

Interpretable artificial intelligence approach for understanding shear strength in stabilized clay soils using real field soil samples

Mohamed Noureldin^{a*}, Aghyad Al Kabbani^b, Alejandra Lopez^a, Leena Korkiala-Tanttu^a

^a Civil Engineering Department, Aalto University, Espoo 02150, Finland

^b Computer Science Department, Aalto University, Espoo 02150, Finland

*Corresponding author. E-mail: Mohamed.noureldin@aalto.fi

© The Author(s) 2025, corrected publication 2025

ABSTRACT Deep mixing, also known as deep stabilization, is a widely used ground improvement method in Nordic countries, particularly in urban and infrastructural projects, aiming to enhance the properties of soft, sensitive clays. Understanding the shear strength of stabilized soils and identifying key influencing factors are essential for ensuring the structural stability and durability of engineering structures. This study introduces a novel explainable artificial intelligence framework to investigate critical soil properties affecting shear strength, utilizing a data set derived from stabilization tests conducted on laboratory samples from the 1990s. The proposed framework investigates the statistical variability and distribution of crucial parameters affecting shear strength within the collected data set. Subsequently, machine learning models are trained and tested to predict soil shear strength based on input features such as water/binder ratio and water content, etc. Global model analysis using feature importance and Shapley additive explanations is conducted to understand the influence of soil input features on shear strength. Further exploration is carried out using partial dependence plots, individual conditional expectation plots, and accumulated local effects to uncover the degree of dependency and important thresholds between key stabilized soil parameters and shear strength. Heat map and feature interaction analysis techniques are then utilized to investigate soil properties interactions and correlations. Lastly, a more specific investigation is conducted on particular soil samples to highlight the most influential soil properties locally, employing the local interpretable model-agnostic explanations technique. The validation of the framework involves analyzing laboratory test results obtained from uniaxial compression tests. The framework demonstrates an ability to predict the shear strength of stabilized soil samples with an accuracy surpassing 90%. Importantly, the explainability results underscore the substantial impact of water content and the water/binder ratio on shear strength.

KEYWORDS explainable artificial intelligence, Shapley additive explanations, local interpretable model-agnostic explanations, partial dependence plots, stabilized soil, water/binder ratio, water content, shear strength

1 Introduction

The significance of soil stabilization, particularly in clay soils, is crucial for the construction and maintenance of various infrastructures and engineering structures. Clay soils are known for their low bearing capacity, high compressibility, and susceptibility to erosion and sloughing, which can lead to the failure of engineering structures. Proper soil stabilization can also prevent

differential settlement, which can cause significant damage to structures and reduce their service life [1]. Therefore, stabilizing clay soils is essential for ensuring the stability, safety, and durability of infrastructures such as embankments, pipelines, and retaining structures [2].

Deep soil mixing, or deep stabilization method, a widely utilized ground improvement method in Nordic countries [3], has gained increased popularity in recent decades due to expanding construction activities into areas with poor quality soils such as soft clays, muds, and silts. This method involves mixing soft soil with a

selected binder to form column-like structures, typically reaching depths of up to 25 m, although columns under 15 m are more common. While Deep soil mixing can be carbon-intensive and costly when lime-cement binders are used, ongoing research focuses on low-carbon binder alternatives [4]. Prior to field stabilization, determining the appropriate binder type and identifying the most influential soil properties through small-scale laboratory tests is crucial and cost-effective. Advanced techniques such as machine learning (ML) models are essential for highlighting key soil parameters affecting the shear strength of stabilized soil.

ML techniques have emerged as a powerful tool for predicting the shear strength of clay soils [5,6]. These data-driven approaches offer advantages over traditional methods by accounting for complex nonlinear relationships between various soil properties and shear strength [7,8]. For example, Ahmad et al. [9] predicted unconfined compressive strength of expansive clay soil treated with hydrated-lime-activated rice husk ash using several ML techniques. Ly and Pham [10] used direct shear test and support vector machine model to predict soil shear strength. Taffese and Abegaz [11] used optimizable ensemble method and artificial neural network to monitor amended soil reliability for a housing development program. Sert et al. [12] found that decision tree regression is the most accurate ML model to estimate stress and strain of expansive soils stabilized with basalt fiber. Moreover, ML techniques extended to include soil structure interaction applications [13–15].

The integration of ML techniques in geotechnical engineering is rapidly evolving, transforming traditional practices and offering innovative solutions to complex engineering challenges. Phoon and Zhang [16] highlight the future potential of ML in geotechnics, emphasizing how data-driven models could improve predictive accuracy in subsurface investigations. This aligns with the work of Karniadakis et al. [17], who introduced the concept of physics-informed machine learning (PIML), where models incorporate physical laws to enhance predictions in scenarios with limited data—a critical factor in geotechnical applications. Zhang and Goh [18,19] explored the applicability of multivariate adaptive regression splines (MARS) and neural networks for geotechnical analysis, particularly in predicting parameters such as pile drivability and soil behavior, illustrating ML's potential in optimizing site investigations. Recent advancements further illustrate this trend; Zhang et al. [20] developed a physics-informed deep learning model to predict tunneling-induced ground deformations, achieving significant accuracy by embedding geotechnical constraints within the model. Additionally, Zhou et al. [21] leveraged knowledge-based statistics for simulating soil stratigraphy, showing the benefits of hybrid approaches in simulating complex

subsurface conditions. Zhang et al. [22] employed Bayesian optimization to enhance the performance of ensemble learning models for predicting soil strength parameters, underscoring the role of ML in handling data variability and improving model reliability. Collectively, these studies reveal how ML and hybrid methods are redefining predictive analytics in geotechnics, making them essential tools in modern engineering practices.

Recently, explainable artificial intelligence (XAI) has emerged as a pivotal domain in ML, offering transparency and comprehensibility in artificial intelligence (AI) models' decision-making processes. XAI refers to the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its primary goal is to make the behavior of these systems transparent and comprehensible to humans by elucidating the underlying mechanisms of their decision-making processes. Different XAI realization strategies are used in literature. For example, feature Importance (FIM) and Shapley additive explanations (SHAP) assess global FIM, with SHAP providing deeper insights by accounting for interactions and nonlinearities [23,24]. Partial dependence plots (PDP) and individual conditional expectation (ICE) plots reveal feature effects on predictions—PDP shows average effects, while ICE captures individual variations. Accumulated local effects (ALE) further detail nonlinear and interactive feature impacts [23]. Heat maps (HM) and feature interaction analysis (FIA) visualize and identify key feature interactions. Finally, Local interpretable model-agnostic explanations (LIME) offers local explanations, highlighting influential features for each soil sample [23]. Together, these methods provide a thorough understanding of factors driving shear strength predictions.

Anysz et al. [25] demonstrated XAI's capability in analyzing the compressive strength of rammed earth by identifying the key features affecting structural integrity, thus guiding material selection and composition. Naser [26] outlined fundamental XAI techniques, emphasizing their value in enhancing causal inference and reducing interpretative biases—factors essential for engineering applications. The application of XAI in fire safety research by Al-Bashiti and Naser [27] and Tapeh and Naser [28] further underscored its effectiveness, where it enabled the validation of domain-specific theories on fire-induced spalling in concrete, advancing safety analysis through interpretable models. In a similar vein, Naumets and Lu [29] explored regression trees as an explainable tool for construction engineering, finding that such models effectively balance accuracy and interpretability, making them suitable for complex construction assessments.

XAI's potential in supporting risk-based decision-making in construction was highlighted by Zhan et al. [30], who used counterfactual explanations to aid in

interpreting risk predictions, offering stakeholders insights into potential failure factors. In sustainable materials research, Ibrahim et al. [31] applied XAI to model the compressive strength of ternary cement concrete, demonstrating the potential for XAI in sustainable design by providing a transparent analysis of key variables influencing strength. Similarly, Shabbir et al. [23] and Yossef et al. [24] presented XAI frameworks for seismic and composite structural assessments, respectively, enabling engineers to derive interpretable predictions for safer, more robust design practices. Das and Rad [32] provide a comprehensive survey highlighting the opportunities and challenges within the realm of XAI, emphasizing its significance across various domains. Longo et al. [33] delve into the concepts, applications, research challenges, and visions of XAI, shedding light on its multifaceted implications. Notably, Noureldin et al. [34] proposed an XAI-based probabilistic framework for seismic assessment that leverages prediction intervals to quantify uncertainties, making XAI a vital tool in assessing structural resilience under variable conditions. These studies collectively demonstrate that XAI is transforming engineering applications by enhancing model transparency, supporting safer and more sustainable designs, and fostering greater trust in ML-driven decision-making processes across diverse engineering domains.

In geotechnical engineering, XAI holds immense potential for improving understanding and predicting complex soil behaviors. Nasiri et al. [35] demonstrated the applicability of XAI in predicting the uniaxial compressive strength and modulus of elasticity for travertine samples, showcasing its efficacy in geophysical sciences. Furthermore, Dahal and Lombardo [36] presented a pioneering glimpse into the future of landslide susceptibility modeling, leveraging XAI techniques to enhance the interpretability of predictive models in geoscience. Specifically, in the context of geotechnical risk analysis, Liu et al. [37] elucidated the importance of causal discovery and reasoning facilitated by XAI methodologies, thereby aiding in informed decision-making processes. Hence, the integration of XAI in geotechnical engineering not only augments predictive capabilities but also enhances the understanding of complex soil behaviors, thereby contributing to more reliable and interpretable models for applications such as shear strength prediction in stabilized clay soils.

One notable gap in XAI studies within engineering, especially in geotechnical engineering, is the absence of a robust framework capable of providing a comprehensive approach for prediction, explainability, and trustworthiness of results. Such a framework should emphasize the significance of each input feature in prediction while ensuring the reliability of the employed XAI tools through comparative analysis. Additionally,

there is a lack of integration between global and local XAI analysis techniques. Furthermore, the absence of a framework that leverages various XAI techniques to complement each other is evident. This is critical because no single XAI tool can unveil all hidden patterns in the data and the ML black box model.

To address the aforementioned shortcomings in the literature, the current study proposes an XAI framework that utilizes multiple XAI techniques synergistically. The framework, first, employs various statistical methods to have deeper insight into the data set. Secondly, it compares multiple ML models for predicting soil shear strength. Thirdly, global FIM is explored through several XAI techniques to unveil hidden interactions within the model and the data set. Moreover, the relationship between soil input features and shear strength is scrutinized to uncover underlying dependencies. Feature interaction and correlation analyses are conducted to illustrate the extent of dependency among input soil features and their impact on soil shear strength. Finally, local model behavior is examined for specific soil samples to verify consistency in the importance of features between local and global perspectives. The framework is validated using real data set obtained from stabilization tests done in 1990s [38,39].

