing more accurate predictions.

The root mean square error (RMSE) is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{14}$$

and represents the standard deviation of prediction errors in the same unit as UCS. Similar to MSE, lower RMSE values indicate better model accuracy, and values approaching zero correspond to improved predictions.

The mean absolute error (MAE) is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{15}$$

and quantifies the average magnitude of prediction errors. Lower MAE values indicate better performance and reflect smaller average deviations between predicted and actual values.

The mean absolute percentage error (MAPE) is calculated as:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{16}$$

and expresses the prediction error as a percentage. Lower MAPE values indicate higher predictive accuracy, with values closer to zero representing better model performance.

The variance accounted for (VAF) is defined as:

$$VAF = \left(1 - \frac{Var\left(y_i - \hat{y}_i\right)}{Var\left(y_i\right)}\right) \times 100 \tag{17}$$

and measures the percentage of variance explained by the model. Higher VAF values indicate better performance, with values approaching 100% representing an accurate model.

2.7. Explainability and Sensitivity Analysis

To improve model interpretability and examine whether the learned relationships are consistent with the mechanisms of soil stabilization, explainability and sensitivity analyses were performed after model development. These analyses were applied to the selected best-performing model after completing model comparison and performance evaluation.

Model explainability was investigated using the SHapley Additive exPlanations (SHAP) method, which is based on cooperative game theory [41,42]. SHAP assigns an importance value to each feature by quantifying its contribution to the model output. The prediction of a model can be expressed as an additive explanation model:

$$f(x) = \varphi_0 + \sum_{i=1}^{M}\varphi_i \tag{18}$$

where φ_0 is the expected model output and φ_i represents the contribution of feature i to the prediction.

The SHAP value for each feature is computed as:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f(S \cup \{i\}) - f(S)] \quad (19)$$

where S is a subset of all features, M is the total number of features, and f(S) represents the model prediction using only the features in subset S. This formulation computes the average marginal contribution of each feature across all possible feature combinations, ensuring fair and consistent attribution.

In this study, SHAP analysis was used for both global and local interpretation. Global SHAP analysis was used to rank the relative importance of the input and engineered features in UCS prediction. Local SHAP analysis was used to explain individual predictions by identifying the features that increased or decreased the predicted UCS for a specific sample. For models trained on the logarithmic scale of UCS, SHAP values were interpreted in terms of their contribution to the model output on the log-transformed scale, while final predictions were back-transformed to the original UCS scale for engineering interpretation.

In addition to SHAP analysis, one-way sensitivity analysis was conducted to evaluate the response of model predictions to controlled changes in individual input variables. In this procedure, one variable was varied across its observed range while the remaining variables were kept constant at their median baseline values. The resulting variation in predicted UCS was used to quantify the relative influence of each input variable and to examine whether the model response followed physically meaningful trends.

Two-dimensional partial dependence analysis was also performed to examine interaction effects between selected variables. In particular, 2D partial dependence contour plots were used to visualize the combined influence of stabilizer contents and curing days on predicted UCS. This analysis supports the interpretation of nonlinear interaction effects that cannot be fully captured by one-way sensitivity analysis alone.

The combined use of SHAP, one-way sensitivity analysis, and 2D partial dependence analysis provides complementary insight into model behavior. SHAP quantifies feature contributions at global and local levels, sensitivity analysis evaluates model response to controlled input variations, and 2D partial dependence analysis reveals interaction patterns between key variables. Together, these methods help assess whether the predictive model is not only accurate, but also physically interpretable and consistent with the expected behavior of stabilized kaolin clay.

2.8. Uncertainty Analysis

To evaluate the reliability of the predictive framework, uncertainty analysis was performed after model selection using the same outer-fold validation structure adopted for model assessment. For each outer fold, the model was trained using the corresponding outer-training data, while uncertainty propagation was applied only to the held-out outer-test samples. The statistical properties required for perturbation, including the covariance matrix, feature standard deviations, and valid input bounds, were estimated exclusively from the outer-training data to avoid information leakage.

Input uncertainty was propagated using Monte Carlo simulation. For each outer-test sample, 5000 perturbed realizations were generated using a correlated Gaussian noise structure:

$$x^* = x + \varepsilon \quad (20)$$

$$\varepsilon \sim N\left(0, k^2 \sum \text{train}\right) \quad (21)$$

where x is the original input vector, x^* is the perturbed input vector, k is the perturbation intensity, and $\sum$train is the covariance matrix estimated from the corresponding outer-training fold. Three perturbation levels were considered: low (k = 0.10), moderate (k = 0.25), and high (k = 0.50).

Two uncertainty settings were examined. In the FullInput setting, all raw input variables, including NS, QL, curing days, LL, PI, OMC, and MDD, were perturbed. In the GeoPropertyOnly setting, perturbation was applied only to geotechnical index and compaction properties, namely LL, PI, OMC, and MDD, while NS, QL, and curing days were kept fixed. After each perturbation, the engineered features were recalculated to preserve the functional relationships among the raw and physics-guided variables.

Prediction intervals were constructed from the Monte Carlo prediction distributions using percentile-based 90% and 95% intervals. The quality of the intervals was evaluated using prediction interval coverage probability (PICP), mean prediction interval width (MPIW), prediction interval normalized average width (PINAW), average coverage error (ACE), and coverage width-based criterion (CWC). These metrics were used to assess both the reliability and sharpness of the uncertainty intervals. In addition, model uncertainty was assessed using group-level bootstrap resampling within the outer-training folds. In this procedure, bootstrap samples were generated at the group level rather than at the individual-sample level to preserve the grouped structure of the experimental data. The resulting ensemble of bootstrap models was used to construct prediction intervals for the outer-test samples. This procedure enabled comparison between input-driven uncertainty and model-instability uncertainty under the same group-aware validation framework.

3. Results

3.1. Model Performance and Prediction Accuracy

The predictive performance of the developed models is summarized in Table 6. The reported values represent the mean and standard deviation of the evaluation metrics across the outer folds of the nested GroupKFold cross-validation framework. Therefore, the results reflect model performance on group-held-out test folds rather than on a single random train-test split. The corresponding actual-versus-predicted UCS plots are shown in Figure 4, while the residual diagnostic plots are presented in Figure 5.

Table 6. Outer-fold predictive performance of the developed models based on nested GroupKFold cross-validation

Model	R^2 (Mean ± SD)	MSE (Mean ± SD)	MAE (Mean ± SD)	RMSE (Mean ± SD)	MAPE (%) (Mean ± SD)	VAF (%) (Mean ± SD)
MLR	0.430 ± 0.880	61470.01 ± 44585.45	178.92 ± 59.47	235.30 ± 87.33	23.96 ± 10.95	60.17 ± 54.33
SVR	0.722 ± 0.418	29594.55 ± 20502.23	127.11 ± 45.39	164.87 ± 54.92	15.46 ± 8.92	78.82 ± 28.25
RF	0.562 ± 0.591	48500.98 ± 29933.36	156.69 ± 41.12	212.00 ± 66.68	18.74 ± 9.54	67.11 ± 39.04
XGBoost	0.768 ± 0.165	33446.15 ± 6910.62	136.75 ± 11.72	182.14 ± 18.40	15.16 ± 4.09	82.01 ± 8.32
CatBoost	0.794 ± 0.189	29309.45 ± 13120.02	126.14 ± 31.71	167.06 ± 41.84	14.64 ± 4.85	83.49 ± 11.81
BP-ANN	0.808 ± 0.101	30303.45 ± 8742.45	129.24 ± 19.83	172.61 ± 25.26	14.07 ± 0.47	83.84 ± 8.07
CatBoost-7PG	0.795 ± 0.163	28912.45 ± 7419.98	122.81 ± 14.05	169.03 ± 20.69	13.43 ± 3.83	84.56 ± 7.86
CatBoost-6PG	0.841 ± 0.123	23809.63 ± 10324.12	110.67 ± 26.04	151.11 ± 34.94	11.97 ± 3.75	87.38 ± 7.24

The results show a clear difference between the linear benchmark model and the nonlinear models. MLR produced the weakest overall performance, with R^2 = 0.430 ± 0.880, RMSE = 235.30 ± 87.33 kPa,

and MAPE = 23.96 ± 10.95%. The large standard deviation of R^2 also indicates substantial fold-to-fold variability, suggesting that a simple linear relationship is not sufficient to represent the nonlinear strength development of NS-QL stabilized kaolin clay.

Among the conventional nonlinear models, SVR considerably improved the prediction accuracy compared with MLR, reducing RMSE to 164.87 ± 54.92 kPa and increasing R^2 to 0.722 ± 0.418. However, the relatively large standard deviations of SVR indicate that its performance was still sensitive to the held-out group composition. RF showed weaker generalization than SVR, with RMSE = 212.00 ± 66.68 kPa and R^2 = 0.562±0.591, indicating that the bagging-based tree ensemble was less effective for this grouped dataset.

The boosting-based models generally provided more stable and accurate predictions. XGBoost achieved R^2 = 0.768 ± 0.165 and RMSE = 182.14 ± 18.40 kPa, while the baseline CatBoost model improved the RMSE to 167.06 ± 41.84 kPa and achieved R^2 = 0.794 ± 0.189. BP-ANN also performed competitively, with R^2 = 0.808 ± 0.101, RMSE = 172.61 ± 25.26 kPa, and MAPE = 14.07 ± 0.47%. However, its RMSE remained higher than that of the best physics-guided model.

The effect of physics-guided feature engineering is observed by comparing the baseline CatBoost, CatBoost-7PG, and CatBoost-6PG models. CatBoost-7PG, which included the initial seven engineered features, achieved R^2 = 0.795 ± 0.163, RMSE = 169.03 ± 20.69 kPa, and MAPE = 13.43 ± 3.83%. Although this model slightly reduced MAE and MAPE compared with the baseline CatBoost, its RMSE did not improve. After removing the redundant QL_Curing feature through ablation analysis, CatBoost-6PG achieved the best overall performance, with the highest R^2 = 0.841 ± 0.123, the lowest RMSE = 151.11 ± 34.94 kPa, the lowest MAE = 110.67 ± 26.04 kPa, and the lowest MAPE = 11.97 ± 3.75%. The corresponding VAF value of 87.38 ± 7.24% also indicates that CatBoost-6PG explained the largest proportion of UCS variance among the evaluated models. Nevertheless, the standard deviations of the outer-fold metrics indicate that prediction performance still varied across held-out mixture-curing groups. This variability is expected because the dataset is relatively small and some outer folds may contain high-strength mixtures or nonlinear mixture responses that are more difficult to predict. Therefore, the reported mean performance should be interpreted together with the fold-to-fold variability rather than as a single fixed accuracy value.

Figure 4 provides a visual comparison of actual and predicted UCS values for all models. The MLR predictions show the largest scatter around the ideal 1:1 line, particularly at higher UCS values, confirming the limited ability of the linear model to capture nonlinear strength behavior. RF also shows noticeable dispersion and wider deviations from the ideal line. In contrast, SVR, XGBoost, CatBoost, and BP-ANN show improved clustering around the ideal prediction line. The CatBoost-6PG model exhibits the closest overall agreement with the 1:1 line, particularly across the low-to-intermediate UCS range where most samples are concentrated. The ±20% reference bounds further show that CatBoost-6PG retains a larger proportion of predictions within an acceptable error band compared with the other models.

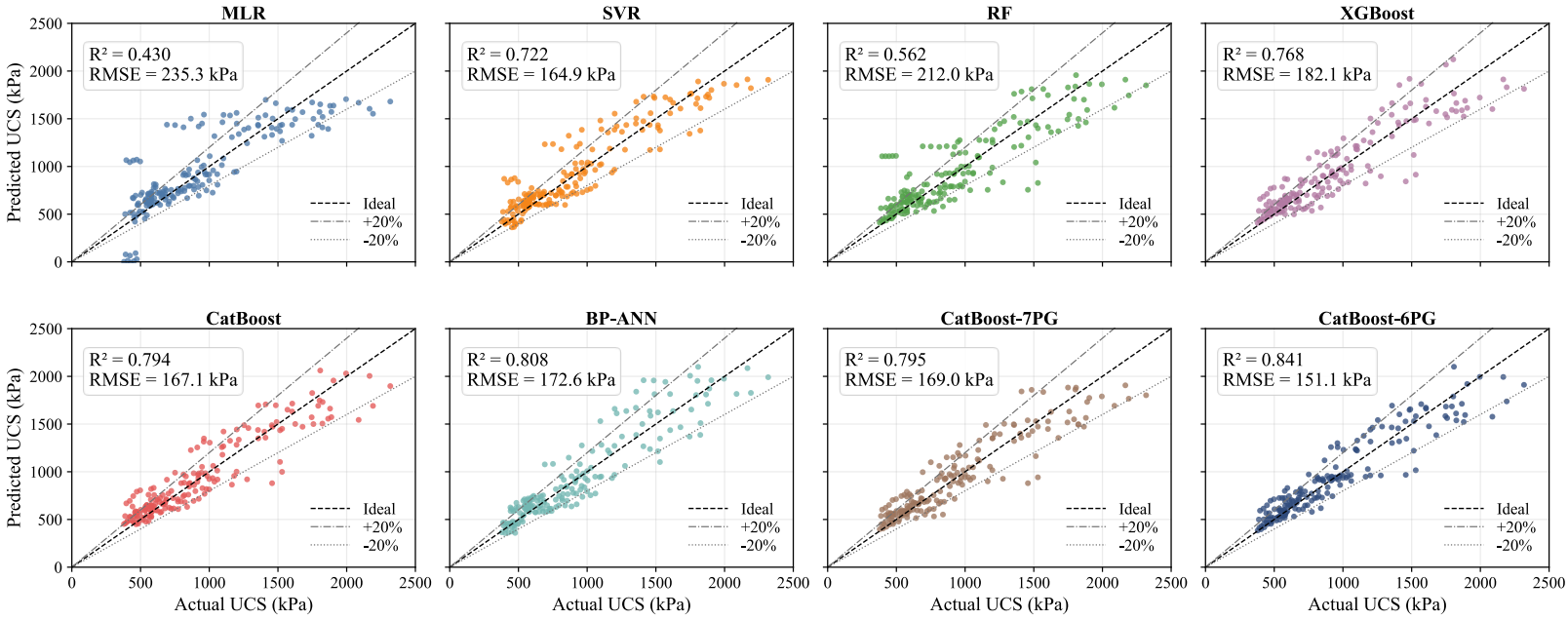


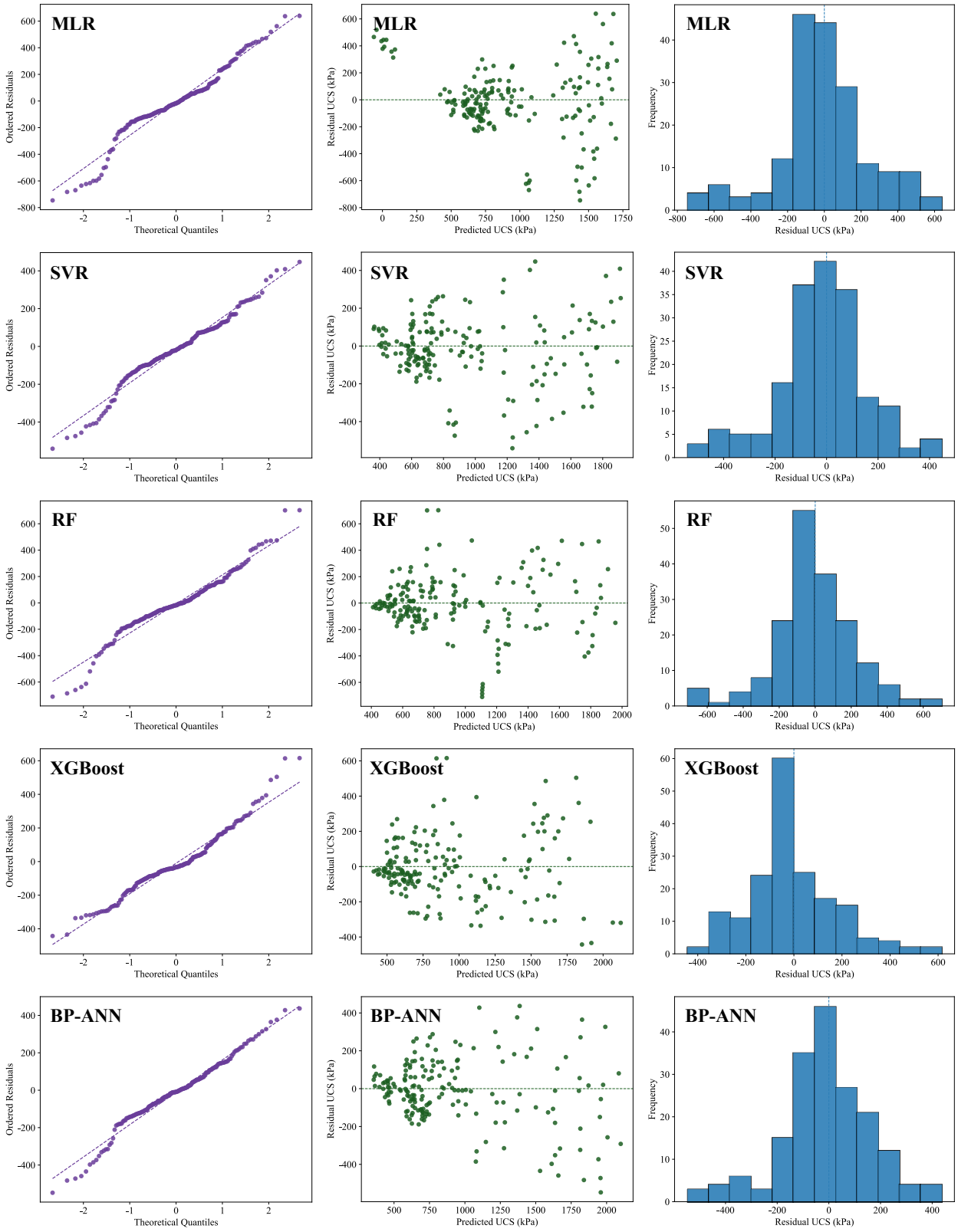
Figure 4. Actual versus predicted UCS values for the developed models using outer-fold test predictions from nested GroupKFold cross-validation

The residual diagnostics in Figure 5 provide a more detailed comparison of model behavior across different error characteristics. For the MLR model, the residuals are widely scattered around zero, with values extending approximately from about −700 to +600 kPa. This wide range indicates high variability in prediction errors and confirms the limited ability of the linear model to capture the nonlinear UCS response. The ordered residuals versus theoretical quantiles plot, also known as the Q–Q plot, shows noticeable deviations from the reference line at the tails, suggesting non-normal error behavior and the presence of large residuals. This behavior is also reflected in the residual histogram, which spans a broad range and indicates a high dispersion of errors.

For the BP-ANN model, the residual spread is somewhat reduced compared with MLR, although relatively large deviations are still observed, particularly at higher predicted UCS values. The central part of the Q–Q plot shows better alignment with the reference line, but deviations at the extremes remain evident, indicating that large prediction errors are still present for some samples. In contrast, ensemble and boosting models such as XGBoost and CatBoost exhibit more compact residual distributions. For example, the residuals of XGBoost are more concentrated around zero than those of MLR and RF, and its histogram shows a narrower error distribution. The Q–Q plots of the boosting-based models also show improved alignment with the reference line, particularly in the central region.

A further refinement is observed in the physics-guided CatBoost models. CatBoost-7PG shows a relatively compact residual distribution compared with the baseline CatBoost model, indicating that the initially introduced physics-guided features improve part of the error structure. However, the most balanced residual behavior is observed for CatBoost-6PG. Its residuals are more concentrated around zero, and its residual histogram is narrower and more centered than those of the other models. This behavior is consistent with its lower RMSE of 151.11 kPa and lower MAE of 110.67 kPa reported in Table 6. The Q–Q plot of CatBoost-6PG also shows closer alignment with the reference line in the central region, although deviations

at the tails remain, particularly for high-strength mixtures. Overall, the residual diagnostics indicate that the improved performance of CatBoost-6PG is mainly associated with a reduction in the spread of prediction errors and more stable predictions across the group-held-out folds.


MLR
Ordered Residuals
Theoretical Quantiles
MLR
Residual UCS (kPa)
Predicted UCS (kPa)
MLR
Frequency
Residual UCS (kPa)
SVR
Ordered Residuals
Theoretical Quantiles
SVR
Residual UCS (kPa)
Predicted UCS (kPa)
SVR
Frequency
Residual UCS (kPa)
RF
Ordered Residuals
Theoretical Quantiles
RF
Residual UCS (kPa)
Predicted UCS (kPa)
RF
Frequency
Residual UCS (kPa)
XGBoost
Ordered Residuals
Theoretical Quantiles
XGBoost
Residual UCS (kPa)
Predicted UCS (kPa)
XGBoost
Frequency
Residual UCS (kPa)
BP-ANN
Ordered Residuals
Theoretical Quantiles
BP-ANN
Residual UCS (kPa)
Predicted UCS (kPa)
BP-ANN
Frequency
Residual UCS (kPa)

669

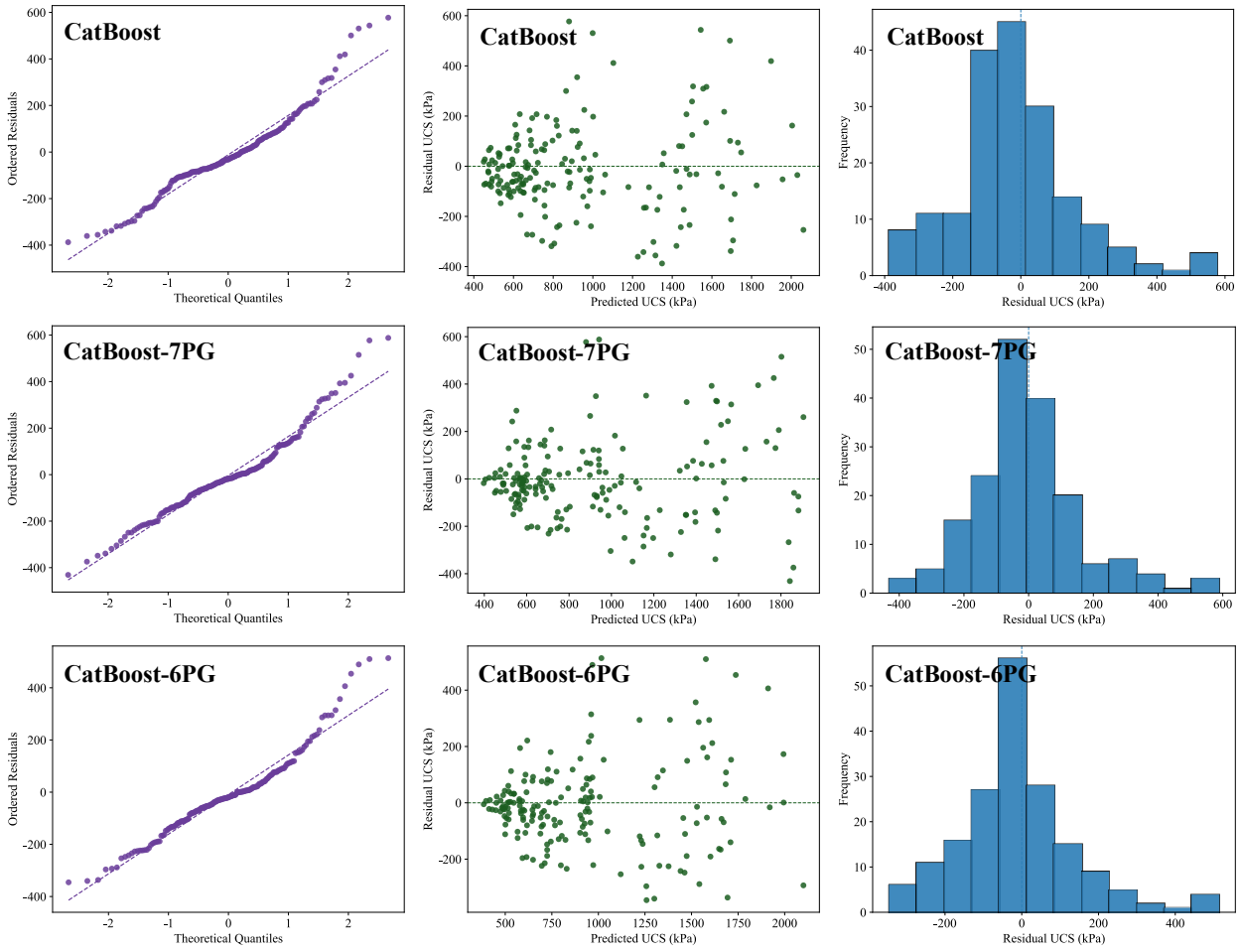


Figure 5. Residual diagnostic plots for the developed models based on outer-fold test predictions from nested GroupKFold cross-validation, including Q–Q plots, residuals versus predicted UCS, and residual histograms

3.2. Generalization Gap Analysis

The generalization behavior of the developed models was evaluated by comparing training RMSE and outer-test RMSE across the nested GroupKFold folds, as summarized in Table 7. The generalization gap was calculated as:

$$\Delta\text{RMSE} = \text{RMSE}_{\text{outer-test}} - \text{RMSE}_{\text{train}} \quad (22)$$

where a larger positive gap indicates a greater difference between fitting performance on the training folds and predictive performance on the group-held-out outer-test folds. The relative gap was also computed on a fold-wise basis to provide a normalized measure of the generalization difference.

A clear trend is observed across the models: high-capacity tree-based boosting models achieved very low training errors but showed larger generalization gaps when evaluated on group-held-out folds. For example, XGBoost produced the lowest training RMSE among the conventional models, with 11.23 ± 18.12 kPa, but its outer-test RMSE increased to 182.14 ± 18.40 kPa, resulting in the largest RMSE gap of 170.91 ± 11.65 kPa. A similar pattern is observed for the baseline CatBoost model, which showed a very low training RMSE of 13.58 ± 10.24 kPa but a substantially higher outer-test RMSE of 167.06 ± 41.84 kPa. These results indicate that

although boosting models have strong fitting capacity, they may also fit fold-specific patterns when evaluated under a stricter group-based validation scheme.

Table 7. Training-outer-test generalization gap analysis based on nested GroupKFold cross-validation

Model	Training RMSE (Mean ± SD)	Outer Test RMSE (Mean ± SD)	RMSE Gap (Mean ± SD)	Relative Gap (%) (Mean ± SD)
MLR	188.49 ± 16.80	235.30 ± 87.33	46.81 ± 104.01	7.82 ± 43.42
SVR	117.58 ± 14.14	164.87 ± 54.92	47.29 ± 46.31	24.16 ± 19.81
RF	59.51 ± 26.35	212.00 ± 66.68	152.49 ± 46.19	72.11 ± 5.64
XGBoost	11.23 ± 18.12	182.14 ± 18.40	170.91 ± 11.65	94.35 ± 8.50
CatBoost	13.58 ± 10.24	167.06 ± 41.84	153.48 ± 44.97	90.78 ± 9.00
BP-ANN	116.18 ± 10.89	172.61 ± 25.26	56.42 ± 18.12	32.08 ± 6.93
CatBoost-7PG	40.00 ± 15.74	169.03 ± 20.69	129.03 ± 21.71	76.32 ± 9.40
CatBoost-6PG	24.36 ± 18.18	151.11 ± 34.94	126.75 ± 18.28	82.60 ± 7.06

RF also shows a notable generalization gap, with training and outer-test RMSE values of 59.51 ± 26.35 kPa and 212.00 ± 66.68 kPa, respectively. This indicates that the ensemble was able to fit the training folds reasonably well, but its predictive performance decreased considerably for unseen mixture-curing groups. In contrast, SVR and BP-ANN showed smaller RMSE gaps of 47.29 ± 46.31 kPa and 56.42 ± 18.12 kPa, respectively. However, their lower generalization gaps did not necessarily correspond to the best predictive performance, as their outer-test RMSE values remained higher than that of CatBoost-6PG.

The MLR model exhibited the smallest average RMSE gap among the evaluated models. However, this should not be interpreted as superior generalization. Its training RMSE was already high (188.49 ± 16.80 kPa), and its outer-test RMSE was the highest among all models (235.30 ± 87.33 kPa). Therefore, the small gap mainly reflects the limited flexibility and underfitting tendency of the linear model rather than strong predictive capability.

The comparison between CatBoost-7PG and CatBoost-6PG provides insight into the effect of physics-guided feature refinement. CatBoost-7PG achieved a training RMSE of 40.00 ± 15.74 kPa and an outer-test RMSE of 169.03 ± 20.69 kPa. After removing the redundant QL_Curing feature, CatBoost-6PG reduced the outer-test RMSE to 151.11 ± 34.94 kPa and achieved the lowest prediction error among all models. Although CatBoost-6PG still shows a relatively large training-testing gap, this behavior reflects the high flexibility of the model under a conservative group-held-out validation framework. More importantly, its outer-test error is the lowest, indicating that the retained physics-guided features improved generalization to unseen groups.

Overall, the generalization gap analysis shows that a small training-testing gap alone is not sufficient to identify the best model. Models with low flexibility, such as MLR, may show smaller gaps but poor accuracy, while high-capacity models may show larger gaps but stronger predictive performance. Among the evaluated models, CatBoost-6PG provided the most favorable balance by achieving the lowest outer-test RMSE, MAE, and MAPE, while maintaining improved generalization compared with the initial CatBoost-7PG feature set. The remaining gap highlights the difficulty of predicting UCS for unseen mixture-curing groups using a relatively small experimental dataset and motivates the subsequent uncertainty analysis.

3.3. Ablation Study

To identify a compact and effective subset of physics-guided engineered features, a sequential backward ablation study was conducted using the CatBoost framework. The ablation analysis was applied only to the engineered features, while the original geotechnical input variables were retained in all scenarios. The initial engineered feature set included Binder_Total, NS_Curing, QL_Curing, BinderTotal_Curing, log_Curing, MDD_OMC_Ratio, and PI_LL_Ratio.

The ablation procedure followed the same nested GroupKFold framework used for model evaluation. In each outer fold, all replicate samples belonging to the same mix design-curing condition were kept together to reduce replicate-level information leakage. For each candidate feature subset, CatBoost hyperparameters were optimized using PSO within the inner GroupKFold loop. The model was trained on log-transformed UCS values, and predictions were back-transformed to the original UCS scale before calculating performance metrics. Feature elimination was guided primarily by the mean RMSE across the outer test folds.

The sequential backward procedure started from the complete seven-feature configuration, denoted as CatBoost-7PG. At each step, one remaining engineered feature was removed, and the resulting candidate models were independently re-optimized and evaluated. A feature removal was accepted only if it reduced the mean outer-test RMSE relative to the current feature set. The procedure was terminated when no additional one-feature removal improved the RMSE.