2 The proposed explainable artificial intelligence framework

The overall proposed framework is shown in Fig. 1. The proposed XAI framework orchestrates a seamless progression from foundational insights to granular interpretations, ensuring a comprehensive understanding of the intricate relationship between input features and soil shear strength. Initiated by an in-depth exploration of the data set's statistical characteristics (Probability density function (PDF) and Histograms), it lays a foundation for subsequent analyses. This foundational understanding informs the global examination of data (SHAP and FIM), unraveling the paramount influence of input parameters on shear strength predictions. Subsequently, the framework transitions to assessing the interplay between input and output parameters, shedding light on nuanced relationships through techniques like PDP, ICE, and ALE. The understanding gleaned from these steps is then leveraged to explore the intricate web of relationships among input features. Through HM and feature interaction analyses, multivariate patterns crucial to shear strength emerge. Finally, the framework culminates in a local assessment, employing LIME to probe specific data points, providing invaluable insights into local model behavior. This meticulously designed framework not only elucidates the underlying mechanisms driving shear strength but also bestows a

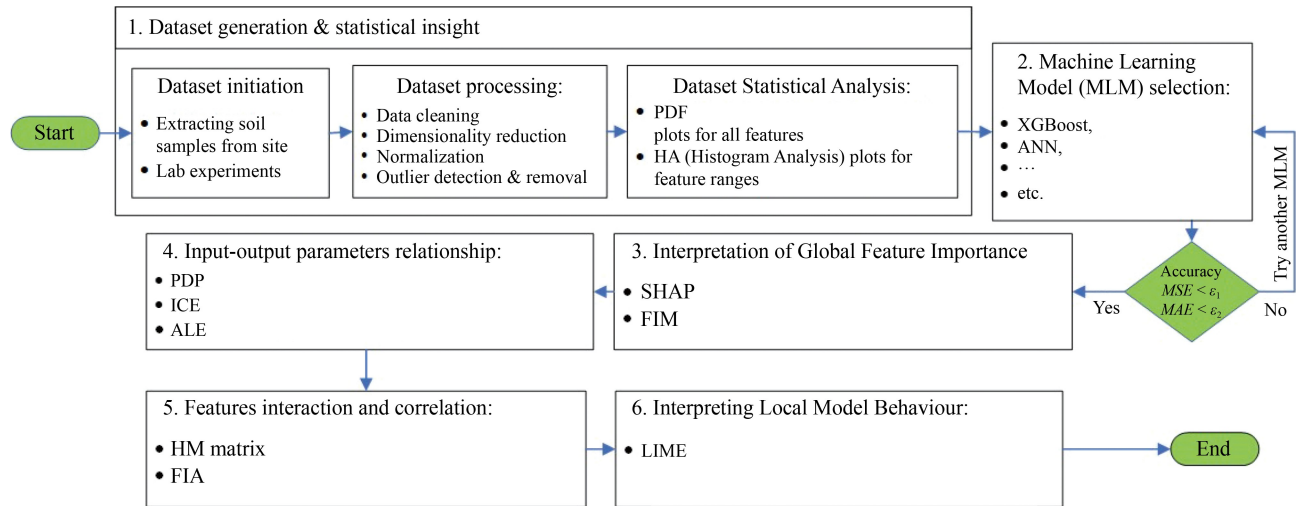


Fig. 1 The overall proposed framework.

powerful tool for interpreting the ML model, thus significantly advancing the field of soil mechanics and engineering structures. The steps of the framework are explained in detail in the following the steps:

2.1 Data set generation and statistical insight

2.1.1 Data set processing

In this stage, collected data is structured to train the ML model, ensuring input features (soil properties) and corresponding output (soil shear strength) are well-organized. Dimensionality reduction initially simplifies complex data sets by decreasing the number of sparse features or variables, making them more manageable for analysis. After that, outlier detection and removal are conducted for identifying and handling data points that deviate significantly from the rest of the data set. Outliers can distort the results of statistical analyses and ML models, making it important to either correct or remove them. After that, all the data set is normalized by adjusting the scale of input features to a uniform range, typically between 0 and 1. This equalizes the contribution of all features during the model's learning process, preventing any single variable from exerting undue influence due to differences in scale. Finally, the data set is partitioned into two distinct subsets: a training set (70%) and a testing set (30%). The training set is utilized to train the model, while the testing set is reserved for assessing its performance. This random splitting ensures that both subsets accurately represent the overall data set and guards against potential issues of overfitting.

2.1.2 Statistical properties of the data set

After data set initiation and processing, statistical insights are gained by constructing PDF curves and HA. PDF

allows to gain deep insights into the distribution of data across various soil properties. By understanding the underlying distribution, we can better grasp the inherent variability in shear strength. In addition, PDF analysis reveals crucial probability characteristics inherent in the data set. These characteristics shed light on the likelihood of different values occurring within the range of each soil property. This information is pivotal in understanding the statistical behavior of the data set. Employing HA in the framework provides a visual representation of the distribution of data for each soil property (input features). This graphical depiction offers an intuitive understanding of how data points are distributed within specific ranges. It serves as a valuable tool for visualizing the variability and discerning patterns and trends in shear strength of the soil. These analyses collectively contribute to a robust understanding of the data set's statistical properties and distribution characteristics, laying a solid foundation for the subsequent stages of our explainable AI framework.

2.2 Machine learning model selection

In this step different MLMs are trained to select the model providing the highest accuracy. Accuracy metrics used are the mean square error (MSE), and the mean absolute error (MAE), which can be represented mathematically as follows [40]:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (2)$$

where n represents the total number of observations, y_i denotes the actual value, and $\hat{y}_i$ represents the predicted

value. These metrics provide a quantitative evaluation of the model's predictive performance, with lower values indicating better accuracy. Common MLMs are tried first such as XGBoost (eXtreme Gradient Boosting), ANN (Artificial Neural Network), etc. After trying several MLMs, the model showing the highest accuracy will be selected. The accuracy is controlled by fulfilling the minimum error requirements (i.e., $MSE < e_1$, $MAE < e_2$), where e_1 , e_2 are thresholds defined by the designer to guarantee high accuracy of the selected MLM model.

2.3 Interpretation of global feature importance

In this step, the investigation of global FIM is enriched by employing both FIM and SHAP techniques in tandem. This intentional combination enhances our understanding of each feature's pivotal role in the analysis, with SHAP providing additional depth to the insights obtained from FIM. Together, these techniques illuminate the significance of individual features in driving the overall soil shear strength outcomes.

2.3.1 Feature importance

FIM provides a ranking of soil features based on their importance, which is related to their direct influence on shear strength prediction. The importance of a feature is measured by calculating the increase in prediction error of the model after accounting for the feature in the analysis. A feature is considered "important" if shuffling its values causes the model error to increase and is considered "unimportant" if shuffling its values causes no change to the model error [41],

$$\begin{aligned} e_{\text{orig}} &= L(y, \hat{f}(X)), \\ e_{\text{perm}} &= L(Y, \hat{f}(X_{\text{perm}})), \\ FI_j &= \frac{e_{\text{perm}}}{e_{\text{orig}}} \text{ or } FI_j = e_{\text{perm}} - e_{\text{orig}}, \end{aligned} \quad (3)$$

where $\hat{f}$ is trained model; X is feature matrix; y is target vector; $L(y, \hat{f})$ is error measure; X_{perm} is feature matrix, and it is generated using each $j \in \{1, \dots, p\}$ feature in the data; e_{orig} is original model error; e_{perm} is estimate error based on the permuted data's predictions; FI_j is permutation FIM. The importance measure considers all interactions with other features automatically. When a feature is allowed, it also eliminates the effects of its interactions with other features. This implies that the importance of permuting a feature takes into account not only its individual impact but also the combined impact of its interactions on the model's performance.

Despite FIM provides a ranking of soil features based

on their importance, but it does not provide information about how each feature affects individual predictions or the interactions between features. This additional layer of insight provides a comprehensive understanding of the critical factors driving soil shear strength variations. This is what SHAP concept will provide as explained in the next section.

2.3.2 Shapley additive explanations

SHAP is a recently popular game theory concept that aims to explain the prediction of an instance x with computing each feature's contribution to the prediction. The SHAP explanation technique uses coalitional game theory to calculate Shapley values. The three desirable features described by SHAP are: 1) local accuracy: the output of the explanation model must at least match that of the initial model; 2) missingness: features that were not present in the initial input must have no effect; 3) consistency: if we change a model so that it is more dependent on a specific feature, the significance of that feature should not be diminished, regardless of the value of other features. Mathematically, the explanation model $g(z')$ with simplified input z' for initial model $\hat{f}(x)$ is defined as [41],

$$\hat{f}(x) = g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j, \quad (4)$$

where $x = h_x(z')$ is the original method's mapping function; the model output is $\phi_0 = f(h_x(0))$, which excludes all simplified inputs; M is the number of the input characteristics and $\phi_j \in \mathbb{R}$ is the characteristic attribution for one characteristic j (i.e., the SHAP values). The Shapley value ($\phi_j(\text{val})$) is defined as:

$$\phi_j(\text{val}) = \sum_{S \subseteq \{1, \dots, p\} | (j)} \frac{|S|!(p-|S|-1)!}{p!} \cdot (\text{val}(S \cup \{j\}) - \text{val}(S)), \quad (5)$$

where S is the number of entries that are not zero; p is the number of features (i.e., the main soil characteristics); $\text{val}(S)$ is the predicted value of feature values in set S that are marginalized over features not in set S .

based on that, the Shapley value will be provided for each input feature of the clay soil data set. Interpretation of global FIM is shown graphically by arranging the soil characteristics based on their respective Shapley values. This graphical representation provides a clear visualization of the relative importance of each input feature within the clay soil data set. The feature with the highest Shapley value is considered the most influential in determining the shear strength of this particular soil. Furthermore, the SHAP figure not only indicates the magnitude of influence each soil characteristic has on

shear strength, but also specifies whether an increase or decrease in the feature corresponds to an increase in shear strength.

2.4 Input-output parameters relationship

In this step, we employ three different techniques: PDP, ICE, and ALE. Each technique complements and reinforces the insights gained from the others, providing a comprehensive understanding of the interplay between parameters and shear strength. PDP allows us to meticulously analyze the individual effects of parameters on shear strength. By uncovering trends and relationships, PDP offers a valuable view of how each parameter contributes to the overall outcome. ICE further enriches our analysis by examining the influence of individual parameters in isolation. This technique excels at capturing the variability in predictions, offering a detailed perspective on how each parameter impacts shear strength across different scenarios. ALE takes our understanding to a deeper level by revealing the cumulative effects of parameters. By identifying nonlinear relationships, ALE provides critical insights into how combinations of parameters collectively influence shear strength. A brief description and the mathematical representation of each technique is explained below.

2.4.1 Partial dependence plots

PDP visualizes the relationship between a feature and the model's predictions while marginalizing the effects of other variables. It shows how the model's output changes as the selected feature's value varies, allowing analysts to understand the feature's impact on the predictions. PDP focuses on a set of features denoted as S and treat the remaining features as random variables (C). The function $\hat{f}_s(x_s)$ shows the impact of the features in set S on the prediction. To understand this, the mathematical expression defining the partial dependence function for regression models is given as [34,42]:

$$\hat{f}_s(x_s) = E_{X_c}[\hat{f}(x_s, X_c)] = \int \hat{f}(x_s, X_c) d\mathbb{P}(X_c), \quad (6)$$

where $\hat{f}_s(x_s)$ represents the partial dependence function we want to plot, x_s denotes the specific values of the features in set SS for which we seek the effect on predictions, X_c represents the remaining features treated as random variables, $\hat{f}(x_s, X_c)$ is the ML model's output considering both x_s and X_c , $E_{X_c}[\hat{f}(x_s, X_c)]$ calculates the expectation over the distribution of X_c , effectively averaging out the impact of the random variables, and $d\mathbb{P}(X_c)$ represents the probability measure over X_c .

2.4.2 Individual conditional expectation

ICE plots are used for visualizing supervised learning

models. Unlike PDP, which show average effects globally, ICE plots provide a granular view, illustrating the impact of a feature on predictions for each instance separately. The plots provide insights into the fitted values across the covariate range and offer a visual test for additive structure in the model, surpassing the limitations of PDPs [34,43]. Mathematically, for a given data set $\{(x_s^{(i)}, x_c^{(i)})\}_{i=1}^N$, where $x_s^{(i)}$ represents the feature of interest and $x_c^{(i)}$ represents the remaining features, the ICE plot defined by the black box model's prediction function $\hat{f}$ for the i th instance is represented as:

$$\hat{f}_s^{(i)} = \hat{f}(x_s^{(i)}, x_c^{(i)}). \quad (7)$$

2.4.3 Accumulated local effects

ALE focus on the impact of a particular feature without partially considering the impact of other features on a ML model prediction. If there is a powerful relationship among the variables in PDP, the results are unreliable, in this case ALE is used. It means that ALE overcomes some key shortcomings of PDP. ALE allows isolating associated features by using "averages of the differences in predictions" instead of the "averages" used in PDP. ALE is defined for (x_s) as [41]:

$$\begin{aligned} \hat{f}_{S,ALE}(x_s) &= \int_{z_{0,S}}^{x_s} E_{X_c|X_s=x_s}[\hat{f}(X_s, X_c) X_s = z_s] dz_s - \text{constant} \\ &= \int_{z_{0,S}}^{x_s} \left(\int_{x_c} \hat{f}(z_s, X_c) d\mathbb{P}(X_c|X_s = z_s) \right) dz_s - \text{constant} \\ \cdot \hat{f}_s(x_s, x_c) &= \frac{\partial \hat{f}(x_s, x_c)}{\partial x_s}. \end{aligned} \quad (8)$$

The complement set of S is C . The characteristics for which the ALE function should be drawn are x_s , and the other characteristics utilized in the ML model $\hat{f}$ are x_c , which are represented as random variables. The local effects of x_s on $\hat{f}(\cdot)$ at (x_1, x_s) are represented by $\hat{f}_s(x_s, x_c)$, $z_{0,S}$ is a value chosen slightly below the lowest observation, and the conditional density is defined as $\mathbb{P}(x_s, x_c)$. For the actual computation, the z 's are substituted using a grid of intervals over which we calculate the changes in the prediction. The ALE technique estimates the effect by calculating the prediction differences conditional on features S and integrating the derivative over features S rather than directly averaging the predictions. ALE plots are zero-centered, and this makes data visualization simple and works well with correlated data.