Figure 6a shows the accepted backward-elimination path. The initial CatBoost-7PG configuration achieved an outer-test RMSE of 164.82 ± 19.02 kPa. Removing QL_Curing reduced the RMSE to 145.51 ± 26.45 kPa, corresponding to an improvement of 19.31 kPa. This indicates that the direct QL-curing interaction introduced redundant information when the raw QL content, curing days, BinderTotal_Curing, and log_Curing were already included. This redundancy is also physically reasonable because QL was fixed at 3 wt.% in the combined NS-QL mixtures, while its independent variation occurred only in the QL-only mixtures.

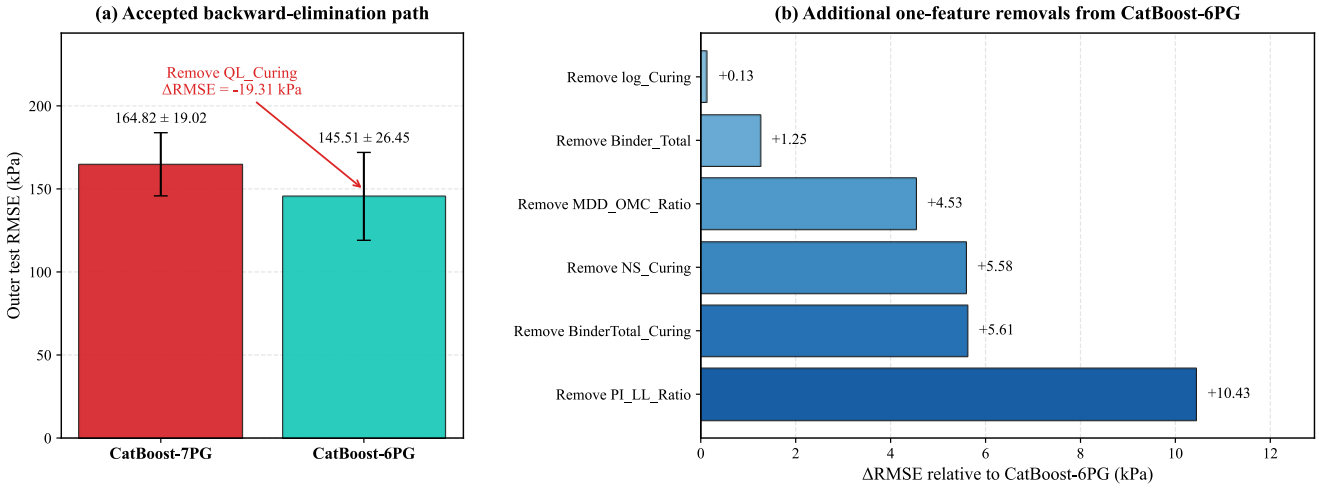


Figure 6. Sequential backward ablation results for the physics-guided CatBoost model: **(a)** Accepted backward-elimination path from CatBoost-7PG to CatBoost-6PG; and **(b)** RMSE changes for additional one-feature removals from CatBoost-6PG

After removing QL_Curing, an additional one-feature removal analysis was performed on the remaining six engineered features. As shown in Figure 6b, all further removals produced positive ΔRMSE values relative to CatBoost-6PG, indicating that no additional feature removal improved the model. The closest case was the removal of log_Curing, which increased RMSE only slightly by 0.13 kPa. In contrast, removing PI_LL_Ratio caused the largest deterioration, with ΔRMSE = +10.43 kPa, confirming the importance of the relative plasticity descriptor. Therefore, the backward elimination process stopped after removing QL_Curing.

The selected configuration, denoted as CatBoost-6PG, retained six engineered features: Binder_Total, NS_Curing, BinderTotal_Curing, log_Curing, MDD_OMC_Ratio, and PI_LL_Ratio. These features represent total binder content, NS-curing interaction, total binder-curing interaction, nonlinear curing-related strength development, compaction state, and relative plasticity, respectively. The retained feature set therefore provides a compact and physically meaningful representation of the main mechanisms controlling UCS development. A certain degree of correlation among the retained engineered features is expected because several of them describe related binder-curing mechanisms. For example, NS_Curing and BinderTotal_Curing both include curing days and therefore partly overlap in their information content. However, the additional one-feature removal analysis showed that removing any of the six retained engineered features increased the outer-test RMSE relative to CatBoost-6PG. Therefore, despite partial correlation, the retained features provided complementary predictive information under the nested GroupKFold evaluation.

It should be noted that the RMSE values reported in Figure 6 correspond to the ablation-selection procedure, whereas the final model-comparison results in Table 6 were obtained from the independently reported nested GroupKFold evaluation of all developed models. In the final comparison, CatBoost-6PG achieved the best overall performance, with R^2 = 0.841 ± 0.123, RMSE = 151.11 ± 34.94 kPa, MAE = 110.67 ± 26.04 kPa, MAPE = 11.97 ± 3.75%, and VAF = 87.38 ± 7.24%. Therefore, the ablation results support the removal of the redundant QL_Curing feature, while the final model-comparison results confirm CatBoost-6PG as the best-performing model.

The ablation study shows that the full seven-feature physics-guided configuration contained redundant information. The selected CatBoost-6PG model provided a better balance between predictive accuracy, reduced feature dimensionality, and geotechnical interpretability. Accordingly, CatBoost-6PG was adopted as the final physics-guided CatBoost model for the subsequent explainability, sensitivity, and uncertainty analyses.

3.4. Reduced-Input Assessment for Preliminary UCS Screening

To examine whether the modeling framework can provide useful information under an early-stage mix design condition, two reduced-input CatBoost scenarios were evaluated as supplementary analyses. In these scenarios, the available inputs were limited to the primary mix design variables, namely NS (wt.%), QL (wt.%), and curing days. The purpose of this comparison was not to replace the final CatBoost-6PG model, but to assess how much predictive capability remains when geotechnical index and compaction variables are not yet available.

CatBoost-3Input used only NS, QL, and curing days. CatBoost-3Input+PG used the same three variables together with binder- and curing-related engineered features, including Binder_Total, NS_Curing,

BinderTotal_Curing, and log_Curing. Both reduced-input scenarios were trained and evaluated using the same nested GroupKFold validation, PSO optimization, target transformation, and grouping strategy adopted for the final model. The results are summarized in Table 8.

Table 8. Reduced-input assessment for preliminary UCS screening and comparison with the CatBoost-6PG model

Model	Input scenario	R^2	RMSE (kPa)	MAE (kPa)	MAPE (%)	VAF (%)
CatBoost-3Input	NS, QL, curing days	0.667 ± 0.469	182.67 ± 53.17	136.87 ± 38.14	16.07 ± 7.75	74.76 ± 32.15
CatBoost-3Input+PG	NS, QL, curing days + binder-curing PG features	0.685 ± 0.393	194.20 ± 49.78	147.24 ± 39.06	17.32 ± 6.84	75.10 ± 27.03
CatBoost-6PG	Full input + 6 PG features	0.841 ± 0.123	151.11 ± 34.94	110.67 ± 26.04	11.97 ± 3.75	87.38 ± 7.24

Note: Values represent mean ± standard deviation across the five outer GroupKFold folds.

The results show that the primary mix design variables alone provide meaningful predictive information. CatBoost-3Input achieved R^2 = 0.667 ± 0.469 and RMSE = 182.67 ± 53.17 kPa. This performance indicates that a nonlinear model can provide a useful preliminary estimate of UCS when only binder dosage and curing days are available. For context, this reduced-input model also performed better than the full-input linear MLR benchmark reported in Table 6, showing that nonlinearity is important even when the input space is limited.

However, the final CatBoost-6PG model remained clearly more accurate and stable. Compared with CatBoost-3Input, CatBoost-6PG increased the mean R^2 from 0.667 to 0.841 and reduced RMSE from 182.67 to 151.11 kPa. The MAPE also decreased from 16.07% to 11.97%. More importantly, the standard deviation of R^2 decreased from 0.469 to 0.123, indicating more stable generalization across held-out mix-curing groups. Therefore, the reduced-input model may be useful for early-stage screening, whereas the full CatBoost-6PG model is more appropriate when reliable UCS prediction is required. This distinction is practically important because LL, PI, OMC, and MDD require additional laboratory characterization. Therefore, the final model does not eliminate laboratory work; rather, it can reduce the number of UCS tests once basic index and compaction properties are available.

CatBoost-3Input+PG produced a slightly higher mean R^2 than CatBoost-3Input, but its error-based metrics were not improved. This suggests that binder- and curing-related engineered features alone cannot fully compensate for the absence of LL, PI, OMC, and MDD. Accordingly, the improved performance of CatBoost-6PG is attributed to the combined contribution of nonlinear learning, geotechnical characterization variables, and selected mechanism-inspired features.

3.5. Comparison of Developed Models

The final comparison of the developed models was performed using the outer-fold predictions obtained from the nested GroupKFold cross-validation framework. Figure 7 presents the Taylor diagram, which compares the models based on their correlation with the measured UCS values, standard deviation, and centered root mean square difference [43]. Models located closer to the reference point provide better agreement with the experimental UCS values.

As shown in Figure 7, CatBoost-6PG is positioned closest to the reference point among the evaluated models, indicating the most favorable combination of correlation, centered error, and standard deviation. This confirms that the selected physics-guided feature configuration improved the agreement between predicted and measured UCS values. CatBoost, CatBoost-7PG, BP-ANN, and SVR are also located relatively close to the reference point, reflecting their ability to capture nonlinear UCS behavior. In contrast,

MLR and RF show larger deviations from the reference point, indicating weaker agreement with the measured UCS values and less stable predictive behavior.

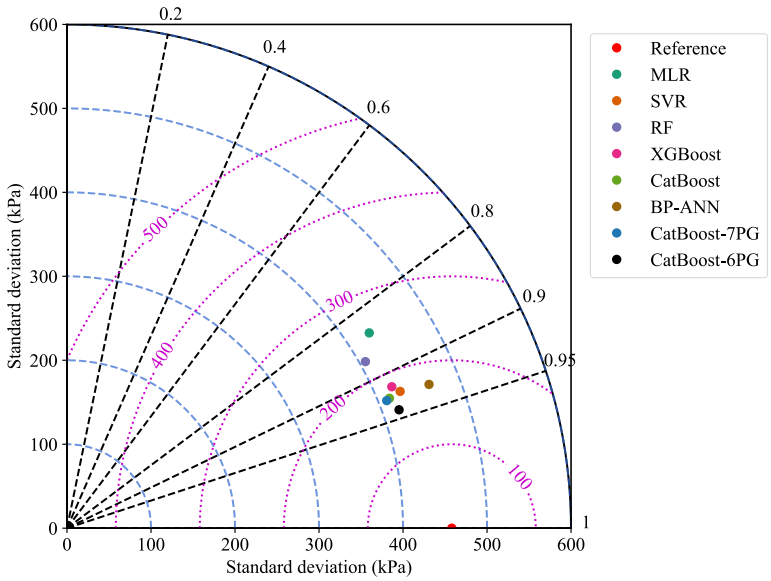


Figure 7. Taylor diagram comparing the developed models based on outer-fold predictions

The circular metric plots in Figure 8 further support this ranking. Based on the outer-fold performance metrics, CatBoost-6PG achieved the highest R^2 value of 0.841 and the lowest RMSE of 151.11 kPa among all developed models. BP-ANN showed the second-highest R^2 value of 0.808, followed by CatBoost-7PG and the baseline CatBoost models with R^2 values of 0.795 and 0.794, respectively. However, in terms of RMSE, CatBoost-6PG provided the lowest prediction error, followed by SVR, CatBoost, CatBoost-7PG, BP-ANN, XGBoost, RF, and MLR.

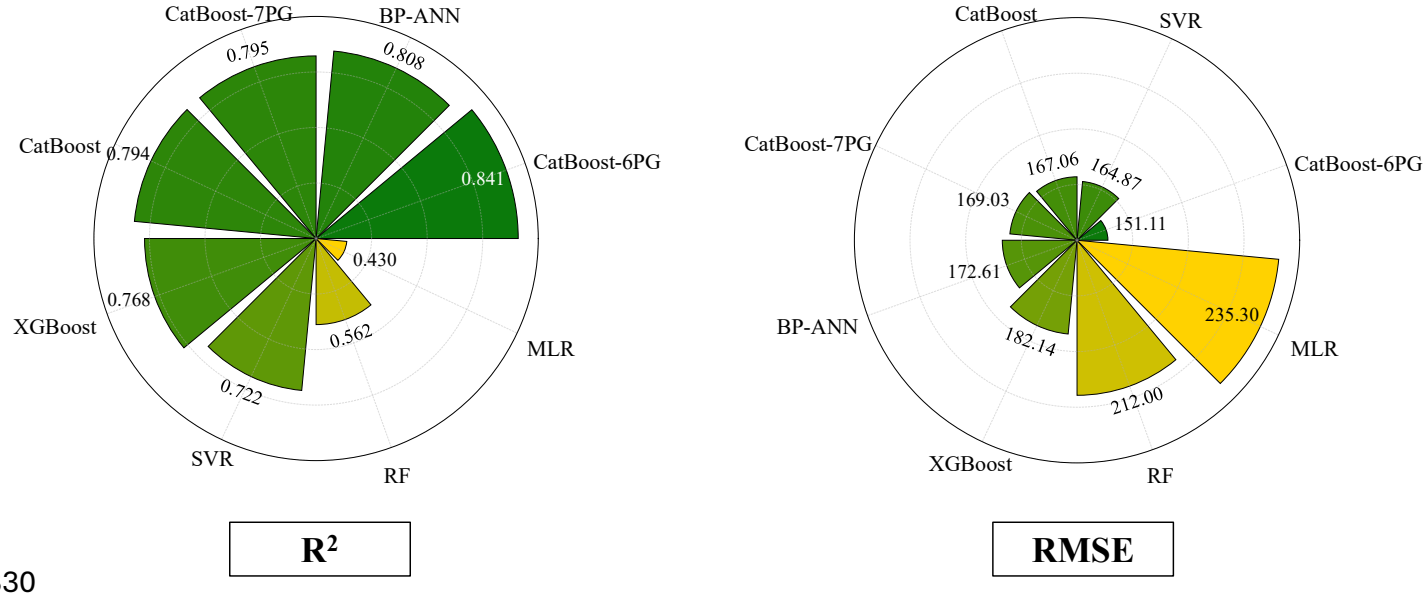


Figure 8. Circular metric plots comparing the developed models in terms of R^2 and RMSE

Compared with the baseline CatBoost model, CatBoost-6PG reduced the RMSE from 167.06 to 151.11 kPa, corresponding to an improvement of approximately 9.5%. Compared with CatBoost-7PG, the RMSE decreased from 169.03 to 151.11 kPa, corresponding to an improvement of approximately 10.6%. This confirms that removing the redundant QL_Curing feature improved generalization and reduced prediction error, while preserving the physically meaningful information contained in the remaining engineered features.

3.6. Comparison with Previous Studies

The performance of the proposed CatBoost-6PG model was compared with previously reported ML-based models for UCS prediction, as summarized in Table 9. Since the reviewed studies were developed using different dataset sizes, input variables, validation strategies, and UCS ranges, direct comparison based only on R^2 or RMSE may be misleading. Therefore, two additional indicators were used: the uncertainty index U and the normalized RMSE (NRMSE).

$$U = \sqrt{1-R^2} \tag{23}$$

$$NRMSE = \frac{RMSE}{(UCS_{max}-UCS_{min})} \times 100 \tag{24}$$

The uncertainty index U represents the unexplained portion of prediction variance; therefore, lower values indicate lower unexplained variance. NRMSE normalizes RMSE by the UCS range, allowing models developed over different strength scales to be compared more fairly.