2.5 Features interaction and correlation

In this section, the deliberate combination of two

techniques, HM and FIA, allows for a comprehensive exploration of feature interactions and correlations. The HM visually presents correlations among input features of the soil, offering insights into multivariate patterns. Concurrently, the FIA method uncovers synergies and conflicts between features, shedding light on their collective impact on shear strength prediction. Together, these techniques provide a holistic understanding of the intricate relationships within the data set, offering valuable implications for accurate shear strength predictions.

HM relies on color gradients or intensity variations to represent the magnitude or density of data points within a matrix or grid. The values in the data set are typically normalized and assigned colors based on their relative magnitude. Feature interaction refers to the relationships and dependencies between different features or variables in a ML model. It explores how the interaction of certain features can affect the model's predictions by measuring the interaction strength, e.g., H-statistic [44]. H-statistic quantifies the strength of interactions, e.g., for two cases, i.e., two-way interaction and total interaction. The interaction strength statistic, H_{jk}^2 , between clay soil features j and k for two-way interaction is given by:

$$H_{jk}^2 = \frac{\sum_{i=1}^n [PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)})]^2}{\sum_{i=1}^n PD_{jk}^2(x_j^{(i)}, x_k^{(i)})}. \quad (9)$$

For total interaction with feature j and all others:

$$H_j^2 = \frac{\sum_{i=1}^n [\hat{f}(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)})]^2}{\sum_{i=1}^n \hat{f}^2(x^{(i)})}. \quad (10)$$

Equation (9) numerator: $PD_{jk}(x_j^{(i)}, x_k^{(i)})$ is the two-way partial dependence function of features j and k evaluated at the i th data point. This represents the average effect of features j and k on the model's prediction, holding all other features constant. $PD_j(x_j^{(i)})$ is the partial dependence function of feature j evaluated at the i th data point. This represents the average effect of feature j on the model's prediction, averaging over all other features. $PD_k(x_k^{(i)})$ is the partial dependence function of feature k evaluated at the i th data point. The difference between these terms is squared to emphasize the strength of the interaction. The summation over i calculates the total interaction strength across all data points.

Equation (9) denominator: $PD_{jk}^2(x_j^{(i)}, x_k^{(i)})$ is the squared two-way partial dependence function. The summation

over i calculates the total variance explained by the interaction between j and k .

A higher value of H indicates a stronger interaction between features j and k . An H -statistic of 0 indicates no interaction, while 1 implies all variance comes from interactions between these two features.

Equation (10) numerator: $\hat{f}(x_i)$ represents the model's prediction for the i th data point. $PD_{-j}(x_{-j}^{(i)})$ is the average marginal effect of all features except j evaluated at the i th data point. The difference between these terms in the numerator is squared to emphasize the strength of the interaction. The summation over i calculates the total interaction strength between feature j and all other features.

Equation (10) denominator: $\hat{f}^2(x^{(i)})$ is the squared model prediction. The summation over i calculates the total variance explained by the model.

A higher value of H indicates a stronger interaction between feature j and all other features.

An H -statistic of 0 indicates no interaction, while 1 implies all variance in the model's predictions is due to interactions.

2.6 Interpreting local model behavior

In this step a more specific investigation is made on a particular soil sample to highlight which of the input features (for this specific sample) contributed to the predicted shear strength. Unlike global explanations, which provide insights into the model's behavior across the entire data set, LIME focuses on a single data point. This becomes especially valuable when we seek to comprehend the rationale behind a particular shear strength prediction, a critical aspect in practical applications. It's important to note that the features highlighted as significant in a local examination (such as LIME) may not necessarily align with those identified in a global analysis (like SHAP). LIME trains local interpretable models to ensure individual predictions in a number of stages. This procedure involves perturbing the inputs, followed by observing the general (black box) model's output in order to understand how the predictions alter with various observations [45].

$$\text{explanation}(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g), \quad (11)$$

where f is the model to be explained; g explanation model; π_x (proximity measure) defines how large the neighborhood around the x instance is; L loss function; $\Omega(g)$ penalizes complexity of the explanation model; G is the family of potential explanations. The purpose is to minimize the loss function L , which assesses how close the explanation matches the prediction of the original model f given a proximity measure π_x .

3 Application of the framework on real clay soil shear strength prediction

In this section, the framework is validated by examining a real clay soil samples.

3.1 Data set initiation and processing

1) Data set initiation

The laboratory test results (based on the stabilization tests done in 1990s [38,39] consist of uniaxial compression test results for cylindrical samples (height 100 mm, diameter 50 mm), from which the shear strength is calculated. These samples are prepared by mixing soft soil with binder and compacting the mixture into tubes. Subsequently, the samples undergo curing for various time spans.

2) Data set processing

The gathered data is structured in a manner suitable for training the ML model. This ensures that both the input features (soil properties) and their corresponding output (soil shear strength) are appropriately organized. Dimensionality reduction is performed on the input feature space to decrease the number of features or variables in the data set that exhibit sparsity across their input domain. Based on that the main input features used are water content (%), humus (%), clay content (%), liquidity limit (%), binder (type), amount of binder (% of wet mass of clay), amount of binder (% of dry mass of clay), water/binder ratio (%), curing time (d). The output response quantity is the stabilized clay shear strength (kN/m²). After that, data cleaning and processing is conducted; this includes: 1) outlier detection and removal: data points that significantly deviated from the majority were identified and removed to prevent them from unduly influencing the model's training; 2) handling missing values: missing values were imputed using appropriate techniques, such as mean imputation or interpolation, depending on the nature of the missing data; 3) feature selection and engineering: redundant features were identified and removed, and consistent units were established for similar features. Features with insufficient data points were excluded to ensure the reliability of the model. The final data set, comprising 817 records and 9 independent features, was carefully curated to meet the requirements of the ML models. The count of binders used in the data set is shown in Fig. 2.

Figure 3 shows the histograms of the input features. The soil property histograms reveal interesting trends. Water content and liquidity limit are skewed toward drier conditions, while curing times primarily focus on shorter durations. Humus content exhibits two distinct populations, potentially indicating two different types. Clay content shows a uniform distribution, while both wet and dry binder amounts favor lower percentages.

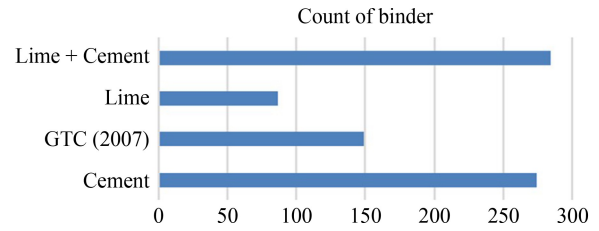


Fig. 2 Count of the binders in the data set.

Finally, the water-to-binder ratio concentrates on lower values, suggesting efficient binder usage at lower water content.

Figure 4 shows PDF plots of the input features. The soil characteristics exhibit diverse distributions: water content skewed with a majority of data points showing low levels; humus content bimodal, with peaks at lower and higher percentages; and clay content evenly distributed. Liquidity limit presents higher probability at lower values pattern, while the amount of binder is left-skewed for both wet and dry percentages. The water/binder ratio shows a prominent peak at lower ratios. Additionally, curing time displays a highly skewed distribution, suggesting shorter curing times.

3.2 Machine learning predictive model selection

Upon investigating multiple ML models, it was concluded that XGBoost and ANN models [46–48] demonstrated the highest accuracy levels (refer to Table 1). Consequently, XGBoost was selected as the preferred model for the proposed framework.

XGBoost is a powerful machine-learning algorithm particularly suited for regression tasks [49]. XGBoost is an ensemble learning algorithm that combines the predictions of multiple decision trees. It works by sequentially training an ensemble of weak learners (typically decision trees) and combining their predictions. The objective function to be minimized by XGBoost is as follows:

$$\mathcal{L}(\theta) = \sum_{i=1}^n l(\mathcal{Y}_i, \hat{\mathcal{Y}}_i) + \sum_{i=1}^k \Omega(f_i), \quad (12)$$

where n is the number of training examples, k is the number of trees in the ensemble, $\mathcal{Y}_i$ is the true output for the i th example, $\hat{\mathcal{Y}}_i$ is the predicted output for the i th example, l is the loss function that measures the difference between true and predicted values, $\Omega(f_i)$ is the regularization term that penalizes the complexity of each tree. The loss function, which measures the squared difference between the true and predicted strength values is given as follows:

$$l(\mathcal{Y}_i, \hat{\mathcal{Y}}_i) = (\mathcal{Y}_i - \hat{\mathcal{Y}}_i)^2. \quad (13)$$

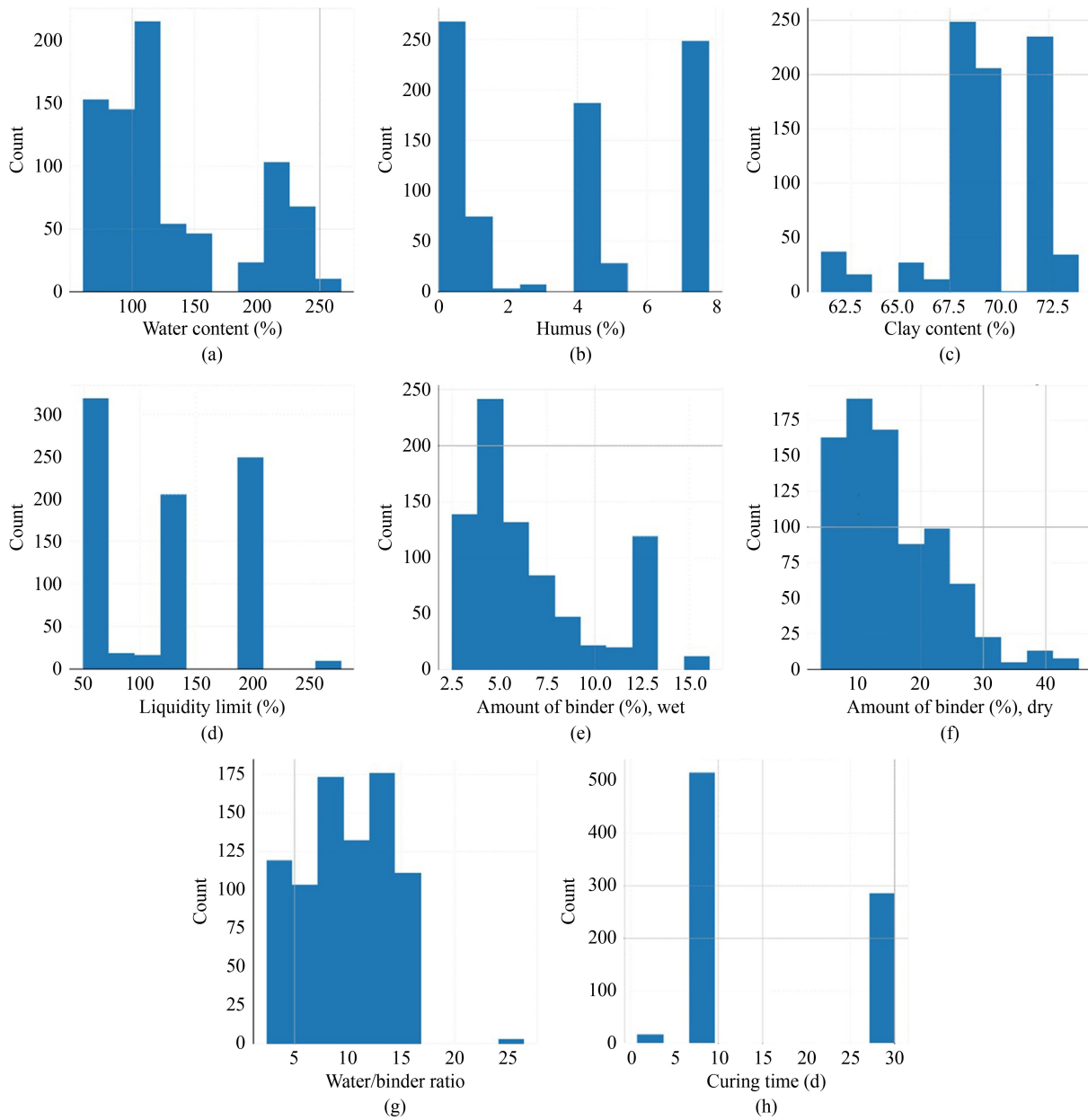


Fig. 3 Histograms of the input features: (a) water content; (b) humus; (c) clay content; (d) liquidity limit; (e) amount of binder, wet; (f) amount of binder, dry; (g) water/binder ratio; (h) curing time.

The regularization term discourages overfitting by penalizing complex trees and is given by:

$$\Omega(f_i) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2, \quad (14)$$

where T is the number of leaves in the tree, w_j are the weights associated with the leaves, γ and λ are hyperparameters controlling the strength of the regularization. By optimizing the objective function using techniques like gradient boosting, XGBoost efficiently constructs an ensemble of trees that collectively provide accurate predictions for the strength of based on the main input parameters. In addition, the regularization terms

ensure that the model remains robust and avoids overfitting.

4 Interpretability analysis and model insights

4.1 Interpretation of global feature importance

In this section, steps 3–6 of the framework are explained graphically. First, the interpretation of the global FIM is provided by SHAP and FIM. Figure 5 shows the SHAP plot and Fig. 6 shows the FIM of the input features.

The SHAP plot provides a comprehensive overview of

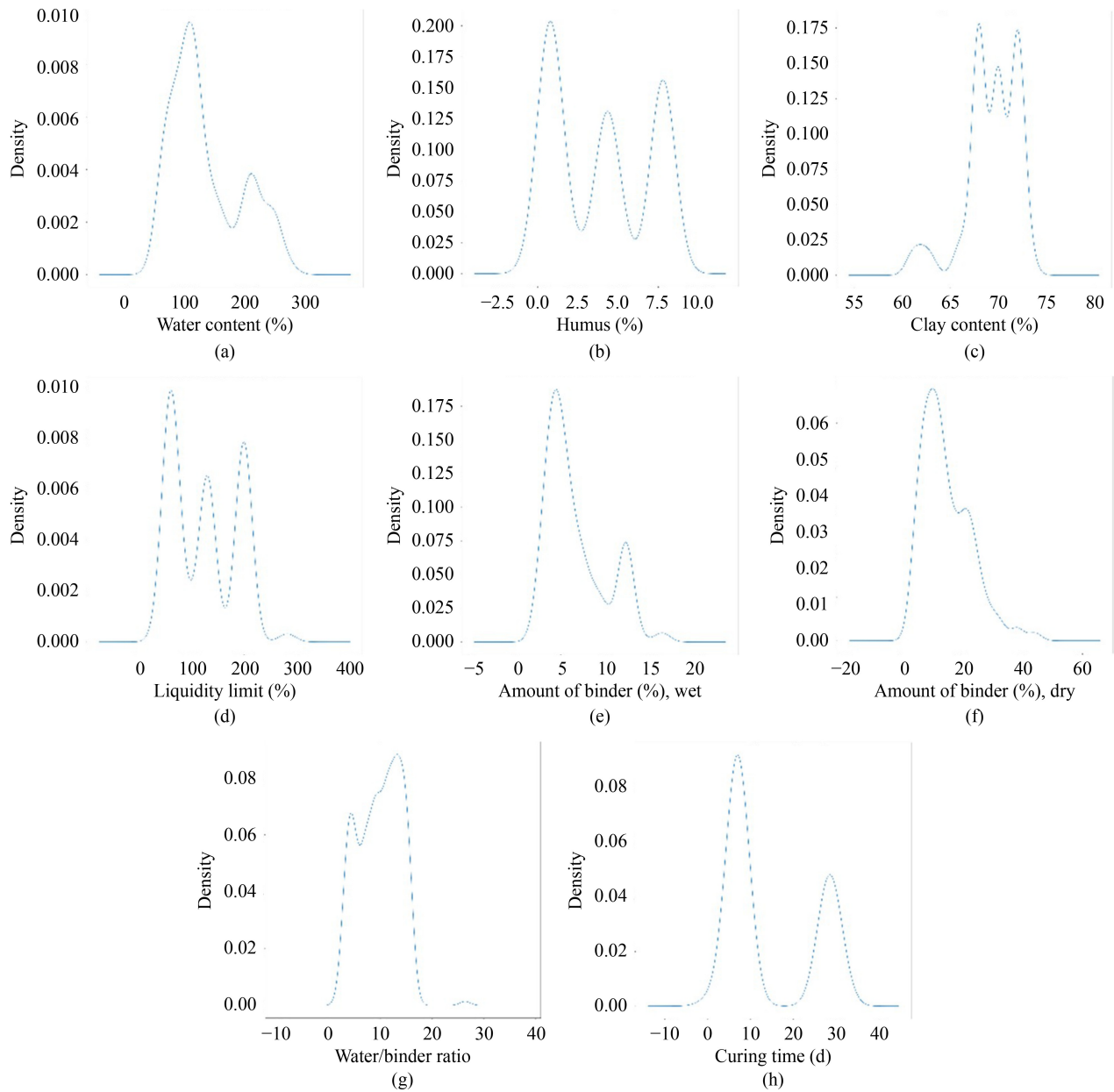


Fig. 4 PDF plots of the input features: (a) water content; (b) humus; (c) clay content; (d) liquidity limit; (e) amount of binder, wet; (f) amount of binder, dry; (g) water/binder ratio; (h) curing time.

Table 1 Accuracy of the ML models

Accuracy metrics	XGBoost			ANN		
	Test	Training	All	Test	Training	All
R^2	90.27%	94.44%	93.56%	90.12%	91.67%	91.37%
$RMSE$	8.55%	6.11%	6.67%	8.62%	7.47%	7.72%
MAE	5.62%	4.36%	4.61%	5.79%	5.13%	5.26%

the importance of different input features in determining the shear strength of the soil. In this plot, features with positive SHAP values (represented by blue dots) indicate a positive relationship with shear strength. Conversely, features with negative SHAP values (represented by red dots) have a negative relationship with shear strength. The impact of higher humus content on the model's

prediction is consistently notable, with its presence consistently pushing the shear strength upwards, whereas lower humus content demonstrates comparatively less influence on the resulting output. However, the effect of water content exhibits a more nuanced pattern, as both high and low values can lead to either positive or negative impacts on the model's prediction, indicating a notable sensitivity to variations in water content with a mixed pattern evident predominantly on the left side of the feature spectrum. Conversely, the water/binder ratio exerts a predominantly negative influence on the model output, particularly evident in the concentration of a prominent red area on the left side of the zero line, suggesting that lower water/binder ratios tend to yield

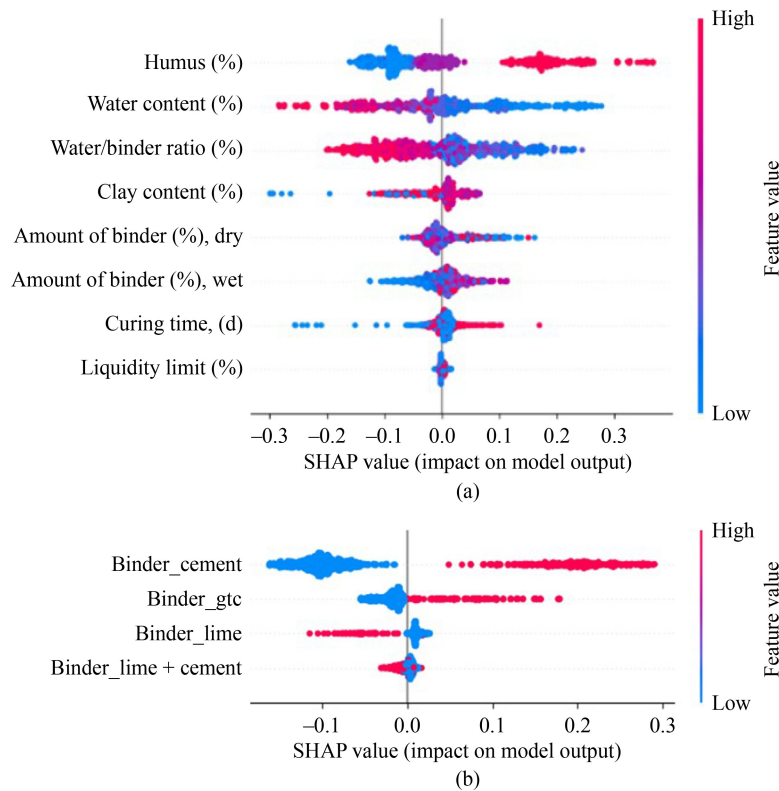


Fig. 5 SHAP plots for (a) all input features without binders; (b) binders.

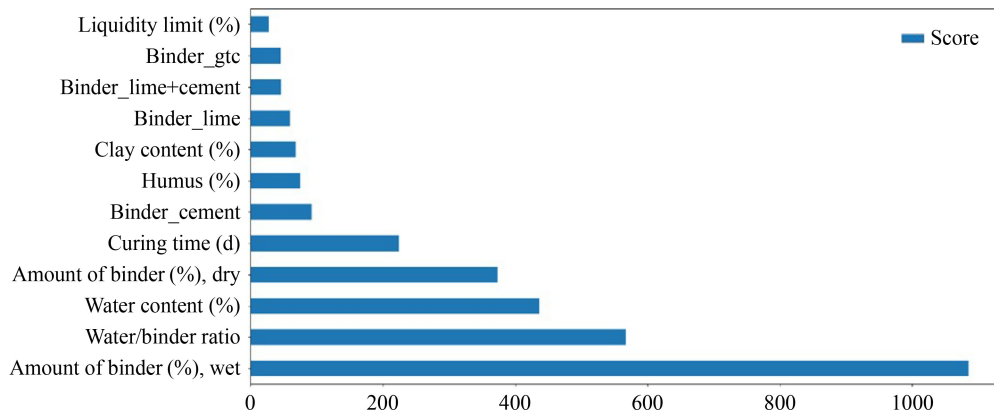


Fig. 6 FIM plot for all input features.

higher shear strength outputs, potentially reflecting more efficient binder usage with reduced water content. Moreover, higher clay content tends to positively affect shear strength values, whereas lower clay content demonstrates scattered influence across the data set. The impact of the amount of binder, whether dry or wet, exhibits variability, with no clear trend indicating predominant influence of high or low values on the resulting outcome, although lower values of the amount of binder ((%), wet) appear to correlate with increased shear strength values, as indicated by scattered blue dots on the negative side of the outcome. Additionally, curing time weakly impacts the model output in a positive direction, with a subtle blue tinge observed on the right

side, although scatter on the negative side remains evident. While the precise impact of the liquidity limit remains somewhat ambiguous due to the mixture of blue and red tones observed across the feature range, it appears to have a generally limited influence on the resulting shear strength value overall.