Table 9. Comparison of the proposed model with previously reported ML-based UCS prediction studies

ML Model	Dataset Size	Num. of Inputs	Validation strategy	R^2	RMSE (kPa)	UCS Range (kPa)	U	NRMSE (%)	References
XGBoost	200	7	Hold-out validation	0.950	77.0	50-3400	0.223	2.30	[14]
GB-PSO	390	5	Hold-out validation	0.966	95.1	100-3282	0.183	2.99	[5]
SVR	45	8	Hold-out validation + 5-fold CV	0.820	437.9	0.5-1000	0.424	43.81	[33]
PI-FNN	109	3	Leave-one-region-out CV	0.902	537.0	23-5800	0.313	9.29	[19]
HC-based RF	109	3	Leave-one-region-out CV	0.930	452.0	23-5800	0.264	8.30	[19]
RF	171	7	Nested 5×5 CV	0.947	291.2	157-3138	0.230	9.77	[20]
XGBoost	341	5	Hold-out validation	0.910	155.1	74-3171	0.300	5.01	[13]
RF	190	6	Hold-out validation	0.930	300.0	100-5400	0.264	5.66	[12]
CatBoost-6PG	180	7 + 6 eng.	Nested GroupKFold CV	0.841	151.1	379-2317	0.399	7.80	This study

As shown in Table 9, previously reported ML-based UCS prediction studies have achieved a wide range of predictive performance depending on the dataset size, input variables, material system, and validation strategy. Several studies reported high R^2 values greater than 0.90, particularly when using ensemble or boosting-based models such as XGBoost, GB-PSO, RF, and CatBoost. For example, the XGBoost and GB-PSO models achieved R^2 values of 0.950 and 0.966, respectively, with low NRMSE values of 2.30% and 2.99%. These results confirm the strong capability of nonlinear ML models to capture complex UCS behavior in stabilized soils.

However, the validation strategies used in previous studies differ considerably. Some studies relied mainly on hold-out validation, while others used additional cross-validation or regional validation

procedures. Among the reviewed studies, the RF model evaluated using nested 5×5 cross-validation provided one of the most rigorous conventional validation frameworks, with R^2 = 0.947, RMSE = 291.2 kPa, and NRMSE = 9.77%. Similarly, the PI-FNN and HC-based RF models were evaluated using leave-one-region-out cross-validation, which provides a more challenging assessment of regional generalization compared with a simple hold-out split.

In the present study, CatBoost-6PG achieved R^2 = 0.841, RMSE = 151.1 kPa, U = 0.399, and NRMSE = 7.80% under nested GroupKFold cross-validation. Although the nominal R^2 value is lower than some previously reported models, this result should be interpreted in the context of the stricter group-aware validation strategy. In the nested GroupKFold framework, samples belonging to the same mixture–curing group were kept within the same fold, reducing replicate-level information leakage and providing a more conservative estimate of generalization to unseen mixture-curing conditions.

The comparison also shows that RMSE alone is not sufficient for evaluating models developed over different UCS ranges. For instance, some models reported lower RMSE values but were trained and tested over wider UCS domains, while others showed higher RMSE values but maintained competitive NRMSE values after normalization. The proposed CatBoost-6PG model achieved an NRMSE of 7.80%, which is comparable to or lower than several previously reported models, despite being evaluated under a more conservative group-held-out validation framework.

Overall, the comparison indicates that the proposed CatBoost-6PG model provides competitive predictive performance while adopting a group-aware validation strategy that is more conservative than conventional hold-out validation and standard cross-validation. More importantly, the present framework goes beyond point prediction by integrating physics-guided feature engineering, ablation-based feature refinement, SHAP explainability, sensitivity analysis, 2D partial dependence analysis, and uncertainty quantification. Therefore, the main contribution of this study is not only improved UCS prediction, but also the development of a physically interpretable and reliability-aware modeling framework for NS-QL stabilized kaolin clay.

4. Discussion

4.1. Interpretation of Model Behavior (SHAP)

4.1.1. Global Feature Importance

Figure 9a presents the global feature importance of the selected CatBoost-6PG model based on mean absolute SHAP values. Since the model was trained on the logarithmic scale of UCS, the SHAP values represent feature contributions to the model output in the log-transformed UCS space. The ranking shows that PI (%) is the most influential variable, followed by log_Curing, PI_LL_Ratio, and curing days. This indicates that plasticity-related characteristics and curing-dependent strength development are the dominant factors controlling UCS prediction in the stabilized kaolin clay system.

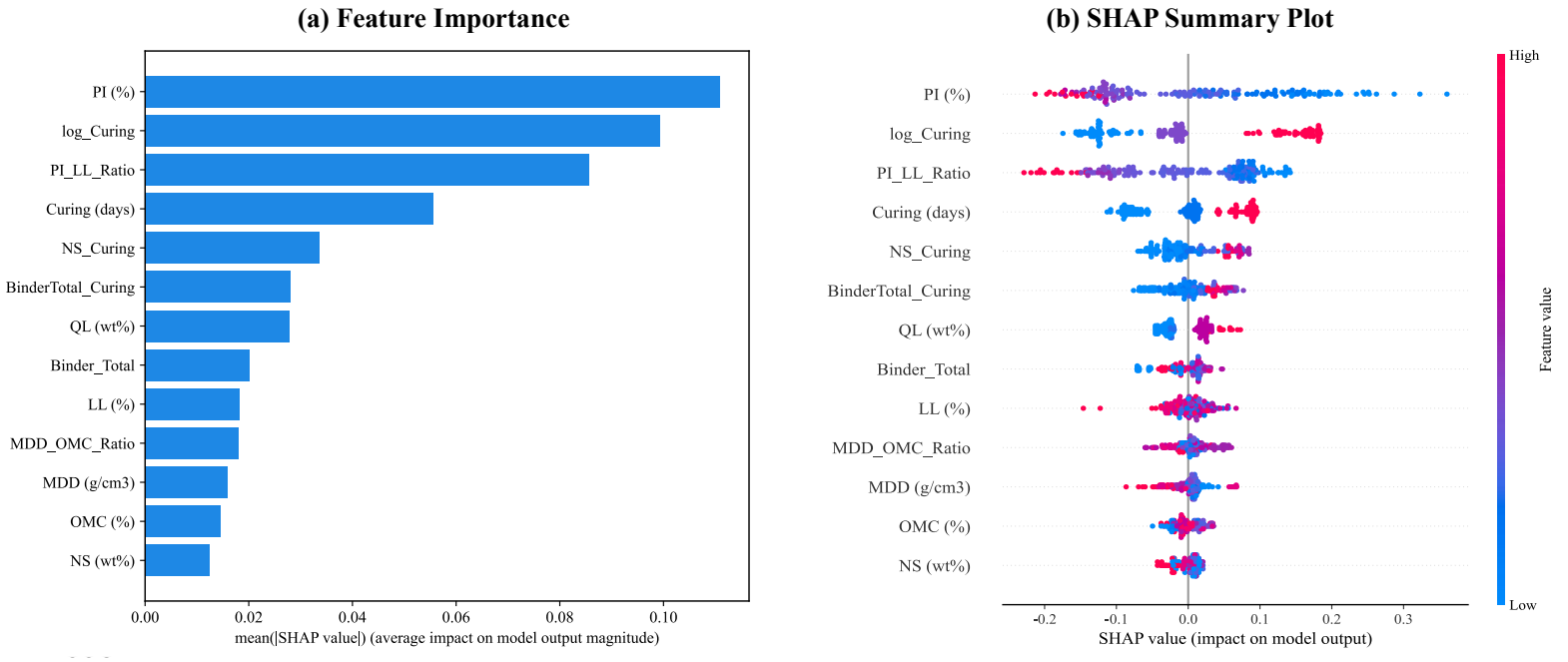


Figure 9. SHAP analysis results for CatBoost-6PG model on full dataset: **(a)** Feature importance, and **(b)** Beeswarm summary plot

The high importance of PI is physically meaningful because plasticity reflects the consistency, water affinity, and particle interaction characteristics of the treated soil. A higher PI generally indicates a more plastic soil structure, which can reduce the efficiency of bonding and hinder strength development [9]. The strong contribution of curing days and log_Curing confirms the importance of time-dependent reactions, including lime activation, pozzolanic reactions, and progressive formation of cementitious bonds. The simultaneous importance of curing days and log_Curing suggests that the model captured both the direct effect of curing duration and the nonlinear nature of strength gain with curing.

The PI_LL_Ratio also shows a high contribution, indicating that relative plasticity provides additional information beyond the absolute PI and LL values. This result supports the use of ratio-based physics-guided variables in the feature set. Among the binder-related variables, NS_Curing and BinderTotal_Curing show moderate importance, suggesting that the strength contribution of NS and total binder content depends strongly on curing days. This behavior is consistent with the expected mechanism of stabilized soils, where binder dosage alone is not sufficient to describe strength gain unless its interaction with curing is also considered [26].

QL content and Binder_Total exhibit intermediate contributions. This indicates that the amount of alkaline activator and total stabilizer dosage influences UCS prediction, but their effects are less dominant than plasticity and curing-related variables. In contrast, LL, MDD_OMC_Ratio, MDD, OMC, and NS content alone show lower global importance. The relatively low importance of NS alone suggests that NS does not act independently; instead, its contribution becomes more meaningful when combined with curing days, as reflected by the higher importance of NS_Curing. Similarly, the lower importance of MDD alone indicates that density by itself is not the primary driver of UCS variation in this stabilized system.

Figure 9b provides additional insight into the direction and distribution of feature effects. Higher curing days and higher log_Curing values are generally associated with positive SHAP values, indicating that

longer curing periods increase the predicted UCS. This trend is consistent with the progressive development of pozzolanic and cementitious reactions over time [8,44]. The effect is not purely linear, which explains the importance of including log_Curing as a nonlinear curing variable.

In contrast, PI shows a strong inverse relationship with predicted UCS. Higher PI values are mostly located on the negative side of the SHAP axis, whereas lower PI values tend to contribute positively. This indicates that increased plasticity reduces the predicted UCS, which agrees with the experimental behavior of stabilized fine-grained soils [26]. A similar pattern is observed for PI_LL_Ratio, where higher relative plasticity tends to shift SHAP values toward the negative side, confirming the detrimental effect of plasticity on strength development.

For binder-related interaction features, higher NS_Curing and BinderTotal_Curing values generally shift SHAP values toward the positive side. This suggests that the contribution of NS and total binder dosage becomes more effective when sufficient curing duration is provided. QL content also tends to show positive SHAP contributions at higher values, although its effect is narrower than that of curing and plasticity-related features. This supports the role of QL as an alkaline activator while indicating that its effect is controlled by interactions with curing and the overall mixture system.

Features such as OMC, MDD, LL, and MDD_OMC_Ratio exhibit relatively narrow SHAP ranges, indicating weaker individual effects on the model output. The slight negative tendency observed for higher MDD values suggests that increased dry density alone does not necessarily correspond to higher UCS in this dataset. This behavior should be interpreted as an indirect effect of stabilization, where binder addition may increase water demand and modify the compacted structure while simultaneously improving strength through chemical bonding.

Overall, the SHAP results show that the CatBoost-6PG model learned physically meaningful relationships. The model assigns dominant importance to plasticity and curing-related variables, captures interaction-driven effects through NS_Curing and BinderTotal_Curing, and gives lower importance to variables whose individual effects are secondary. These findings support the interpretability of the selected physics-guided model and confirm that its predictions are consistent with the expected mechanisms of NS-QL stabilization.

4.1.2. Feature Effects Based on SHAP Dependence Analysis

The SHAP dependence plots in Figure 10 provide a more detailed view of how key variables influence the predictions of the selected CatBoost-6PG model. Since the model was trained on the logarithmic scale of UCS, the SHAP values represent the contribution of each feature to the model output in the log-transformed UCS space. The plots reveal several nonlinear and interaction-driven behaviors that are consistent with the expected mechanisms of stabilized clay strength development.

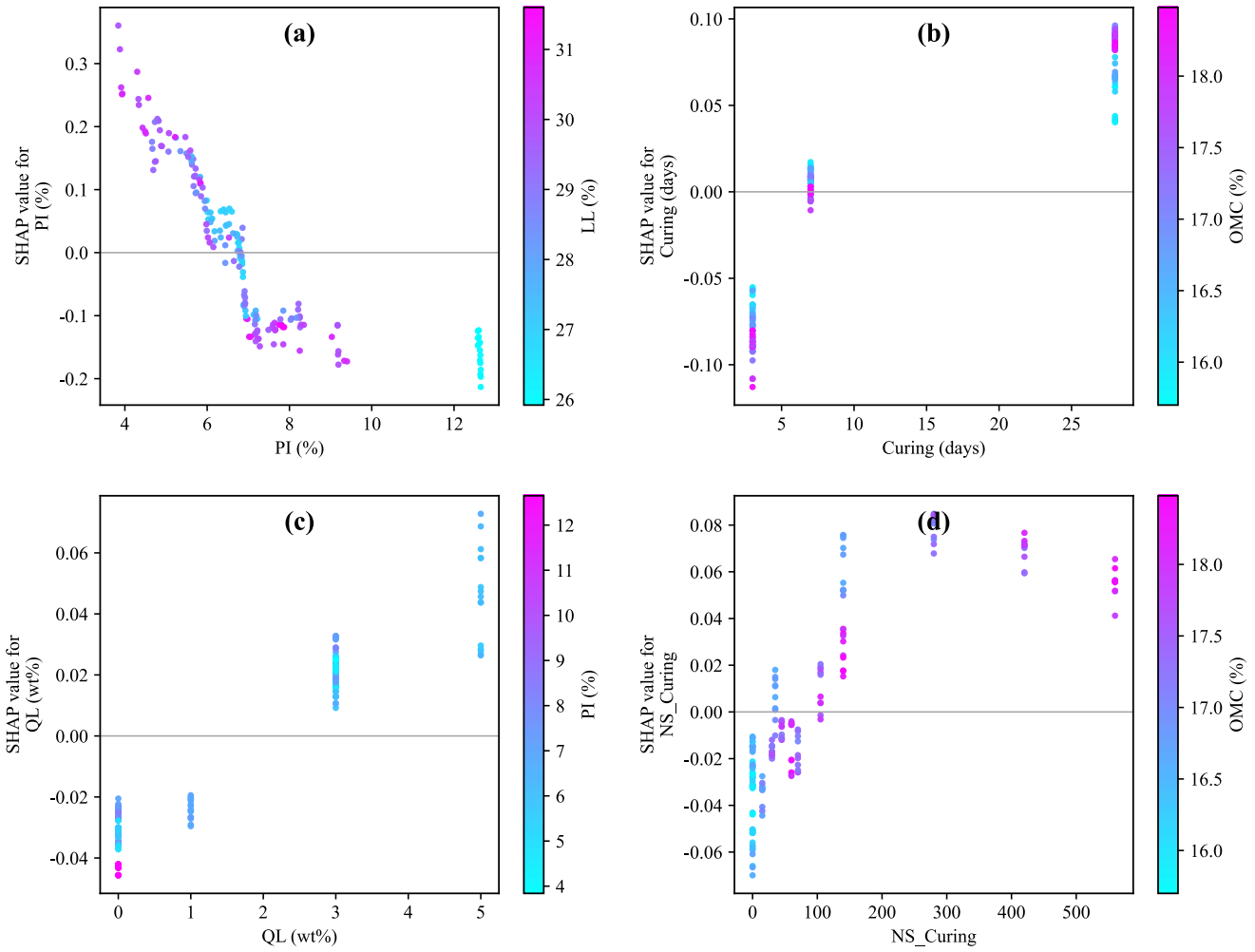


Figure 10. SHAP dependence plots for key variables in the selected CatBoost-6PG model: **(a)** PI, colored by LL; **(b)** curing days, colored by OMC; **(c)** QL content, colored by PI; and **(d)** NS_Curing, colored by OMC

For PI, a strong inverse relationship is observed (Figure 10a). As PI increases from approximately 4-6% to values greater than about 7%, the SHAP values decrease markedly from positive values to negative contributions. The transition zone occurs around PI ≈ 6.5–7%, where the SHAP values cross the zero line. This indicates that lower plasticity contributes positively to UCS prediction, whereas higher plasticity reduces the predicted strength. The color gradient associated with LL further suggests that the influence of PI is coupled with the consistency characteristics of the soil. This behavior is physically meaningful because increased plasticity generally reflects higher water affinity and weaker particle bonding, which can reduce the efficiency of stabilization [45,46].

Curing days show a distinct stepwise positive effect, as presented in Figure 10b. At 3 days, SHAP values are negative, indicating that early curing contributes to lower predicted UCS. At 7 days, the SHAP values are close to zero, while at 28 days they shift clearly toward the positive side. This separation confirms the important role of curing duration in strength development. The trend reflects the progressive formation of cementitious products and the continuation of pozzolanic reactions over time. The color variation with

OMC also indicates that curing-related effects may be influenced by the moisture condition of the compacted mixtures.

For QL content (Figure 10c), a nonlinear increasing trend is observed. At 0 wt.% QL, SHAP values are mostly negative, indicating the absence of alkaline activation. Increasing QL content to 3 wt.% shifts the SHAP values toward positive contributions, while 5 wt.% generally provides the highest positive SHAP values. This confirms the beneficial role of QL in improving UCS through alkaline activation, flocculation, and pozzolanic reaction processes [44]. However, the magnitude of the QL effect is smaller than that of PI and curing-related variables, suggesting that lime content alone does not fully control UCS development and must be interpreted together with curing days and other mixture properties.

The NS_Curing feature reveals a clear interaction-driven effect (Figure 10d). At low NS_Curing values, SHAP values are mostly negative, while increasing NS_Curing generally shifts the contribution toward the positive side. This indicates that the effect of NS becomes more beneficial when combined with sufficient curing duration. In other words, NS content alone has limited explanatory power, but its interaction with curing days captures the time-dependent contribution of NS to strength development. The color gradient associated with OMC suggests that this interaction may also be affected by water demand and compaction-related moisture conditions.

As a result, the dependence analysis confirms that UCS prediction is governed by a combination of strong individual effects and interaction-based mechanisms. Plasticity has a dominant negative influence, curing days provide a positive time-dependent contribution, QL improves the chemical stabilization potential, and NS_Curing captures the delayed contribution of NS under curing. These patterns support the physical interpretability of the CatBoost-6PG model and show that the retained physics-guided features provide meaningful information for predicting UCS of NS-QL stabilized kaolin clay.

4.1.3. Interaction Effects Based on SHAP Dependence Plots

The SHAP dependence plots in Figure 11 provide further insight into the coupled effects among stabilizer content, curing days, and soil plasticity. These plots show how the SHAP value of a target feature changes with its magnitude, while the color scale represents a second interacting variable. Therefore, they help identify whether the contribution of a feature is controlled by another physical or mixture-related parameter.

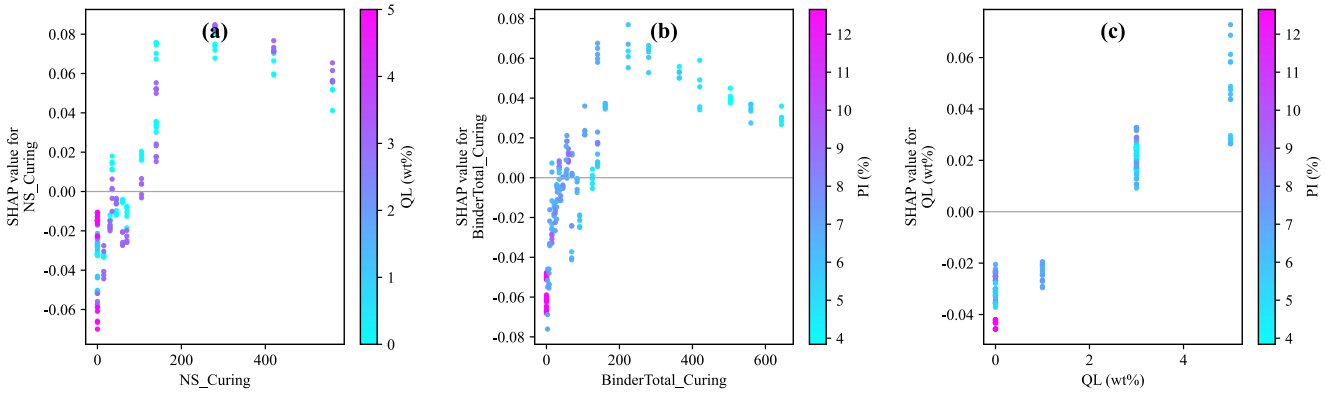


Figure 11. SHAP dependence plots illustrating interaction effects in the selected CatBoost-6PG model: **(a)** NS_Curing colored by QL content; **(b)** BinderTotal_Curing colored by PI; and **(c)** QL content colored by PI

The interaction between NS_Curing and QL content is shown in Figure 11a. At low NS_Curing values, the SHAP values are mostly negative, indicating that NS has a limited or unfavorable contribution when either NS content or curing duration is insufficient. As NS_Curing increases, the SHAP values gradually shift toward the positive side, particularly for mixtures with nonzero QL content. This trend suggests that the effectiveness of NS is enhanced when sufficient curing days are combined with alkaline activation provided by QL. In other words, the contribution of NS is not purely dosage-dependent; it is strongly controlled by curing duration and the presence of QL.

A similar interaction-driven behavior is observed for BinderTotal_Curing and PI (Figure 11b). At low BinderTotal_Curing values, SHAP values are mainly negative, reflecting insufficient combined binder-curing contribution. As BinderTotal_Curing increases, SHAP values become positive, indicating that higher total binder content combined with longer curing days improves UCS prediction. However, the color distribution shows that this positive contribution is moderated by PI. Mixtures with higher PI tend to show lower or less favorable SHAP contributions compared with lower-plasticity mixtures at comparable BinderTotal_Curing levels. This indicates that soil plasticity can limit the efficiency of binder-curing mechanisms, even when the total binder-curing index is relatively high.

The QL-PI dependence plot in Figure 11c further highlights the role of QL in strength development. At QL = 0 wt.%, SHAP values are predominantly negative, indicating limited strength contribution in the absence of alkaline activation. Increasing QL content to 3 wt.% shifts the SHAP values toward the positive range, while 5 wt.% generally produces the highest positive contributions. The color scale indicates that the influence of QL is also affected by PI, with lower-plasticity mixtures tending to show more favorable SHAP contributions. This suggests that QL improves UCS through alkaline activation and pozzolanic reaction processes, but its effectiveness is influenced by the plasticity state of the soil.

In summary, the interaction patterns in Figure 11 demonstrate that UCS prediction is controlled by coupled effects rather than by independent variables alone. The positive influence of NS becomes more evident when combined with curing days and QL activation, while the effectiveness of binder-curing mechanisms is moderated by plasticity. These findings support the inclusion of physics-guided interaction features such as NS_Curing and BinderTotal_Curing in the CatBoost-6PG model and confirm that the model captures physically meaningful mechanisms of NS-QL stabilization.

4.1.4. Local Interpretability

To provide a detailed understanding of how individual features contribute to a single model prediction, SHAP local analysis was performed for a representative moderate-to-high strength sample, as shown in Figure 12. For this sample, the measured UCS was 1164.9 kPa, while the CatBoost-6PG model predicted 1152.0 kPa, corresponding to an absolute error of only 12.93 kPa. This close agreement indicates that the model prediction for this sample is locally reliable and provides a useful case for interpreting the feature-level contributions.

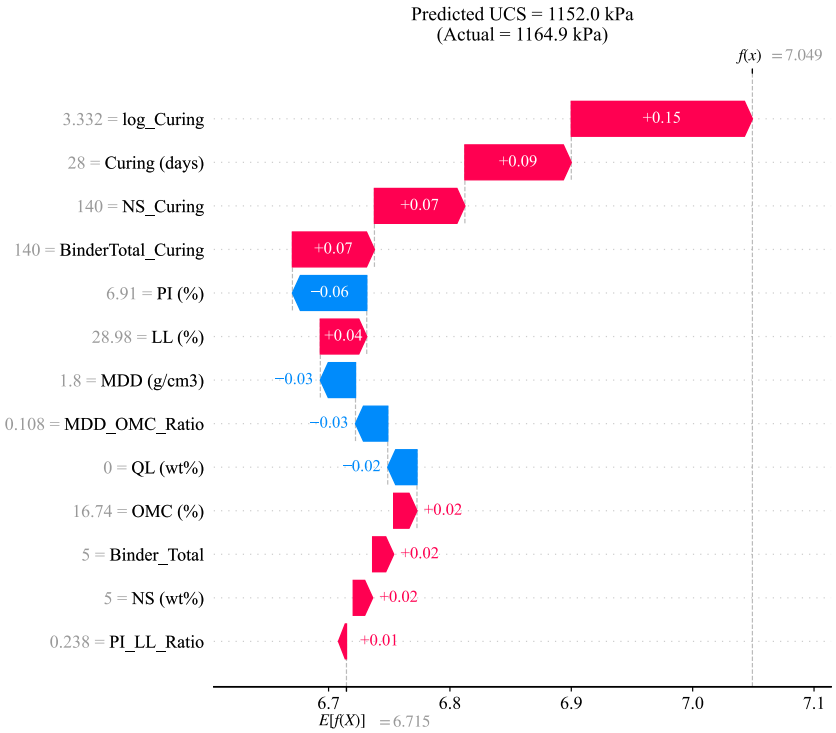


Figure 12. Local SHAP waterfall explanation for a representative moderate-to-high strength sample predicted by the selected CatBoost-6PG model

It should be noted that the SHAP values in Figure 12 are computed on the logarithmic scale of UCS because the model was trained using log-transformed UCS values. Therefore, the base value E[f(x)] = 6.715 and the final model output f(x) = 7.049 correspond to the log-transformed model output. The difference between these two values represents the cumulative contribution of all input and engineered features to the prediction.

The largest positive contribution is provided by log_Curing, with a SHAP contribution of approximately +0.15, followed by curing days and NS_Curing, each contributing positively to the prediction. This confirms that the model attributes the high predicted UCS mainly to curing-related effects and the interaction between NS content and curing days. For this sample, the curing period is 28 days and NS_Curing equals 140, indicating that the model recognizes the contribution of time-dependent stabilization mechanisms.

BinderTotal_Curing also contributes positively, indicating that the combined binder-curing effect supports the predicted strength. Other variables, including LL, OMC, Binder_Total, NS content, and PI_LL_Ratio, show smaller positive contributions. These effects suggest that the local prediction is not controlled by a single variable alone, but by the combined influence of curing duration, binder-curing interactions, and soil index properties.

Several features reduce the prediction relative to the base value. PI has a negative contribution of approximately −0.06, indicating that the plasticity level of this sample reduces the predicted UCS. MDD, MDD_OMC_Ratio, and QL also show negative local contributions. In particular, the negative contribution of QL is physically reasonable because this sample has QL=0 wt.%, meaning that no QL activation is

present. Therefore, although the sample benefits from curing days and NS-curing interaction, the absence of QL limits the predicted strength.

This local SHAP explanation shows that the predicted UCS is mainly increased by curing-related and interaction-based features, while plasticity and the absence of QL act as limiting factors. The balance between these positive and negative contributions leads to a predicted UCS very close to the measured value. This confirms that the CatBoost-6PG model provides not only accurate prediction for this representative sample, but also a physically interpretable explanation of the underlying strength-development mechanisms.

4.2. Sensitivity Analysis and Physical Consistency

To further examine the physical consistency of the selected CatBoost-6PG model, a one-at-a-time sensitivity analysis was performed on the raw input variables. The analysis was conducted around a baseline input defined by the median values of the full dataset: NS = 7.5 wt.%, QL = 2 wt.%, curing days = 7 days, LL = 28.99%, PI = 6.79%, OMC = 17.09%, and MDD = 1.685 g/cm^3. The corresponding baseline predicted UCS was 925.14 kPa. During the sensitivity analysis, each raw input variable was varied over its observed full-dataset range, while the remaining raw variables were fixed at their baseline values. The physics-guided engineered features were recalculated after each raw-feature perturbation to preserve consistency among the dependent variables.

Figure 13a presents the sensitivity range of each raw input variable. PI (%) showed the largest influence on the predicted UCS, with a sensitivity range of 691.72 kPa and a normalized sensitivity index of 1.00. Varying PI changed the predicted UCS from 617.39 to 1309.12 kPa. This strong response confirms the dominant role of plasticity in controlling the strength behavior of the stabilized kaolin clay system. The high sensitivity to PI is also consistent with the SHAP analysis, where plasticity-related variables exhibited substantial contributions to the model output.

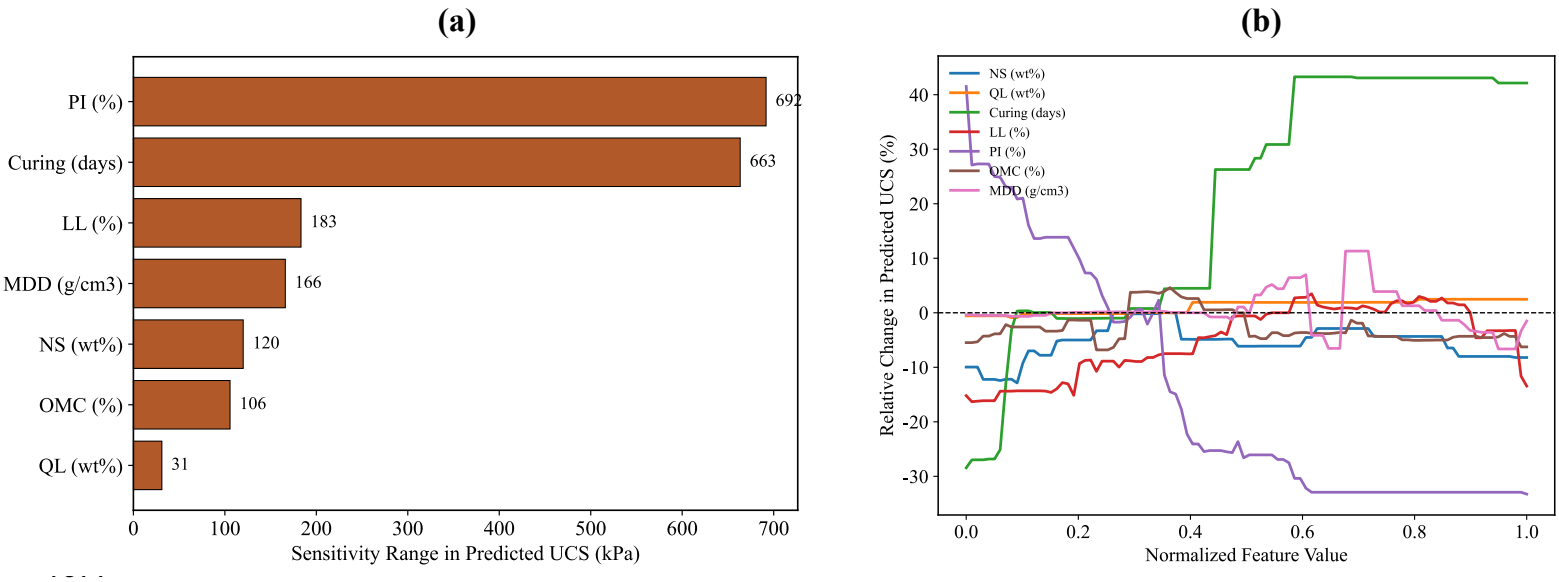


Figure 13. One-at-a-time sensitivity analysis of raw input variables for the selected CatBoost-6PG model: **(a)** sensitivity range in predicted UCS; **(b)** normalized sensitivity curves expressed as relative changes in predicted UCS

Curing days were the second most influential variable, with a sensitivity range of 663.46 kPa and a normalized sensitivity index of 0.959. The predicted UCS varied from 662.05 to 1325.51 kPa across the observed curing range. As shown in Figure 13b, increasing curing days produced a strong positive response in predicted UCS, reflecting time-dependent strength gain associated with stabilization reactions. This behavior is physically consistent with the progressive development of cementitious and pozzolanic bonding over time.