Figure 5(b) shows SHAP plots for the different types of binders. The SHAP value analysis reveals that Binder–Cement has a positive impact on the model's output, indicating that higher percentages of cement in the binder tend to increase the predicted shear strength values. Conversely, lower percentages of cement have a comparatively lesser effect on the model's prediction. Similarly, Binder–GTC demonstrates a positive effect,

with higher scatter observed in the lower values and less scatter in the higher values. In contrast, higher lime content tends to negatively influence model predictions, while lower lime content exhibits a less consistent impact. Moreover, the combination of lime and cement presents a diverse impact, with some samples benefiting from this combination while others do not, leading to variations in the model's output based on the specific mix of lime and cement.

Figure 6 shows the FIM plot, which provides a ranked list of features based on their influence on shear strength prediction. The features are ordered from the most influential (at the bottom) to the least influential (at the top). According to the FIM plot, amount of binder ((%), wet) emerges as the most influential feature, signifying its significant contribution to shear strength prediction. Following are features like water/binder ratio, water content, amount of binder ((%), dry), and Curing time in days, which also hold substantial importance. Interestingly, the type of binder (specifically cement) and the amount of binder (as a percentage of dry clay mass) hold less significant influence. In contrast, Binder–Cement, Humus, and clay content fall in the middle of the plot. While they still contribute significantly to shear strength, their influence is considered to be of moderate importance in comparison to the other features. The FIM plot highlights that features such as liquidity limit and various binder types demonstrate relatively lower importance in influencing the shear strength.

4.2 Input–output parameters relationship

Figures 7–9 show input–output parameters relationship through PDP, ICE, and ALE, respectively. As evident from the PDPs, the relationship between input features and shear strength exhibits distinct patterns. Water content and water/binder ratio demonstrate a decrease in partial dependence with increasing values. Conversely, humus and liquidity limit exhibit an opposite effect, with their partial dependence increasing as values rise. Notably, clay content displays a stepped pattern with steep slopes, indicating a significant impact on shear strength within the range of 65.5% to 68%. Similar behavior is observed for liquidity limit between 135% and 200%. The patterns for features like amount of binder (both wet and dry) and water/binder ratio are notably variable. Additionally, curing time exhibits steep slopes for values below 10 d, indicating a pronounced influence on shear strength within this timeframe.

ICE plots allow for the exploration of sample-specific trends, showcasing how individual observations respond to changes in input features. As shown in Fig. 8, ICE plots for almost all features align closely with the aggregate trend represented by the PDPs. This consistency reinforces the robustness and reliability of the

insights derived from the PDPs. This suggests that the average response to changes in each feature is a reliable representation of the overall trend. Also, this indicates that the average response is a representative trend that can be generalized to broader scenarios for each feature. Moreover, ICE plots reveals that there are no significant outlier scenarios in the data set. The consistent alignment between ICE and PDP for all features reaffirms the robustness of the insights gained. It indicates that the averaged response provided by PDP reliably captures the essence of the relationship between each input feature and shear strength.

ALE plots, as shown in Fig. 9, follow the same trends of PDPs and ICE plots and provide not only the central trend but also include 95% confidence intervals. This adds an important dimension of uncertainty, allowing for a more understanding and reliability of the relationship between input features and shear strength. ALE plots include indicators of sample density along the horizontal axis. This visual aspect helps to understand regions of the input feature space where data points are densely or sparsely distributed. This information can be crucial in identifying areas where predictions may be more or less reliable. The inclusion of confidence intervals aids in understanding the range of uncertainty associated with predictions. This information is vital in decision-making processes, as it allows for a more informed assessment of the reliability of model predictions. For example, in the ALE plot of clay content it is clear that between 63% and 66% there is a wider confidence interval compared to other percentages. a wider confidence interval indicates more uncertainty and lower precision in the predictions. In such cases, the model's predictions may have a greater range of potential shear strength values, making them less accurate. Conversely, A narrower confidence interval indicates higher precision and less uncertainty in the predictions. Therefore, when the confidence interval is narrow, it suggests that the model's predictions are more accurate and reliable. This means that the model is likely to make shear strength predictions that are closer to the true values.

It is noteworthy to mention that ALE plots focus on capturing the marginal effect of a single feature while taking into account the interactions with other features (not like PDPs, which do not consider the interactions between different features). Also, ALE plots provide a more localized view of the relationship and do not assume linearity (as PDPs) and allow for nonlinear relationships between the feature and the output. This provides a more accurate representation of the feature's effect especially if there is interaction between features and a nonlinear relationship between the input features and the shear strength. It is important to note that binder types cannot be represented through PDP, ICE, and ALE plots because of their categorical nature. Unlike

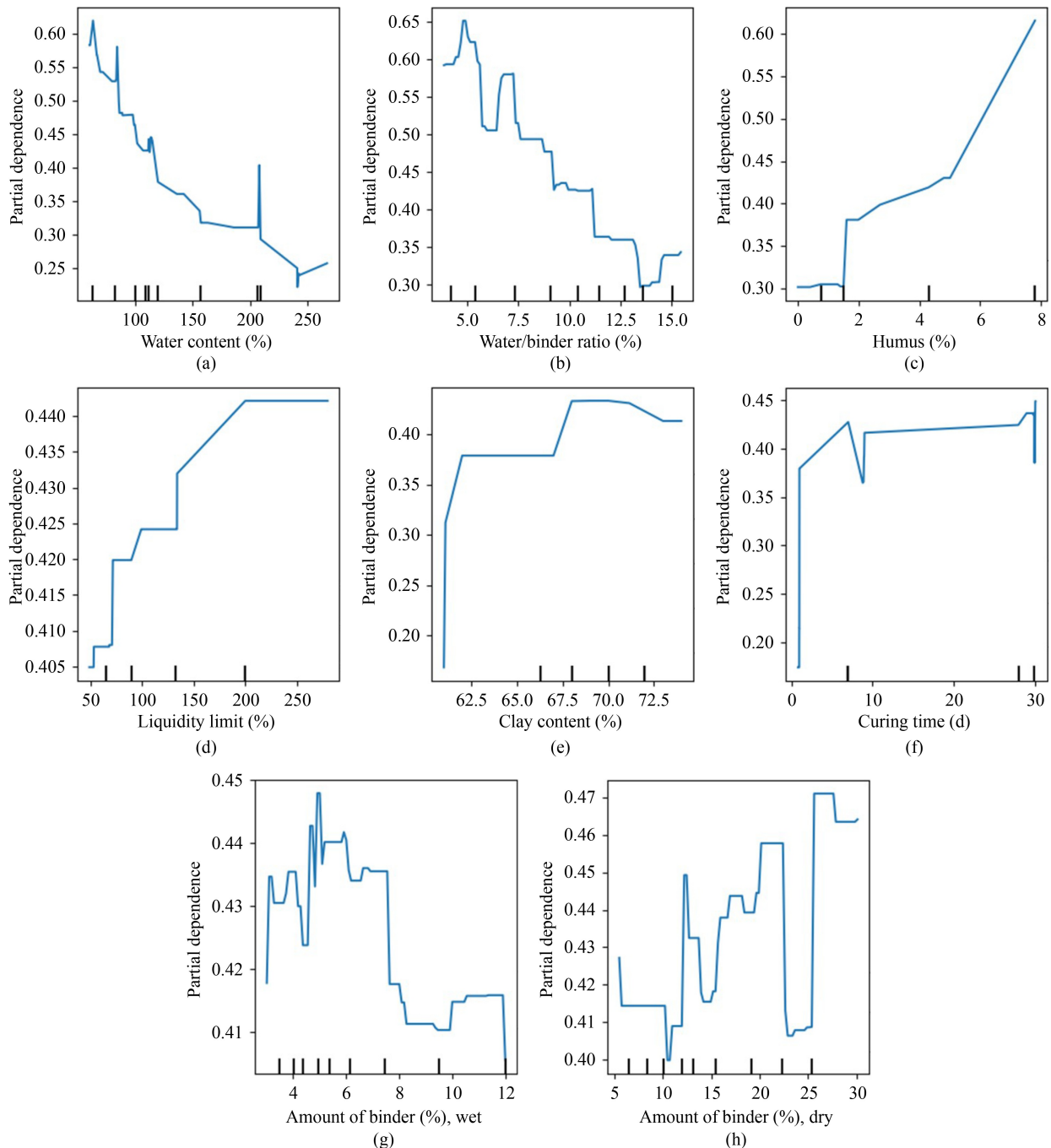


Fig. 7 PDPs plots: (a) water content; (b) water/binder ratio; (c) humus; (d) liquidity limit; (e) clay content; (f) curing time; (g) amount of binder, wet; (h) amount of binder, dry.

numerical or ordinal features, categorical features lack inherent numerical values or ordering, making it challenging to visualize their effects using these plot techniques. As a result, alternative methods are required to explore the influence of categorical features like binder types on the predicted outcome.

4.3 Features interaction and correlation

The input feature interaction and correlation are

represented by the HM, [Fig. 10](#), and the FIA, [Table 1](#). The observed correlations in the HM highlight the interplay between key soil parameters. Notably, higher clay content tends to coincide with lower percentages of binder, both in wet and dry states, as well as a reduced liquidity limit. This suggests a relationship between these characteristics, potentially influencing the overall shear strength of the soil. Also, a negative correlation exists between water/binder ratio and amount of binder (wet and dry). Conversely, a noteworthy positive correlation

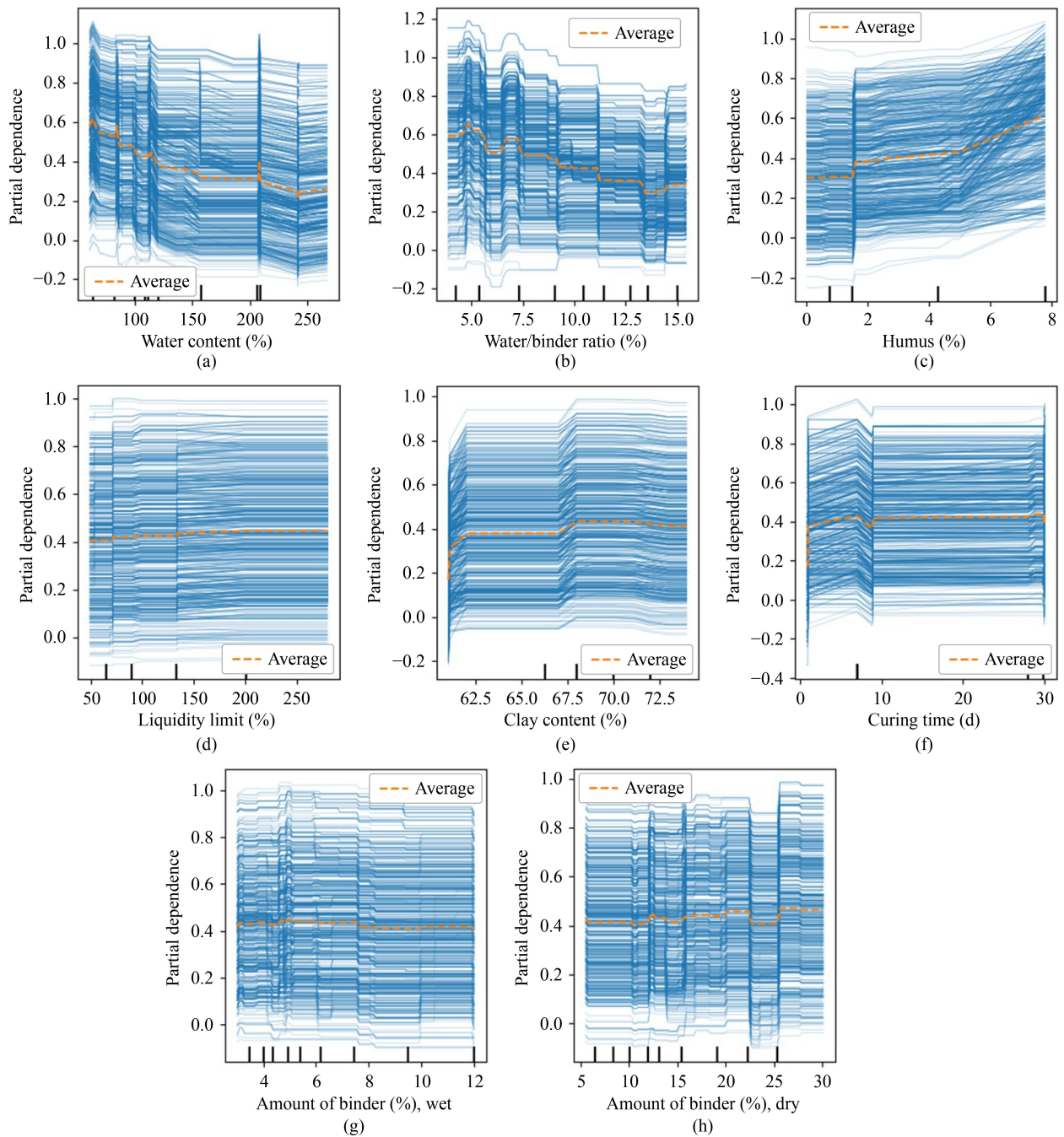


Fig. 8 ICE plots: (a) water content; (b) water/binder ratio; (c) humus; (d) liquidity limit; (e) clay content; (f) curing time; (g) amount of binder, wet; (h) amount of binder, dry.

emerges between humus content and water content, and also, water content and liquidity limit. This indicates a potential avenue for further investigation into their combined effects on soil behavior. The absence of strong correlations for other parameters indicates less influence on shear strength.