The remaining variables showed lower sensitivity compared with PI and curing days. LL (%) ranked third, with a sensitivity range of 183.25 kPa, followed by MDD with 166.11 kPa, NS content with 120.07 kPa, OMC with 105.67 kPa, and QL content with 31.12 kPa. The moderate influence of LL, MDD, and OMC suggests that soil consistency and compaction-moisture conditions still contribute to UCS variability, although their effects are secondary compared with plasticity and curing-related mechanisms. The relatively low direct sensitivity of QL content around the median baseline indicates that QL dosage alone did not dominate the local model response; rather, its influence is mainly expressed through interaction with curing days, binder-related variables, and the overall mixture condition.

Figure 13b shows the normalized sensitivity curves as relative changes in predicted UCS. The response curves further confirm the physically meaningful behavior of the CatBoost-6PG model. Higher PI values led to a marked reduction in predicted UCS, while curing days produced a strong positive response. NS content showed a weaker and nonlinear effect, suggesting that the contribution of NS is interaction-dependent rather than uniformly monotonic across the FullInput range. The relatively flat response of QL content further supports the ablation result, where the direct QL_Curing feature was identified as redundant and removed from the final model.

Overall, the sensitivity analysis supports the physical consistency of the CatBoost-6PG model. Although the sensitivity ranking is not identical to the SHAP ranking because the two methods evaluate model behavior differently, both analyses identify plasticity and curing-related effects as dominant controls on UCS prediction. Therefore, the final model not only provides the best predictive performance among the evaluated models, but also produces responses that are consistent with the underlying geotechnical stabilization mechanisms.

4.3. Two-Dimensional Partial Dependence Analysis

Two-dimensional partial dependence analysis was performed to further examine the interaction effects between the most influential variables identified by SHAP and sensitivity analyses. The resulting contour plots are presented in Figure 14. In these plots, the color gradient represents the mean predicted UCS obtained from the selected CatBoost-6PG model after back-transformation to the original UCS scale. Therefore, the contours provide an averaged view of how pairs of variables jointly influence the predicted UCS while accounting for the distribution of the remaining input features.

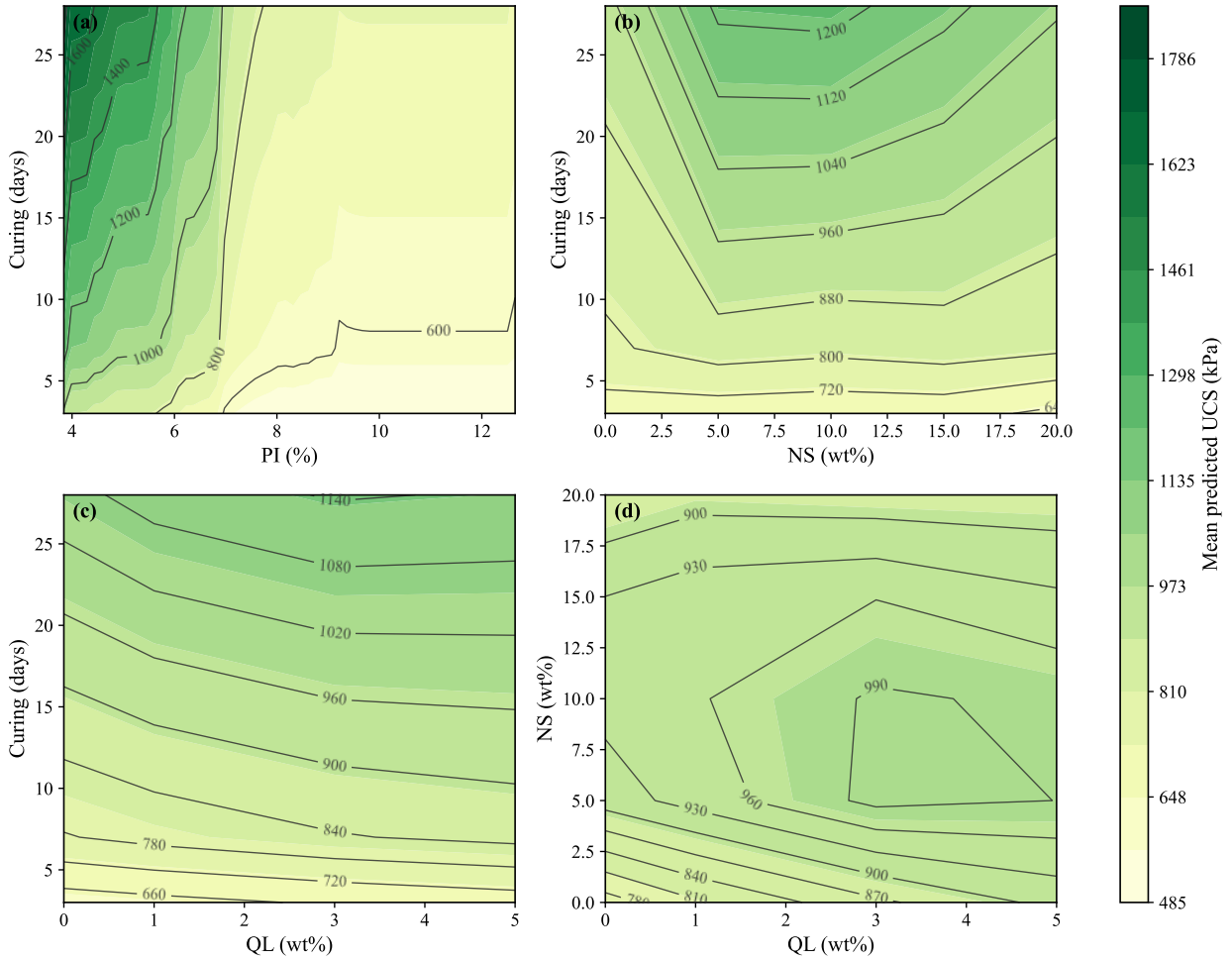


Figure 14. 2D-Partial dependence plot of the CatBoost-6PG model: **(a)** Curing days vs PI (%); **(b)** Curing days vs NS (wt.%); **(c)** Curing days vs QL (wt.%); **(d)** NS (wt.%) vs QL (wt.%)

Figure 14a shows the combined effect of PI and curing days. A clear inverse relationship is observed between PI and predicted UCS. The highest predicted UCS values are concentrated at low PI values and longer curing days, whereas increasing PI substantially reduces the predicted strength even when curing days increase. This pattern confirms the dominant negative role of plasticity observed in both SHAP and sensitivity analyses. The contour transition is particularly steep between PI values of approximately 5-7%, indicating that small changes in plasticity within this range can produce considerable changes in predicted UCS. This behavior is physically meaningful because lower plasticity is associated with reduced water affinity and more favorable particle interaction, allowing stabilization reactions to contribute more effectively to strength development [44,45].

Figure 14b illustrates the interaction between NS content and curing days. Predicted UCS generally increases with curing days, confirming the strong time-dependent nature of strength development. However, the effect of NS content is nonlinear. At longer curing periods, intermediate NS contents produce relatively high predicted UCS values, while the response becomes less favorable at very high NS contents. This indicates that NS contributes to strength improvement when sufficient curing is available, but its effect is not simply monotonic with dosage. The result supports the interpretation that NS acts through time-

dependent pozzolanic contribution and interaction with the alkaline environment rather than as an independently dominant variable.

Figure 14c presents the combined influence of QL content and curing days. The predicted UCS increases markedly with curing days at all QL levels, again confirming the importance of curing-related mechanisms. Increasing QL from 0 to approximately 3 wt.% improves the predicted UCS, especially at longer curing periods. However, further increase to 5 wt.% does not produce a proportional additional improvement across the entire domain. This suggests that the beneficial effect of QL is most effective up to an intermediate dosage, beyond which the response tends to plateau. Such behavior is consistent with the role of QL as an alkaline activator: sufficient lime promotes flocculation and pozzolanic reactions, but excessive lime content may not necessarily lead to further strength gain within the investigated range [26].

Figure 14d shows the joint effect of QL and NS contents. The predicted UCS is relatively low when both stabilizers are absent or present at low levels. The response improves at intermediate QL and NS combinations, particularly around QL values near 3 wt.% and moderate NS contents. This confirms that the strength development of the stabilized kaolin clay is governed by coupled binder effects rather than by either stabilizer alone. The contour pattern also suggests that very high stabilizer contents do not automatically maximize UCS, highlighting the importance of balanced NS-QL proportions for effective stabilization.

Overall, the 2D partial dependence results confirm the main findings obtained from SHAP and sensitivity analyses. Plasticity has a strong limiting effect on UCS, curing days provide the dominant positive contribution, and the effects of NS and QL are nonlinear and interaction-dependent. The highest predicted strength regions are associated with longer curing days, lower PI, and balanced stabilizer combinations. These results support the physical consistency of the CatBoost-6PG model and demonstrate its ability to capture nonlinear interaction patterns relevant to NS-QL stabilization.

4.4. Reliability and Uncertainty Analysis

Reliability and uncertainty analyses were performed for the selected CatBoost-6PG model using the Monte Carlo input-perturbation and group-bootstrap procedures described in Section 2.8. The results are presented in Tables 10-12 and Figures 15-17, including deterministic baseline performance, Monte Carlo prediction intervals under different perturbation scenarios, and group-bootstrap model uncertainty. The deterministic outer-fold predictions used as the baseline for uncertainty analysis achieved a pooled R^2 of 0.889, RMSE of 152.10 kPa, and MAE of 112.79 kPa, as summarized in Table 10. These values provide the reference predictions around which uncertainty intervals were constructed.

Table 10. Deterministic outer-fold performance of the CatBoost-6PG model used as the baseline for uncertainty analysis

Evaluation	R^2	RMSE (kPa)	MAE (kPa)
Outer-test mean	0.839	150.38	112.45
Outer-test SD	0.136	24.23	21.62
Pooled outer-fold predictions	0.889	152.1	112.79

The Monte Carlo uncertainty results are summarized in Table 11 and Figure 15. As shown in Figure 15a, the mean prediction uncertainty increased systematically with perturbation intensity. Under the FullInput setting, the mean Monte Carlo standard deviation increased from 50.18 kPa at k = 0.10 to 108.65 kPa at k = 0.25, and further to 190.14 kPa at k = 0.50. A similar but consistently lower uncertainty level

was observed for the GeoPropertyOnly setting, where the corresponding mean Monte Carlo standard deviations were 41.41, 80.60, and 132.66 kPa, respectively. This indicates that perturbing all input variables produces a broader uncertainty envelope that better reflects the combined variability of binder dosage, curing days, consistency index, and compaction variables.

Table 11. Monte Carlo uncertainty metrics for the CatBoost-6PG model under different perturbation settings and intensity levels

Perturbation mode	Scenario	k	PICP90	PICP95	MPIW95 (kPa)	PINAW95	ACE95	CWC95	Mean MC std. (kPa)	Mean MC CV (%)
FullInput	Low	0.1	0.394	0.489	195.37	0.101	-0.461	$1.04e^{+09}$	50.18	5.81
FullInput	Moderate	0.25	0.694	0.756	408.54	0.211	-0.194	$3.52e^{+03}$	108.65	12.39
FullInput	High	0.5	0.867	0.894	710.63	0.367	-0.056	6.26	190.14	21.11
GeoPropertyOnly	Low	0.1	0.361	0.406	156.92	0.081	-0.544	$5.38e^{+10}$	41.41	4.65
GeoPropertyOnly	Moderate	0.25	0.556	0.644	295.02	0.152	-0.306	$6.57e^{+05}$	80.6	8.79
GeoPropertyOnly	High	0.5	0.756	0.778	473.56	0.244	-0.172	$1.34e^{+03}$	132.66	14.13

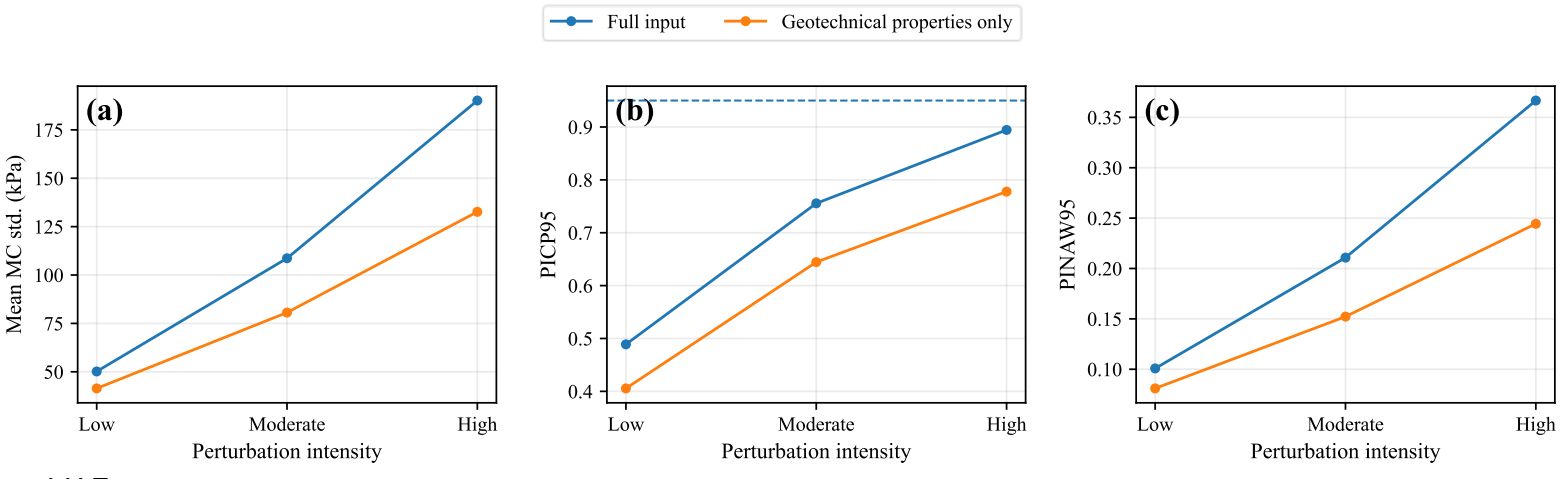


Figure 15. Scenario-level uncertainty comparison for the CatBoost-6PG model: **(a)** mean Monte Carlo standard deviation; **(b)** PICP95; and **(c)** PINAW95 under different perturbation intensities

The improvement in interval coverage is shown in Figure 15b. For the FullInput setting, PICP95 increased from 0.489 in the low-perturbation scenario to 0.756 in the moderate scenario and 0.894 in the high scenario. In contrast, the GeoPropertyOnly setting produced lower PICP95 values of 0.406, 0.644, and 0.778 for the same perturbation levels. This shows that perturbing only LL, PI, OMC, and MDD underestimates the total predictive uncertainty. Including NS, QL, and curing days provides a more complete representation of input-driven uncertainty because these variables directly affect the binder dosage, curing-related reactions, and the derived physics-guided features.