The ranking of features using the FIA, Table 2, offers valuable insights into their relative importance in influencing shear strength. In the provided table, “gain” refers to the measure of importance assigned to each input feature in predicting clay shear strength.

Essentially, it indicates how much each factor contributes to the accuracy of the predictive model. Higher gain values signify greater importance, suggesting that those features have a more significant impact on the outcome being predicted. The “FScore” quantifies a feature’s predictive importance, with higher values indicating greater relevance to the outcome. “wFScore” adjusts for data set imbalances, highlighting features with more substantial predictive contributions. The “Average wFScore” provides a collective assessment of a feature’s predictive power across all data points. “Average Gain”

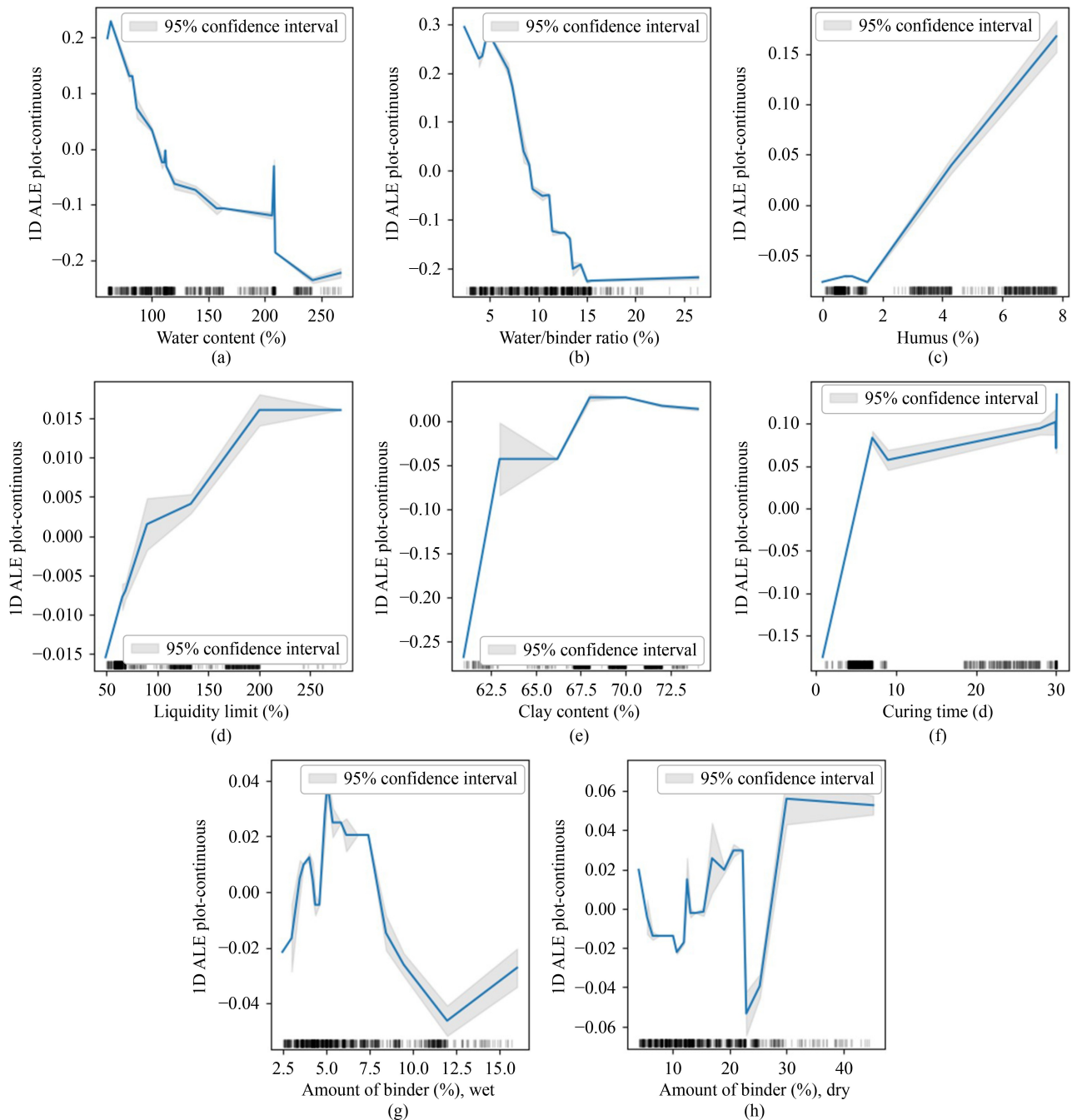


Fig. 9 ALE plots: (a) water content; (b) water/binder ratio; (c) humus; (d) liquidity limit; (e) clay content; (f) curing time; (g) amount of binder, wet; (h) amount of binder, dry.

offers a consolidated measure of importance, reflecting the average gain value for each feature. “Expected Gain” estimates potential model improvements by selecting a feature for node splitting, with higher values indicating greater enhancement potential. Lastly, “Gain Rank” ranks features based on their gain values, facilitating a relative comparison of their importance in predicting the outcome. The feature’s ranking takes into account all of those parameters.

The water/binder ratio emerges as the most influential parameter, with a gain of 19.9, reflecting its critical role

in governing hydration processes and mechanical properties relevant to shear strength. Following closely, the water content (%), with a gain of 15.0, underscores its substantial impact on mixture workability and strength development, thus indirectly affecting shear strength. Additionally, the type and proportion of cement binder, represented by Binder–Cement with a gain of 12.6, significantly influence the observed interaction, highlighting the importance of binder selection in engineering applications and its subsequent impact on shear strength. Notably, the presence of humus, with a

gain of 8.7, and the amount of binder ((%), dry) and ((%), wet), with gains of 7.2 and 6.3 respectively, further contribute to the variability in clay shear strength. Moreover, clay content (%), with a gain of 5.3, can directly influence the shear strength properties of the clay material, while curing time (d), with a gain of 4.3, affects the development of clay strength over time. Specific binder formulations, such as Binder-GTC (2007) with a gain of 3.6, also play a moderate role in determining clay shear strength. However, binders like “Lime” and “Lime + Cement” exhibit relatively minor influences, with gains of 0.9 and 0.2, respectively, alongside the liquidity limit

(%) with a gain of 0.2. This comprehensive analysis provides nuanced insights into the intricate relationships between input parameters and clay shear strength.

4.4 Interpreting local model behavior

Table 3 provides insights into the LIME for three randomly selected samples. Each column in the table corresponds to a distinct sample, while the rows denote the ranking of the most influential features affecting the outcome for that particular sample. Notably, the rankings of important features vary locally across the different samples, indicating a dynamic relationship between input variables and the predicted outcome. For instance, in sample 1, “Binder Lime” holds the highest importance, suggesting its significant impact on the shear strength within that specific context. Conversely, in sample 2, “Water/binder ratio” takes precedence, highlighting its dominant influence on the predicted shear strength for that sample. Similarly, in sample 3, “Amount of binder (% wet)” emerges as the most critical feature. This variability in FIM underscores the complex and context-dependent nature of the relationship between input variables and the predicted outcome. Such variability can arise due to differences in sample characteristics, environmental conditions, or underlying mechanisms driving the outcome of interest. Therefore, considering local interpretations of model outputs is crucial for gaining a comprehensive understanding of the factors

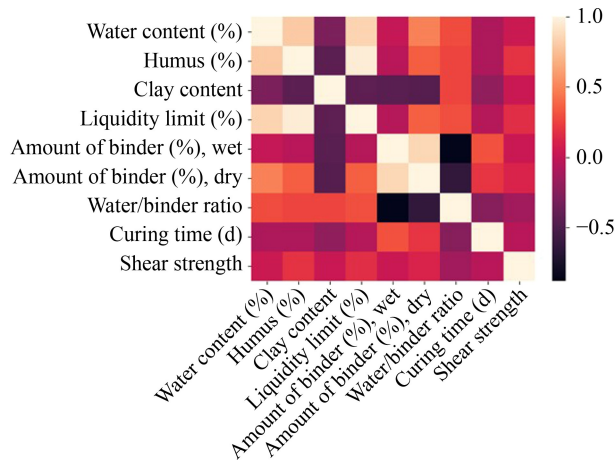


Fig. 10 HM of input features interaction.

Table 2 Feature interaction analysis (FIA)

Interaction	Gain	FScore	wFScore	Average wFScore	Average gain	Expected gain	Gain rank
Water/binder ratio	19.95	567.00	156.52	0.28	0.04	5.91	1
Water content (%)	14.98	436.00	59.99	0.14	0.03	3.04	2
Binder_Cement	12.55	92.00	17.13	0.19	0.14	12.29	3
Humus (%)	8.70	75.00	15.76	0.21	0.12	2.54	4
Amount of binder ((%), dry)	7.20	373.00	115.74	0.31	0.02	3.83	5
Amount of binder ((%), wet)	6.26	1085.00	159.63	0.15	0.01	0.60	6
Clay content (%)	5.31	68.00	11.05	0.16	0.08	1.33	7
Curing time (d)	4.30	224.00	30.38	0.14	0.02	0.56	8
Binder_GTC (2007)	3.64	45.00	5.42	0.12	0.08	0.77	9
Binder_Lime	0.89	59.00	6.22	0.11	0.02	0.12	10
Liquidity limit (%)	0.15	28.00	7.20	0.26	0.01	0.01	11
Binder_Lime + Cement	0.15	46.00	9.21	0.20	0.00	0.02	12

Table 3 LIME outputs for random samples

Ranking of most important features	1	2	3
1	binder lime	water/binder ratio	water/binder ratio
2	curing time	binder_GTC (2007)	amount of binder ((%), wet)
3	water content	humus	amount of binder ((%), dry)
4	water/binder ratio	water content	liquidity limit

influencing the predicted outcomes and informing decision-making processes accordingly. It is worth noting that water/binder ratio appears consistently across all three samples, suggesting its universal importance in influencing the predicted outcomes, which may indicate its fundamental role in the predicted shear strength. Conversely, water content appears in two out of the three samples, indicating its significant but potentially context-specific influence on the predicted shear strength, highlighting its relevance in certain scenarios or conditions within the studied context.

5 Results and discussions

In this section, insights on input features effect on shear strength, interpretability tools comparison, and overall assessment and limitation of the framework are discussed.

5.1 Insights on input features effect on shear strength

5.1.1 Distinguishing key and marginal features in shear strength prediction

The SHAP plot and FIM plot collectively offer a comprehensive understanding of the crucial factors impacting shear strength prediction. Both methods converge on certain key features. Water content and water/binder ratio emerge as the most influential, showcasing their pivotal roles in shaping shear strength. These findings are consistent across both analyses, reinforcing their significance. On the other hand, both the

SHAP plot and the FIM plot suggest that liquidity limit and Binder-Lime + Cement have relatively weak effects on shear strength prediction, demonstrating limited influence compared to other features in both analyses. Among other features, clay content falls within the moderate influence category, demonstrating a significant yet intermediary impact on shear strength prediction compared to other factors.