This improvement in coverage was accompanied by wider prediction intervals. As shown in Figure 15c, the FullInput PINAW95 increased from 0.101 at k = 0.10 to 0.211 at k = 0.25, and 0.367 at k = 0.50. For GeoPropertyOnly, the corresponding PINAW95 values were lower, increasing from 0.081 to 0.152 and 0.244. Therefore, the FullInput-High scenario provided the most reliable interval coverage among the evaluated scenarios, while the GeoPropertyOnly scenarios produced sharper but less reliable prediction intervals. This behavior reflects the expected reliability-sharpness trade-off: broader intervals increase the

probability of capturing measured UCS values, whereas narrower intervals may fail to represent the full experimental variability.

The uncertainty bands for the high perturbation level are shown in Figure 16. In the FullInput case (Figure 16a), the 95% prediction interval covered 89.4% of the measured UCS values, with an average interval width of 710.63 kPa. In contrast, the GeoPropertyOnly case (Figure 16b) covered only 77.8% of the samples, with a narrower average interval width of 473.56 kPa. The wider FullInput interval is therefore more conservative and more reliable, while the GeoPropertyOnly interval is sharper but under-represents the uncertainty associated with stabilizer dosage and curing days.

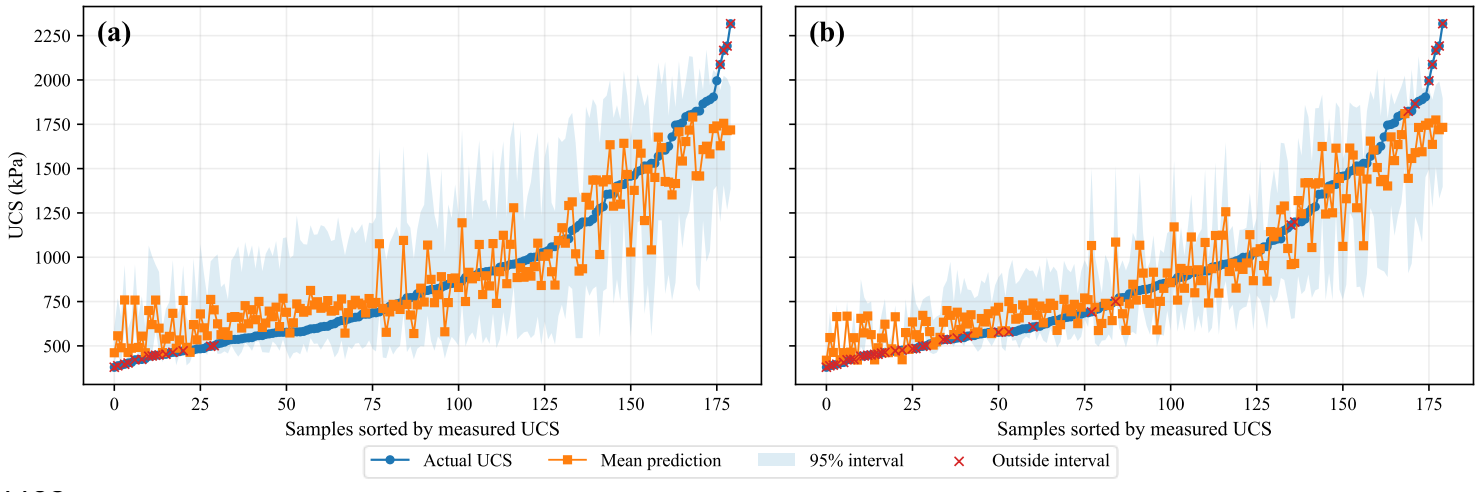


Figure 16. Monte Carlo 95% prediction intervals for the CatBoost-6PG model under the high perturbation level: **(a)** FullInput perturbation; **(b)** GeoPropertyOnly perturbation

Figure 17 compares input-driven uncertainty with model-driven uncertainty. The FullInput Monte Carlo uncertainty band under the high perturbation scenario is shown in Figure 17a, while the group-bootstrap model uncertainty band is shown in Figure 17b. The corresponding group-bootstrap uncertainty metrics are summarized in Table 12. The group-bootstrap analysis used 50 group-level resampled models per fold and produced a PICP95 of 0.600, MPIW95 of 279.61 kPa, and mean bootstrap standard deviation of 78.31 kPa. These values are substantially lower than those obtained from the FullInput-High Monte Carlo scenario, which produced PICP95 = 0.894, MPIW95 = 710.63 kPa, and mean MC standard deviation = 190.14 kPa.

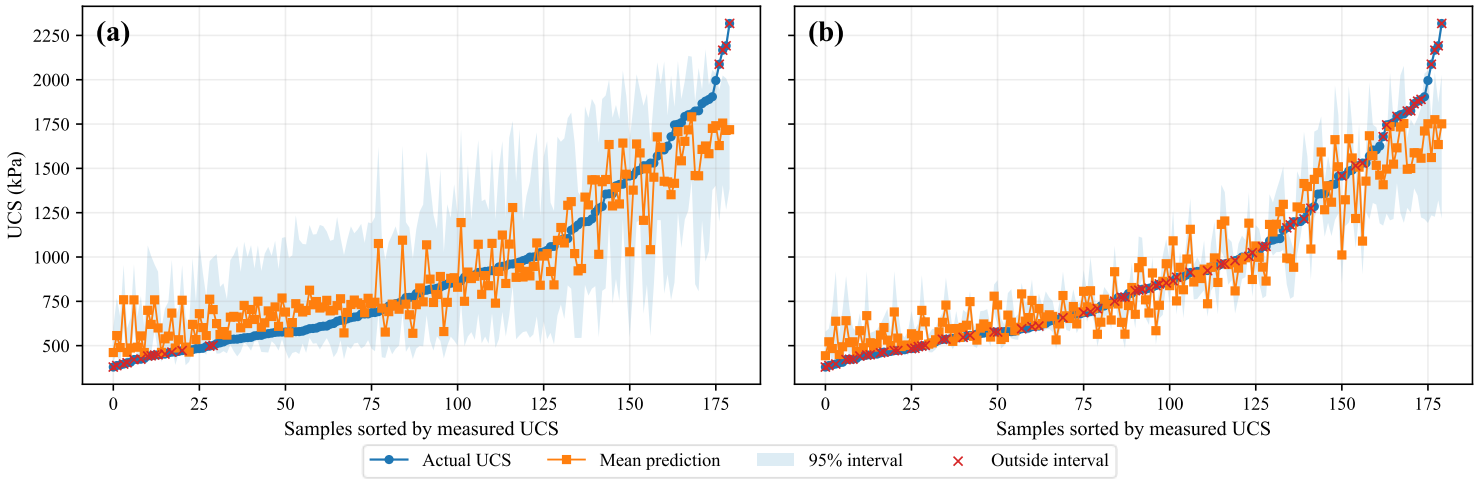


Figure 17. Comparison between input-driven and model-driven uncertainty: **(a)** FullInput Monte Carlo uncertainty under the high perturbation scenario; **(b)** group-bootstrap model uncertainty

Table 12. Group-bootstrap model uncertainty metrics for the CatBoost-6PG model

Metric	Value
Bootstrap type	Group-level bootstrap within each outer-training fold
Bootstrap models per fold	50
PICP90	0.528
PICP95	0.6
MPIW95 (kPa)	279.61
PINAW95	0.144
ACE95	-0.35
CWC95	$5.75e^{+06}$
Mean bootstrap std. (kPa)	78.31
R^2 of bootstrap mean	0.881
RMSE of bootstrap mean (kPa)	157.55
MAE of bootstrap mean (kPa)	115.53
MAPE of bootstrap mean (%)	13.01

The comparison between Monte Carlo and bootstrap results indicates that model instability alone cannot fully explain the prediction uncertainty. The group-bootstrap intervals were considerably narrower and had much lower coverage than the FullInput Monte Carlo intervals. Therefore, the dominant source of uncertainty in this study is the variability of input variables and their nonlinear propagation through the physics-guided feature space, rather than instability of the fitted CatBoost-6PG model alone.

Overall, the uncertainty analysis shows that CatBoost-6PG provides strong deterministic predictions, while the FullInput Monte Carlo scenario offers the most informative uncertainty envelope for reliability assessment. However, even the highest PICP95 value remained below the nominal 0.95 level. Therefore, the prediction intervals should be interpreted as informative but slightly under-calibrated. This finding highlights the need to account for material variability, stabilizer dosage uncertainty, and curing-related variability when using ML models for reliability-oriented UCS prediction.

4.5. Engineering Implications and Sustainability Considerations

The integration of physics-guided engineered features into the ML framework provides advantages beyond predictive accuracy. By incorporating physically meaningful variables such as total binder content, binder-curing interactions, nonlinear curing effects, relative plasticity, and compaction-state indicators, the selected CatBoost-6PG model offers a more interpretable representation of UCS development than a purely black-box data-driven model. This is particularly important in geotechnical engineering, where model predictions must be consistent with known stabilization mechanisms to be useful for design support.

From an engineering perspective, the proposed framework can assist preliminary UCS estimation and mix-design screening. Conventional UCS testing is time-consuming and resource-intensive, especially when several stabilizer combinations and curing days must be evaluated. The developed model provides rapid UCS estimates using mixture variables together with basic consistency and compaction-related properties. Therefore, it should be viewed as a decision-support tool after basic soil characterization, rather than as a replacement for laboratory testing.

A key practical advantage of the physics-guided approach is its ability to represent interaction-based effects. Features such as NS_Curing and BinderTotal_Curing explicitly account for the coupling between stabilizer dosage and curing days. The SHAP, sensitivity, and partial dependence analyses showed that UCS development is controlled not only by individual variables, but also by the interaction among binder content, curing days, and plasticity. This confirms that strength gain in NS-QL stabilized kaolin clay is governed by time-dependent physicochemical processes rather than by isolated mixture parameters alone.

Compared with models based only on raw input variables, the CatBoost-6PG formulation embeds important geotechnical relationships directly into the feature space. This can help the model learn more efficiently from a relatively small experimental dataset and reduce the risk of non-physical trends. The physically consistent responses observed in SHAP dependence plots, sensitivity curves, and partial dependence contours indicate that the model captures meaningful stabilization behavior, including the positive effect of curing days, the limiting influence of plasticity, and the interaction-dependent contribution of NS.

The model also provides useful guidance for material optimization. The results indicate that increasing stabilizer content does not necessarily produce proportional strength improvement across the entire investigated range. Instead, the predicted UCS response depends on balanced NS-QL proportions, curing days, and the plasticity state of the soil. This can help engineers identify efficient mix-design regions that achieve target strength while avoiding unnecessary binder addition.

From a sustainability perspective, reducing excessive binder use can contribute to more efficient resource utilization and may help lower the environmental burden associated with chemical stabilization. Quicklime and cementitious binders are typically associated with energy consumption and carbon emissions during production [47]. Although this study did not conduct a life-cycle assessment, the proposed framework can support sustainability-oriented decision making by helping identify mixtures that provide adequate UCS with more efficient use of stabilizing agents. In this context, NS may serve as a supplementary stabilizing material, while the model helps evaluate its contribution under different curing and QL conditions.

The uncertainty analysis further highlights the importance of reliability-aware design. The comparison between Monte Carlo input perturbation and group-bootstrap model uncertainty showed that input-driven uncertainty was the dominant contributor to prediction intervals. This indicates that engineering decisions should not rely only on deterministic UCS predictions, especially when material properties, stabilizer dosages, or curing conditions may vary in practice. The use of prediction intervals enables a more cautious interpretation of model outputs and supports safer preliminary design decisions.

Overall, the proposed framework bridges data-driven modeling and geotechnical mechanism-based interpretation. The CatBoost-6PG model provides a practical tool for UCS prediction, while SHAP analysis, sensitivity analysis, partial dependence analysis, and uncertainty quantification improve transparency and reliability. This combination supports more informed decision-making for NS-QL soil stabilization, particularly during preliminary mix design and evaluation of strength-development trends.

5. Limitations and Future Work

Despite the competitive performance and physical interpretability of the developed framework, several limitations should be acknowledged. First, the dataset is relatively small and is based on laboratory-scale experiments. Although the nested GroupKFold validation strategy was used to obtain a conservative estimate of model generalization, the dataset may not fully represent field-scale variability, including environmental conditions, compaction heterogeneity, curing environment, moisture fluctuations, and construction-related uncertainties. Therefore, the direct application of the model to field conditions should be performed with caution and preferably supported by additional validation data.

Second, the developed model is specific to kaolin clay stabilized with NS and QL within the investigated ranges of mixture composition, curing days, consistency limits, and compaction properties. Its applicability to other soil types, stabilizers, curing regimes, or strength ranges cannot be assumed without further calibration or external validation. In particular, extrapolation beyond the experimental domain may lead to unreliable predictions, even when the model provides physically meaningful trends within the training range.

Third, the physics-guided engineered features used in this study represent simplified variables of complex stabilization mechanisms. Features such as Binder_Total, NS_Curing, BinderTotal_Curing, log_Curing, MDD_OMC_Ratio, and PI_LL_Ratio improve interpretability and help encode relevant geotechnical relationships, but they do not explicitly model the underlying chemical reactions, mineralogical transformations, pore-structure evolution, or microstructural bonding mechanisms. Therefore, these features should be interpreted as mechanism-inspired proxies rather than full physical representations of the stabilization process.

Fourth, the uncertainty analysis was based on correlated Gaussian input perturbations estimated from the outer-training folds. Although this approach provides a leakage-safe and structured way to propagate input variability, it may not fully capture non-Gaussian behavior, spatial variability, measurement uncertainty, or field-dependent correlations among geotechnical parameters. In addition, the highest PICP95 remained below the nominal 0.95 level, indicating that the prediction intervals were informative but slightly under-calibrated. Future uncertainty frameworks should therefore consider non-Gaussian distributions, Bayesian methods, conformal prediction, or calibration-based approaches for more reliable interval estimation.

Finally, the sensitivity analysis was performed using a one-at-a-time strategy around a median baseline condition. This approach is useful for interpreting the local response of the model, but it does not fully capture higher-order interactions among input variables. Although SHAP and 2D partial dependence analyses were used to examine interaction effects, future work could extend this analysis using global variance-based sensitivity methods or multi-dimensional response-surface exploration.

Future work should focus on expanding the dataset by including additional soil types, stabilizer combinations, curing conditions, and field-scale measurements. Incorporating microstructural, mineralogical, and chemical descriptors could improve the physical basis of the model and provide deeper insight into the stabilization mechanisms. Further research should also evaluate the proposed framework using independent external datasets and compare different uncertainty quantification strategies. Extending the model toward multi-output prediction, reliability-based mix design, and optimization of strength-cost-environmental trade-offs would further enhance its practical value for sustainable geotechnical engineering applications.