5.1.2 The dual effect of some input features on the shear strength: s

Figure 11 shows effect of correlation between 2 features (selected randomly) on the shear strength using the PDP contour plots. For example, the relationship between humus content and water content in the contour PDP plot (Fig. 11(a)) shows that as we move along the x -axis (increasing water content), the clay shear strength tends to decrease. This trend is more pronounced when humus content is low. In other words, higher water content generally leads to less clay shear strength. Generally, an increase in humus content correlates with higher shear strength; however, the relationship is more pronounced with decreasing water content, as evidenced by a more substantial increase in shear strength with lower water content levels.

Another example (Fig. 11(b)) showcases the influence of clay content and the amount of binder ((%), wet) on shear strength. An increase in clay content typically corresponds to higher shear strength, while the opposite trend is observed when the amount of binder ((%), wet) increases, leading to decreased shear strength.

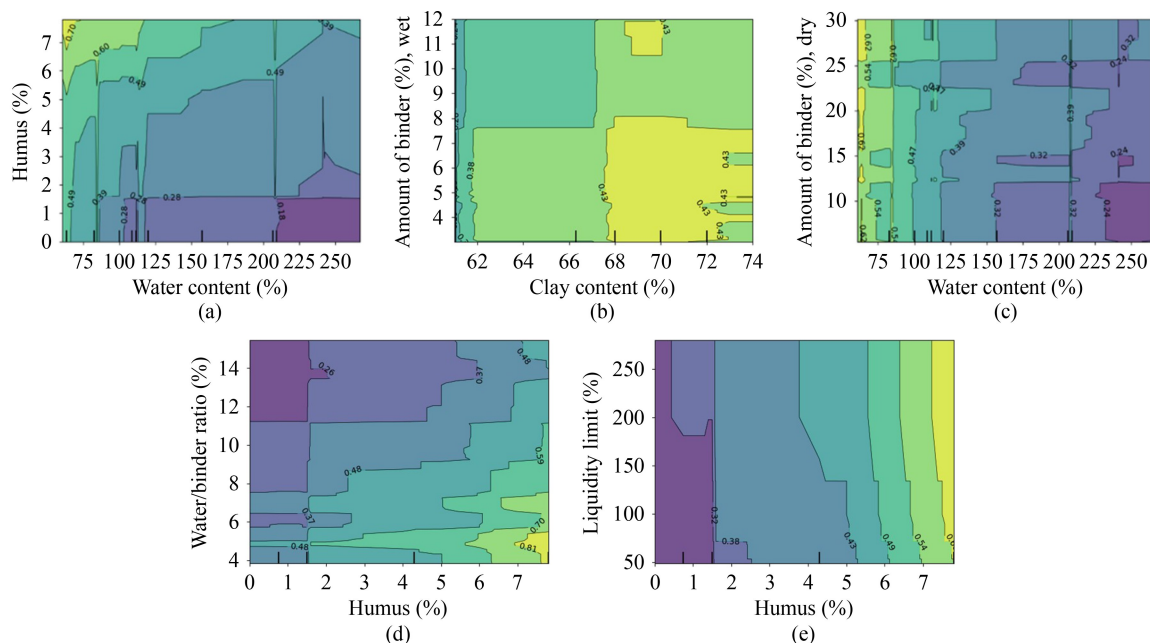


Fig. 11 Dual effect of some features on the shear strength: (a) humus–water content; (b) (amount of binder, wet)–clay content; (c) (amount of binder, dry)–water content; (d) water/binder ratio–humus; (e) liquidity limit–humus.

Interestingly, when both clay content and the amount of binder ((%), wet) are either low or high, the resulting shear strength tends to remain relatively consistent. This suggests a balancing effect wherein similar levels of these two factors mitigate their individual impacts on shear strength, resulting in comparable outcomes.

Figure 11(c) illustrates the dual effect of water content and the amount of binder ((%), dry) on shear strength. It is evident that an increase in water content, coupled with low values of the amount of binder ((%), dry), leads to lower shear strength, while the reverse scenario results in higher shear strength. Similarly, Fig. 11(d) demonstrates the interplay between Humus and water/binder ratio. In cases where Humus levels are high and the water/binder ratio is low, shear strength increases, whereas the opposite trend is observed. Additionally, Fig. 11(e) highlights the combined impact of Humus and liquidity limit on shear strength. As both Humus and liquidity limit increase, shear strength gradually increases, indicating a positive correlation between these factors and shear strength. These observations underscore the complex and interconnected nature of the factors influencing shear strength, wherein varying combinations of input variables lead to distinct outcomes.

Local vs global effect of input features.

The distinction between local and global effects of input features on shear strength lies in the scope and context of analysis. Locally, as elucidated by LIME, FIM varies across individual samples, showcasing the dynamic relationship between input variables and predicted outcomes. This variability stems from nuanced differences in sample characteristics and environmental conditions, reflecting the complex and context-dependent nature of shear strength prediction. Conversely, global FIM rankings, depicted through SHAP plots and the FIM (FIM) plot, provide a broader perspective by aggregating insights across the entire data set. Here, certain features consistently emerge as influential, such as the water content, water/binder ratio, and amount of binder ((%), wet), indicating their robust contribution to shear strength prediction across diverse scenarios. Based on both local and global interpretations, Binder Lime appears to exert a potentially more pronounced effect on shear strength compared to other types of binders (considering FIM and LIME). However, it is noteworthy that the SHAP plot highlights Binder Cement as having a more pronounced effect on shear strength compared to other binder types. This discrepancy underscores the importance of considering multiple analytical approaches to gain a comprehensive understanding of the factors influencing shear strength prediction, as different methods may reveal varying insights or nuances in the data. While local interpretations offer granularity and sensitivity to specific contexts, global analyses offer overarching trends and patterns, facilitating a comprehensive understanding of the factors driving shear strength prediction.

5.2 Interpretability tools comparison

SHAP and FIM plots. Interpretability tools, such as SHAP and FIM plots, offer valuable insights into the relative importance of features in predictive models. However, it's important to recognize that these tools may not provide exact interpretations of important features. As evidenced by discrepancies between the SHAP plot, which highlighted Humus as important, and the FIM plot, which indicated a moderate effect, such discrepancies underscore the need for complementary approaches. While both SHAP and FIM plots are valuable in identifying common important features (such as water content and water/binder ratio), other features may require interpretation using additional XAI tools. Hence, a combination of interpretability tools may be necessary to obtain a comprehensive understanding of the factors influencing model predictions. It's important to remember that FIM plots only provide a relative ranking of features based on their influence on the model's predictions. They don't necessarily represent the absolute contribution of each feature, and interactions between features might not be captured.

PDP, ICE, and ALE. On the other hand, the consistent trend observed in ALE plots, aligning with PDPs and ICE plots (as shown in Fig. 12 for selected features), reaffirms the reliability and robustness of the insights gained. It indicates that the trends observed are not isolated occurrences but are representative of broader patterns in the data.

Both ALE plots and PDPs provide consistent insights regarding the relationship between the feature in question and the model's predictions. This reinforces the understanding that the selected feature has a stable and significant impact on the predicted output, regardless of interactions with other features. Additionally, the similarity between ALE plots and PDPs indicates a degree of linearity in the relationship between the feature and the predicted shear strength.

Another important observation that these plots indicate water content and water/binder ratio are having significant effect on the shear strength (as indicated by SHAP and FIM plots), however PDP, ICE, and ALE plots indicate that amount of binder ((%), wet) does not have much effect on the shear strength (which is opposite to the FIM plot). In this case, the alignment between SHAP and FIM plots regarding water content and water/binder ratio suggests their robust influence, while the divergence with PDP, ICE, and ALE plots regarding the amount of binder ((%), wet) may warrant a closer examination. Consequently, this observation suggests a leaning toward the SHAP plot's ranking for the amount of binder ((%), wet), which places less emphasis on its importance compared to the FIM plot.

FIA and HM plots. The HM offers a visual snapshot of

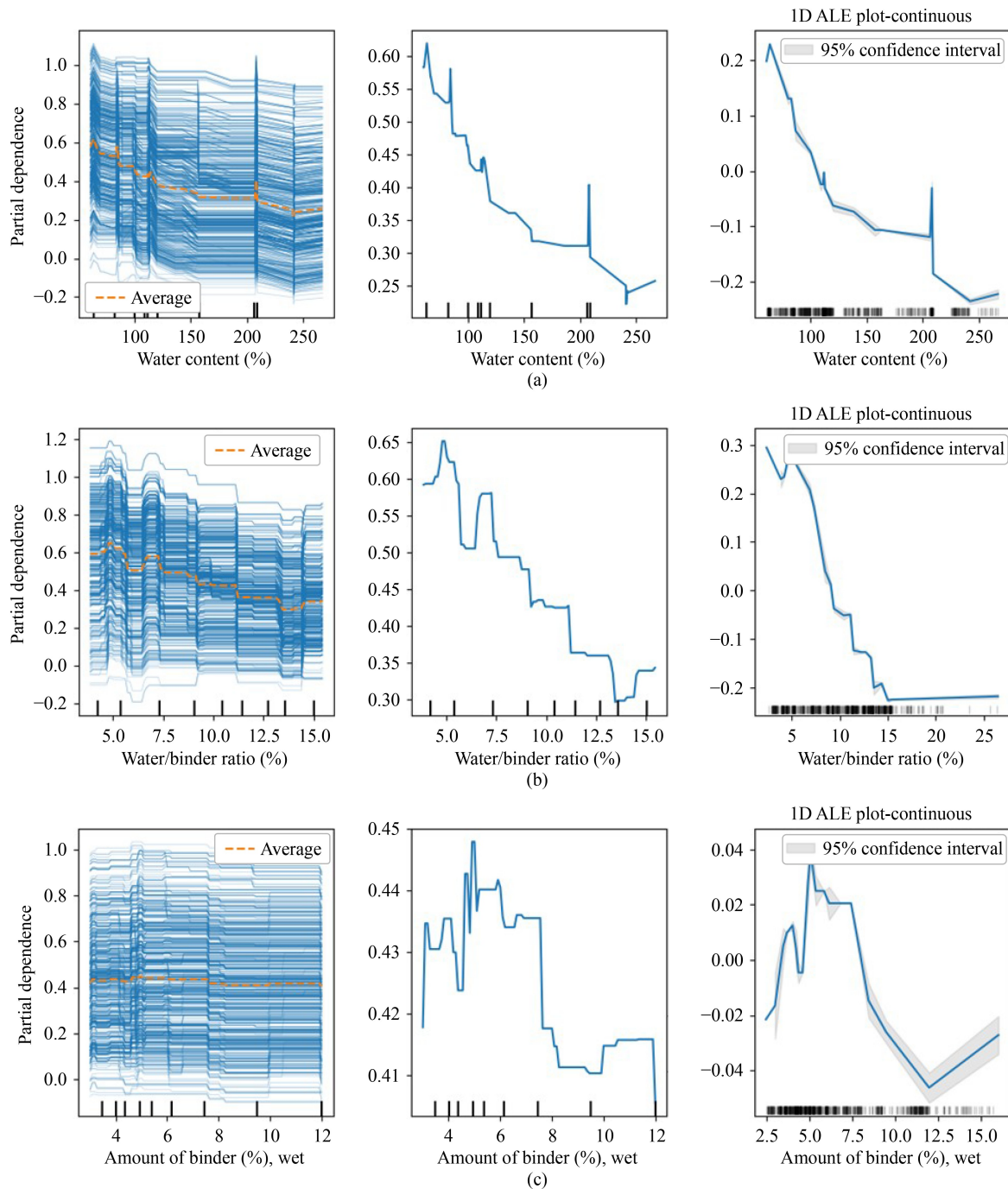


Fig. 12 Comparison PDP, ICE, and ALE for selected features: (a) water content; (b) water/binder ratio; (c) amount of binder, wet.

correlations (qualitative overview), providing an intuitive understanding of relationships, while the FIA employs a numerical ranking (quantitative assessment), offering a precise measure of influence. This complementary approach ensures a comprehensive assessment of the factors influencing shear strength. The comparison between the HM and FIA sheds valuable light on the influential factors impacting shear strength. Both methods affirm the significance of certain features. Notably, the Water/binder ratio emerges as a critical parameter in both analyses, underscoring its pivotal role in shear strength

determination. Similarly, Water content is highlighted by both approaches, reinforcing the importance of moisture content. In contrast, the FIA suggests that Clay content has moderate interaction effect on shear strength, while the HM shows it has important interactions effects on shear strength, especially with amount of binder ((%), wet and dry), liquidity limit, Humus, and water content. This indicates that HM and FIA are complementing one another offering a comprehensive understanding of the complex interplay influencing shear strength in stabilized soils.