6. Conclusions

This study developed an interpretable and uncertainty-aware machine learning framework supported by mechanism-inspired, physics-guided feature engineering for predicting the UCS of kaolin clay stabilized with nano shale and quicklime. A laboratory dataset containing 180 UCS measurements was used, with NS content, QL content, curing days, LL, PI, OMC, and MDD selected as the raw input variables. Six machine learning models, including MLR, SVR, RF, XGBoost, CatBoost, and BP-ANN, were optimized and evaluated using a nested GroupKFold validation strategy. Physics-guided engineered features were then introduced to represent binder dosage, binder-curing interactions, nonlinear curing effects, compaction state, and relative plasticity. Based on the obtained results, the following conclusions can be drawn:

- A nested GroupKFold validation strategy was used to evaluate model generalization under group-held-out mixture-curing conditions. Among the evaluated models, CatBoost-6PG achieved the best overall performance, with $R^2 = 0.841 \pm 0.123$, RMSE = 151.11 ± 34.94 kPa, MAE = 110.67 ± 26.04 kPa, MAPE = 11.97 ± 3.75%, and VAF = 87.38 ± 7.24%.
- The ablation study showed that the initial seven-feature physics-guided configuration contained redundant information. Removing QL_Curing improved the feature structure, leading to the final CatBoost-6PG model with six engineered features: Binder_Total, NS_Curing, BinderTotal_Curing, log_Curing, MDD_OMC_Ratio, and PI_LL_Ratio.
- The reduced-input comparison showed that NS, QL, and curing days can support preliminary UCS screening, but the CatBoost-6PG model achieved higher accuracy and stability by incorporating geotechnical index, compaction, and selected mechanism-inspired variables.
- SHAP, sensitivity analysis, and partial dependence analysis consistently showed that UCS prediction is mainly controlled by plasticity, curing days, and binder-curing interaction effects. Higher PI reduced predicted UCS, while longer curing days and favorable binder-curing combinations increased strength.
- The effects of NS and QL were nonlinear and interaction-dependent rather than purely dosage-controlled. The PDP results indicated that higher predicted UCS values are associated with longer curing days, lower plasticity, and balanced NS-QL combinations.
- The uncertainty analysis showed that input-driven uncertainty was more influential than model-instability uncertainty. The FullInput Monte Carlo scenario at k = 0.50 produced PICP95 = 0.894,

whereas group-bootstrap model uncertainty produced PICP95 = 0.600, highlighting the importance of accounting for material, stabilizer dosage, and curing variability.

- Although the prediction intervals were informative, the highest PICP95 remained below the nominal 0.95 level, indicating slight under-calibration. Future studies should therefore consider more advanced uncertainty calibration methods, such as conformal prediction or Bayesian approaches.
- From an engineering perspective, the proposed framework can support preliminary UCS estimation, mix-design screening, and reliability-aware decision making. It does not replace confirmatory laboratory testing, but it can help reduce the number of trial mixtures during early-stage design.
- The proposed model should be used within the investigated material, dosage, and curing ranges. Application to other soils, stabilizers, curing regimes, or field conditions requires further calibration and external validation.

The results demonstrate that combining physics-guided feature engineering, group-aware validation, explainable machine learning, and uncertainty quantification provides a practical and interpretable framework for UCS prediction of NS-QL stabilized kaolin clay.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions: CRediT

Arash Aminaee: Writing – Original Draft; Software; Methodology; Formal Analysis; Visualization; Investigation; Writing – Review & Editing. **Abolfazl Soltani:** Supervision; Conceptualization; Methodology; Software; Project Administration; Scientific Leadership; Interpretation of Results; Writing – Review & Editing. **Abolfazl Baghbani:** Validation; Writing – Review & Editing. **Parisa Salehan:** Software; Visualization; Writing – Review & Editing. **Hossam Abuel-Naga:** Methodology; Writing – Review & Editing. **Pijush Samui:** Conceptualization; Scientific Input; Methodology.

Data Availability Statement

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Astaneh AG, Abbasi M, Hosseinpour I, Payan M, & Salimi M. Mechanical, Durability, and microstructural study of dispersive soil stabilized by alkali-activated rice husk ash and natural fiber. *Results in Engineering* 29:108679. https://doi.org/10.1016/j.rineng.2025.108679. (2026).

[2] Syed M, GuhaRay A, Chukka SK, & Ahmad S. A laboratory investigation and numerical modeling on fiber reinforced lime and alkaline binder stabilized pavement subgrade soil. *Case Studies in Construction Materials* 20:e03000. https://doi.org/10.1016/j.cscm.2024.e03000. (2024).

[3] Sun X, Yeon Ting MZ, Qin J, & Yi Y. Sustainable production of liquefied stabilized soil using waste clay, fine incineration bottom ash, carbide sludge, and GGBS for backfilling applications. *Journal of Rock Mechanics and Geotechnical Engineering* 18:2238–50. https://doi.org/10.1016/j.jrmge.2025.04.006. (2026).

[4] Yu W, Zhiduo Z, Dingwen Z, He S, & Yan L. Research on strength and deformation properties of high moisture soft soils stabilized by GGBS, cement, fly ash, and sodium silicate. *Constr. Build. Mater.* 448:138244. https://doi.org/10.1016/j.conbuildmat.2024.138244. (2024).

[5] Chen Q, Hu G, & Wu J. Comparative study on the prediction of the unconfined compressive strength of the one-part geopolymer stabilized soil by using different hybrid machine learning models. *Case Studies in Construction Materials* 21:e03439. https://doi.org/10.1016/j.cscm.2024.e03439. (2024).

[6] Luo Z, Xue X, Xiong C, Huang R, Du J, Li J, et al. Data-driven prediction of unconfined compressive strength in stabilized soils using machine learning. *Discover Applied Sciences* 7:1122. https://doi.org/10.1007/s42452-025-07766-9. (2025).

[7] Toualbia Y, Sari-Ahmed B, La Porta G, Makhlouf A, Ghrici M, & Pirulli M. Unconfined compressive strength prediction of lime-stabilized clayey soils under freeze-thaw conditions. *Environmental Geotechnics* 12:517–30. https://doi.org/10.1680/jenge.24.00109. (2025).

[8] Waleed M, & Alshawmar F. Enhancing mechanical properties of low plasticity soil through coal and silica fume stabilization. *Sci. Rep.* 15:9990. https://doi.org/10.1038/s41598-025-94149-0. (2025).

[9] Soltani A, Abed P, & Aminaee A. Experimental studies of kaolin soil improved by calcined sepiolite as a partial replacement: mechanical performances and micro-scale analysis. *Indian Geotechnical Journal* 55:750–65. https://doi.org/10.1007/s40098-024-00913-w.(2025).

[10] Khatti J, & Grover KS. Prediction of UCS of fine-grained soil based on machine learning Part 1: Multivariable regression analysis, Gaussian process regression, and Gene expression programming. Multiscale and multidisciplinary modeling experiments and design 6:199–222. https://doi.org/10.1007/s41939-022-00137-6. (2023).

[11] Thapa I, & Ghani S. Advancing earth science in geotechnical engineering: A data-driven soft computing technique for unconfined compressive strength prediction in soft soil. *Journal of Earth System Science* 133:159. https://doi.org/10.1007/s12040-024-02374-4. (2024).

[12] Onyelowe KC, Moghal AAB, Ebid A, Rehman AU, Hanandeh S, & Priyan V. Estimating the strength of soil stabilized with cement and lime at optimal compaction using ensemble-based multiple machine learning. *Sci. Rep.* 14:15308. https://doi.org/10.1038/s41598-024-66295-4. (2024).

[13] Sari-Ahmed B, Benzaamia A, Ghrici M, & Moghal AAB. Strength prediction of fiber-reinforced clay soils stabilized with lime using XGBoost machine learning. *Civil and Environmental Engineering Reports* 34:157–76. https://doi.org/10.59440/ceer/190062. (2024).

[14] Mohammed AMA, Husain O, Abdulkareem M, Mohd Yunus NZ, Jamaludin N, Mutaz E, et al. Explainable artificial intelligence for predicting the compressive strength of soil and ground granulated blast furnace slag mixtures. *Results in Engineering* 25:103637. https://doi.org/10.1016/j.rineng.2024.103637. (2025).

[15] Zeini HA, Al-Jeznawi D, Imran H, Bernardo LFA, Al-Khafaji Z, & Ostrowski KA. Random forest algorithm for the strength prediction of geopolymer stabilized clayey soil. *Sustainability* 15:1408. https://doi.org/10.3390/su15021408. (2023).

[16] Tiwari P, & Tung S. Data-driven prediction and optimization of unconfined compressive strength in lime-stabilized soils using explainable machine learning. *Asian Journal of Civil Engineering,* https://doi.org/10.1007/s42107-026-01708-5. (2026).

[17] Yuan B, Choo CS, Yeo LY, Wang Y, Yang Z, Guan Q, et al. Physics-informed machine learning in geotechnical engineering: A direction paper. *Geomechanics and Geoengineering* 20:1128–59. https://doi.org/10.1080/17486025.2025.2502029. (2025).

[18] Minasny B, Bandai T, Ghezzehei TA, Huang Y-C, Ma Y, McBratney AB, et al. Soil science-informed machine learning. *Geoderma* 452:117094. https://doi.org/10.1016/j.geoderma.2024.117094. (2024).

[19] Yang H, Hu G, Liu L, Li Y, Deng Y, & Wu J. A physics-informed and SHAP-enhanced modeling framework for predicting strength of Ccement-stabilized soil. *Case Studies in Construction Materials* 23:e05280. https://doi.org/10.1016/j.cscm.2025.e05280. (2025).

[20] Teodoru I-B, Owusu-Yeboah Z, Aniculăesi M, Dascălu AV, Hörtkorn F, Amelio A, et al. Prediction of unconfined compressive strength in cement-treated soils: A machine learning approach. *Applied Sciences* 15:7022. https://doi.org/10.3390/app15137022. (2025).

[21] Baghbani A, Soltani A, Kiany K, & Daghistani F. Predicting the strength performance of hydrated-lime activated rice husk ash-treated soil using two grey-box Mmachine learning models. *Geotechnics* 3:894–920. https://doi.org/10.3390/geotechnics3030048. (2023).

[22] Abdullah GMS, Ahmad M, Babur M, Badshah MU, Al-Mansob RA, Gamil Y, et al. Boosting-based ensemble machine learning models for predicting unconfined compressive strength of geopolymer stabilized clayey soil. *Sci. Rep.* 14:2323. https://doi.org/10.1038/s41598-024-52825-7. (2024).

[23] Aminaee A, Ardakani A, Baghbani A, Abuel-Naga H, & Daghistani F. Hybrid ML–XAI Framework for predicting and interpreting the strength of lime–silica fume stabilized clay for sustainable construction applications. *Buildings* 16:953. https://doi.org/10.3390/buildings16050953. (2026).

[24] Zhang A, Wang L, Zhang W, Zhao C, & Zhang P. Bayesian neural network prediction and uncertainty analysis of bio-cemented soil strength. *Int J. Numer Anal Methods Geomech* 50:1349–66. https://doi.org/10.1002/nag.70170. (2026).

[25] Wan W. Enhancing unconfined compressive strength of stabilized soil with lime and cement prediction through a robust hybrid machine learning approach utilizing Naive Bayes algorithm. *Journal of Engineering and Applied Science* 71:84. https://doi.org/10.1186/s44147-024-00408-8. (2024).

[26] Soltani A, Aminaee A, & Abed P. Improving mechanical properties of low-strength kaolin clay by mixing nano hydro carbonated shale and quicklime. *International Journal of Geosynthetics and Ground Engineering* 10:53. https://doi.org/10.1007/s40891-024-00567-x. (2024).

[27] Hu J, Li Z, Qin X, Dong S, Zhao T, & Li H. Machine learning-based optimization of geopolymer-stabilized soil formulation using blast furnace slag and fly ash. *Mater. Today Commun.* 48:113500. https://doi.org/10.1016/j.mtcomm.2025.113500. (2025).

[28] Salehi M, Bayat M, Saadat M, & Nasri M. Prediction of unconfined compressive strength and California bearing capacity of cement- or lime-pozzolan-stabilised soil admixed with crushed stone waste. *Geomechanics and Geoengineering* 18:272–83. https://doi.org/10.1080/17486025.2022.2040606. (2023).

[29] Noaman MF, Haq M, Khan MA, Ali K, & Kamyab H. Geotechnical and microstructural analysis of high-volume fly ash stabilized clayey soil and machine learning application. *Case Studies in Construction Materials* 21:e03628. https://doi.org/10.1016/j.cscm.2024.e03628. (2024).

[30] Zhang C, Zhu Z, Liu F, Yang Y, Wan Y, Huo W, et al. Efficient machine learning method for evaluating compressive srength of cement stabilized soft soil. *Constr Build. Mater.* 392:131887. https://doi.org/10.1016/j.conbuildmat.2023.131887. (2023).

[31] ASTM International. ASTM C618-22: Standard Specification for Coal Fly Ash and Raw or Calcined Natural Pozzolan for Use in Concrete. vol. 04.02. West Conshohocken, PA: *ASTM International* https://doi.org/10.1520/C0618-22. (2022).

[32] ASTM International. ASTM D6276-19: Standard Test Method for Using pH to Estimate the Soil-Lime Proportion Requirement for Soil Stabilization. vol. 04.09. West Conshohocken, PA: *ASTM International* https://doi.org/10.1520/D6276-19. (2019).

[33] Pasupuleti LN, Penugonda SK, Danikonda AK, & Jyothula H. Composite machine learning models for forecasting UCS of stabilized lateritic soils. multiscale and multidisciplinary modeling. *Experiments and Design* 9:113. https://doi.org/10.1007/s41939-026-01200-2. (2026).

[34] Smola AJ, & Schölkopf B. A tutorial on support vector regression. *Stat. Comput.* 14:199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88. (2004).

[35] Breiman L. & Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324. (2001).

[36] Chen T, & Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining, New York, NY, USA: *ACM* p. 785–94. https://doi.org/10.1145/2939672.2939785. (2016).

[37] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, & Gulin A. CatBoost: Unbiased boosting with categorical features. 32nd Conference on neural information processing systems, *Montréa*, p. 6638–48. (2018).

[38] Rumelhart DE, Hinton GE, & Williams RJ. Learning representations by back-propagating errors. *Nature* 323:533–6. https://doi.org/10.1038/323533a0. (1986).

[39] Poli R, Kennedy J, & Blackwell T. Particle swarm optimization. swarm intelligence 1:33–57. https://doi.org/10.1007/s11721-007-0002-0. (2007).

[40] Xiong J, Gong Y, Liu X, Li Y, Chen L, Liao C, et al. A PSO-XGBoost model for predicting the compressive strength of cement–soil mixing pile considering field environment simulation. *Buildings* 15:2740. https://doi.org/10.3390/buildings15152740. (2025).

[41] Syed M, Ashfaq M, Jamhiri B, Ali U, & Jalal FE. Sustainable engineering of fiber-reinforced geopolymer-treated low plasticity clay linking geomechanics and microstructure through support vector machines. *Sci. Rep.* 15:44399. https://doi.org/10.1038/s41598-025-28438-z. (2025).

[42] Lundberg S, & Lee S-I. A unified approach to interpreting model predictions. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. 31st Conference on neural information processing systems, Red Hook, NY: *Curran Associates Inc.* p. 4768–77. (2017).

[43] Taylor KE. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research Atmospheres* 106:7183–92. https://doi.org/10.1029/2000JD900719. (2001).

[44] Beygi L, & Khazaei J. Soft clay eco-friendly improvement by ground granulated blast furnace slag and quicklime. *Geotechnical and Geological Engineering* 42:2061–74. https://doi.org/10.1007/s10706-023-02661-9. (2024).

[45] Majd Rahimabadi M, & Arabani M. Impact of lime kiln dust on the mechanical properties and freeze-thaw durability of nanoclay-stabilized kaolin soil. *Case Studies in Construction Materials* 24:e05774. https://doi.org/10.1016/j.cscm.2026.e05774. (2026).

[46] Manaviparast HR, Cristelo N, Pereira E, & Miranda T. A comprehensive review on clay soil stabilization using rice husk ash and lime sludge. *Applied Sciences* 15:2376. https://doi.org/10.3390/app15052376. (2025).

[47] Al-Subari BL, Hilal A, & Ekinci A. Life cycle assessment of soil stabilization using cement and waste additives. *Constr. Build. Mater.* 403:133045. https://doi.org/10.1016/j.conbuildmat.2023.133045. (2023).