5.3 Overall assessment and limitation of the framework

More than one tool is needed for assessment. The current study emphasizes the importance of using multiple interpretability tools to gain a comprehensive understanding of the factors influencing shear strength prediction for stabilized soils. Tools such as SHAP, FIM plots, LIME, HM, FIA, PDPs, ICE, and ALE plots offer valuable insights into the relative importance of input features and their impact on shear strength prediction. By combining these tools, researchers can obtain a holistic view of the complex interplay of factors affecting the model's predictions. The current study highlights instances where certain interpretability tools, such as SHAP and FIM plots, may reinforce the patterns identified by other tools. For example, the alignment between SHAP and FIM plots regarding the significance of water content and water/binder ratio suggests their robust influence on shear strength prediction. This consistency across different tools (PDP, ICE, ALE, FIA, and HM) enhances the reliability of the insights gained and provides a more robust understanding of the factors driving shear strength prediction in stabilized soils.

Limitation based on the data set provided in this study. One important limitation that should be considered is the data set-specific nature of the findings. The analysis and interpretations presented in the current study are based on the specific data set used in the study. Researchers should be cautious in extrapolating the findings of the current study to different scenarios and consider the unique characteristics of their own data sets when applying the interpretability tools discussed in the current study. Further validation and testing on diverse data sets are recommended to ensure the robustness and applicability of the framework beyond the current study.

6 Conclusions

The study introduces a novel XAI framework that advances the understanding of the factors influencing the shear strength of stabilized soils, which are widely used in deep stabilization applications in Nordic countries. This framework provides a comprehensive approach, progressing from statistical analysis to ML model training and detailed interpretation through a range of global (FIM, SHAP, PDP, ICE, ALE) and local (LIME) interpretability techniques.

Key findings.

1) Model Performance: XGBoost emerged as the top-performing model, achieving over 90% accuracy in predicting shear strength for clay soils.

2) Influential Features: Water content and the water/binder ratio were identified as primary drivers of shear strength, with the latter emerging as a critical

determinant for soil stability. While “Binder Cement” demonstrated the most substantial effect among binders, certain binders (e.g., Lime + Cement) and factors like liquidity limit were less influential.

3) Feature Interaction: PDP, ALE, and ICE plots confirmed that water content, water/binder ratio, and clay content significantly impact strength within specific ranges, while the HM and FIA underscored water/binder ratio's predominant influence, followed by water content and Binder–Cement.

4) Local Interpretability: LIME analysis highlighted variation in FIM across individual samples, underscoring the necessity of employing diverse interpretability tools to ensure a robust understanding of localized model behavior.

Limitations and Future Directions: A primary limitation of this study is the data set-specific nature of findings, which may restrict generalization to other soil types or geographic regions. Future research could expand the data set to include more diverse soil types and stabilization conditions, enabling broader applicability. Additionally, exploring other XAI methods, such as counterfactual explanations or hybrid interpretability approaches, could further refine insights into feature interactions and improve the framework's adaptability to various soil stabilization applications.

Funding note Open Access funding provided by Aalto University.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Competing interests The authors declare that they have no competing interests.

References

1. Firoozi A A, Guney Olgun C, Firoozi A A, Baghini M S. Fundamentals of soil stabilization. *International Journal of Geo-engineering*, 2017, 8: 1–16
2. Afrin H. A review on different types soil stabilization techniques. *International Journal of Transportation Engineering and Technology*, 2017, 3(2): 19–24
3. López Ramírez A, Korkiala-Tanttu L. Stabilisation of Malmi soft clay with traditional and low-CO₂ binders. *Transportation*

- Geotechnics, 2023, 38: 100920
4. Ramirez A L, Zhang Y, Forsman J, Korkiala-Tanttu L Q A. Stabilization of soft clay with sustainable binders for dry deep mixing design. *Geotechnical Testing Journal*, 2024, 47(1)
 5. Rabbani A, Samui P, Kumari S. Optimized ANN-based approach for estimation of shear strength of soil. *Asian Journal of Civil Engineering*, 2023, 24(8): 3627–3640
 6. Gupta P, Gupta N, Saxena K K, Goyal S. A novel hybrid soft computing model using stacking with ensemble method for estimation of compressive strength of geopolymer composite. *Advances in Materials and Processing Technologies*, 2022, 8(S3): 1494–1509
 7. Lämsivaara T T, Farhadi M S, Samui P. Performance of traditional and machine learning-based transformation models for undrained shear strength. *Arabian Journal of Geosciences*, 2023, 16(3): 183
 8. Pham B T, Hoang T A, Nguyen D M, Bui D T. Prediction of shear strength of soft soil using machine learning methods. *Catena*, 2018, 166: 181–191
 9. Ahmad M, Al-Mansob R A, Ramli A B B, Ahmad F, Khan B J. Unconfined compressive strength prediction of stabilized expansive clay soil using machine learning techniques. *Multiscale and Multidisciplinary Modeling, Experiments and Design*, 2024, 7(1): 217–231
 10. Ly H B, Pham B T. Prediction of shear strength of soil using direct shear test and support vector machine model. *Open Construction & Building Technology Journal*, 2020, 14(1)
 11. Taffese W Z, Abegaz K A. Prediction of compaction and strength properties of amended soil using machine learning. *Buildings*, 2022, 12(5): 613
 12. Sert S, Arslan E, Ocakbaşı P, Ekinci E, Garip Z, Özocak A, Ndepete C P. Stabilization of expansive clays with basalt fibers and prediction of strength by machine learning. *Arabian Journal for Science and Engineering*, 2024, 49: 13651–13670
 13. El-Din MN, Kim J. Simplified seismic life cycle cost estimation of a steel jacket offshore platform structure. *Structure and Infrastructure Engineering*, 2017, 13(8): 1027–44
 14. Noureldin M, Ali T, Kim J. Machine learning-based seismic assessment of framed structures with soil–structure interaction. *Frontiers of Structural and Civil Engineering*, 2023, 17(2): 205–223
 15. Ali T, Eldin M N, Haider W. The effect of soil–structure interaction on the seismic response of structures using machine learning, finite element modeling and ASCE 7–16 methods. *Sensors*, 2023, 23(4): 2047
 16. Phoon K K, Zhang W. Future of machine learning in geotechnics. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 2023, 17(1): 7–22
 17. Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. *Nature Reviews. Physics*, 2021, 3: 422–440
 18. Zhang W G, Goh A T C. Multivariate adaptive regression splines for analysis of geotechnical engineering systems. *Computers and Geotechnics*, 2013, 48: 82–95
 19. Zhang W, Goh A T C. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, 2016, 7: 45–52
 20. Zhang Z, Pan Q, Yang Z, Yang X. Physics-informed deep learning method for predicting tunnelling-induced ground deformations. *Acta Geotechnica*, 2023, 18: 4957–4972
 21. Zhou X, Shi P, Sheil B. Knowledge-based multiple point statistics for soil stratigraphy simulation. *Tunnelling and Underground Space Technology*, 2024, 143: 105475
 22. Zhang W, Wu C, Zhong H, Li Y, Wang L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, 2021, 12: 469–477
 23. Shabbir K, Noureldin M, Sim S H. Data-driven model for seismic assessment, design, and retrofit of structures using explainable artificial intelligence. *Computer-Aided Civil and Infrastructure Engineering*, 2025, 40(3): 281–300
 24. Yossef M, Noureldin M, Alqabbany A. Explainable artificial intelligence framework for FRP composites design. *Composite Structures*, 2024, 118190
 25. Anysz H, Brzozowski Ł, Kretowicz W, Narloch P. Feature importance of stabilised rammed earth components affecting the compressive strength calculated with explainable artificial intelligence tools. *Materials*, 2020, 13(10): 2317
 26. Naser M Z. An engineer's guide to eXplainable artificial intelligence and interpretable machine learning: Navigating causality, forced goodness, and the false perception of inference. *Automation in Construction*, 2021, 129: 103821
 27. Al-Bashiti M K, Naser M Z. Verifying domain knowledge and theories on fire-induced spalling of concrete through eXplainable artificial intelligence. *Construction and Building Materials*, 2022, 348: 128648
 28. Tapeh A T, Naser M Z. Discovering graphical heuristics on fire-induced spalling of concrete through explainable artificial intelligence. *Fire Technology*, 2022, 58(5): 2871–2898
 29. Naumets S, Lu M. Investigation into explainable regression trees for construction engineering applications. *Journal of Construction Engineering and Management*, 2021, 147(8): 04021084
 30. Zhan J, Fang W, Love P E, Luo H. Explainable artificial intelligence: Counterfactual explanations for risk-based decision-making in construction. *IEEE Transactions on Engineering Management*, 2024, 71: 10667–10685
 31. Ibrahim S M, Ansari S S, Hasan S D. Towards white box modeling of compressive strength of sustainable ternary cement concrete using explainable artificial intelligence. *Applied Soft Computing*, 2023, 149: 110997
 32. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence: A survey. 2020, arXiv: 2006.11371
 33. Longo L, Goebel R, Lecue F, Kieseberg P, Holzinger A. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In: *Proceedings of International cross-domain conference for machine learning and knowledge extraction*. Heidelberg: Springer International Publishing, 2020, 1–16
 34. Noureldin M, Abuhmed T, Saygi M, Kim J. Explainable probabilistic deep learning framework for seismic assessment of structures using distribution-free prediction intervals. *Computer-Aided Civil and Infrastructure Engineering*, 2023, 38(12): 1677–1698

35. Nasiri H, Homafar A, Chelgani S C. Prediction of uniaxial compressive strength and modulus of elasticity for Travertine samples using an explainable artificial intelligence. *Results in Geophysical Sciences*, 2021, 8: 100034
36. Dahal A, Lombardo L. Explainable artificial intelligence in geoscience: A glimpse into the future of landslide susceptibility modeling. *Computers & Geosciences*, 2023, 176: 105364
37. Liu W, Liu F, Fang W, Love P E. Causal discovery and reasoning for geotechnical risk analysis. *Reliability Engineering & System Safety*, 2024, 241: 109659
38. Lahtinen P, Kujala K. Lime and Cement Pillar Development Project. 1990 (in Finnish)
39. Kukko H, Ruohomäki J. Stabilization of Clays with Various Binders. Espoo: VTT Publications, 1995 (in Finnish)
40. Gharagoz M, Noureldin M, Kim J. Machine learning-based design of a seismic retrofit frame with spring-rotational friction dampers. *Engineering Structures*, 2023, 292: 16053
41. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Victoria: Leanpub, 2019
42. Friedman J H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2001, 29(5): 1189–1232
43. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 2015, 24(1): 44–65
44. Friedman J, Popescu B E. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2008, 2(3): 916–954
45. Marco S S, Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier. 2016, arXiv:1602.04938
46. Noureldin M, Ali A, Memon S, Kim J. Fragility-based framework for optimal damper placement in low-rise moment-frame buildings using machine learning and genetic algorithm. *Journal of Building Engineering*, 2022, 54: 104641
47. Noureldin M, Ali A, Sim S, Kim J. A machine learning procedure for seismic qualitative assessment and design of structures considering safety and serviceability. *Journal of Building Engineering*, 2022, 50: 104190
48. Noureldin M, Ali A, Nasab M S, Kim J. Optimum distribution of seismic energy dissipation devices using neural network and fuzzy inference system. *Computer-Aided Civil and Infrastructure Engineering*, 2021, 36(10): 1306–1321
49. Chen T, Guestrin C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM, 2016, 785–794