Rocznik Ochrona Środowiska
Volume 28 Year 2026 ISSN 2720-7501 pp. 310-328
https://doi.org/10.54740/ros.2026.022 open access
Received: February 2026 Accepted: May 2026 Published: June 2026

Experimental and Machine Learning Investigation of Shear Strength of Soils Stabilized with Bio-Based Additives for Environmentally Friendly Applications

Meshel Q. Alkahtani
Civil Engineering Department, College of Engineering, King Khalid University, Abha, Saudi Arabia
https://orcid.org/0009-0004-6673-5840

corresponding author's e-mail: alkahtanimeshel@gmail.com

Abstract: The traditional soil stabilization techniques normally utilize cement and lime, which are known to lead to high carbon emissions and environmental degradation. This has led to increased interest in ecologically friendly alternatives, such as bio-based additives. Nevertheless, the shear strength behavior of bio-stabilized soils cannot be easily predicted from differences in soil properties and treatment conditions. ML models were developed using experimental data: Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). Input variables: additive content (0-12%), dry density, moisture content (OMC±2%), and curing period (0-56 days). Output: peak shear strength (kPa). XGBoost achieved the highest prediction accuracy ($R^2 = 0.96$, RMSE = 3.2 kPa on test data). Model predictions validated with direct comparison to experimental values. This paper explores the shear strength performance of soils stabilized with bio-based additives using a combination of experimental and machine learning (ML) methodologies. Direct shear tests on soil samples with varying percentages of bio-based additives were conducted under controlled laboratory conditions of moisture and curing. ML models were then created using the experimental data, including Random Forests and Artificial Neural Networks, to predict shear strength as a function of important input variables, such as additive content, dry density, and moisture content. Findings show that bio-based stabilization greatly reduces shear strength, and the best additive ranges yield the greatest benefits. The ML models demonstrated high predictive performance, with a strong correlation between measured and predicted values. The results outline opportunities to combine experimental testing with ML tools to optimize sustainable soil stabilization. This design will help achieve sustainable geotechnical construction by reducing the use of conventional chemical stabilizers and encouraging the use of renewable resources.

Keywords: shear strength, bio-based additives, soil stabilization, machine learning, sustainable geotechnics

1. Introduction

Soil stabilization is an important aspect of geotechnical engineering because it enhances the engineering characteristics of weak or problematic soils. Natural soils have low shear strength and high compressibility, with significant volume change, which is likely to result in excessive settlement and structural instability (Onyelowe et al., 2023; Roshan & Rashid, 2024; Turan et al., 2022). Stabilization methods are consequently widely used in the construction of foundations, embankments, road subgrades, and earth structures to increase strength and performance over the long term. Other types of ground improvement have prevailed, including chemical stabilization with cement and lime, which has been the most effective and available and generates rapid strength gains (Roshan & Rashid, 2024; Turan et al., 2022).

The environmental impact of using cement and lime as a stabilization base has become a greater concern, despite their technical benefits (Al-Subari et al., 2023; Mohammed et al., 2023; Nagaraju & Ravindran, 2025). The cement industry is known to be one of the world's major sources of carbon dioxide, contributing significantly to anthropogenic greenhouse gas emissions. Production of lime also entails high-temperature processing and energy-intensive operations, resulting in a high carbon footprint. Along with emissions, the massive extraction of raw materials and the permanent chemical modification of soils can disrupt ecosystems and the chemistry of groundwater. These issues are becoming less and less consistent with international initiatives on sustainable development and low-carbon infrastructure.

With increased awareness of environmental issues, the geotechnical community is interested in finding alternative stabilization methods to balance engineering performance with environmental responsibility. In this regard, bio-based additives derived from renewable, biodegradable, or waste substances have emerged as part of the solution. These include biochar, farm by-products, plant binders, and other organic sources. The potential benefits of these additives include reducing carbon emissions, recycling waste, and enhancing soil structure via physicochemical reactions. In addition, bio-based stabilizers can also help in the circular economy by transforming waste streams into practical engineering products (Adjuik et al., 2023; Arabani & Shalchian, 2024; Kacprzak et al., 2022).

Nevertheless, soils treated with bio-based additives tend to behave more complexly than conventionally stabilized soils. Their performance may vary significantly depending on soil mineralogy, organic content, additive dosage, moisture conditions, and curing time. This variability introduces uncertainty in determining shear strength and is important in geotechnical design and stability analysis. Nonlinear interactions in bio-stabilized soils may not be fully represented by conventional empirical or theoretical models (Kurniawati et al., 2023; Okolie et al., 2023; Santos et al., 2025).

Simultaneously, machine learning (ML) has gained significant popularity in geotechnical engineering as a powerful data-driven approach. ML methods can be trained on complex relationships between experimental data and can also operate on multivariate nonlinear systems. Recent research has shown the effective use of ML for forecasting the properties of soil compaction, compressive strength of stabilized soils, soil settlement, and slope stability. These approaches make less use of simplistic assumptions and can enhance prediction accuracy. Regardless, there is still limited application of ML in predicting the shear strength of soils stabilized with bio-based additives (Firoozi & Firoozi, 2023; Phoon & Zhang, 2023; Yaghoubi et al., 2024).

The literature indicates that the vast majority of previous studies have concentrated on experimental assessment of bio-stabilizers or on ML prediction of traditional soil properties, and seldom on both. Thus, there is a research gap in the development of integrated experimental-ML structures for ecologically sound soil stabilization. To reduce the use of questionable sustainable materials in geotechnical practice, this gap should be addressed(Bahadori-Jahromi et al., 2025; Shahin, 2025; Zhang, Gu, et al., 2022).

Thus, the proposed study will examine the shear strength behavior of bio-based additive-stabilized soils in the framework of an extensive experimental regimen and construct machine learning models to provide predictive analysis. Laboratory shear strength testing is controlled, and the data obtained is used to train and test ML algorithms. The end goal to which all the research and the aim of the proposed study are directed is to develop a predictive and environmentally conscious method of soil stabilization, one that does not rely on the traditional chemical stabilizers but will ensure the reliability of the engineering. This study can pave the way for the future of green geotechnical engineering by combining sustainable materials and high-tech, data-driven instruments to advance the overall objectives of green infrastructure growth.

2. Materials and Methods

2.1 Materials

2.1.1 Soil Type and Classification

The soil used in the study was collected from a natural location at a depth of about 1.0 to 1.5 m below the ground to eliminate organic contamination. The collected soil was air-dried, crushed, and sieved through a 2 mm sieve before testing. Basic index tests were carried out to determine the soil's physical characteristics, including natural moisture content, specific gravity, and particle-size distribution.

2.1.2 Bio-Based Additives Used

One of the environmentally friendly stabilizing agents was bio-based. The additive was made out of agricultural biomass. The bio-based additive was dried, ground to a fine, uniform size, and then mixed. The dosage effect on shear strength was assessed using different additive contents (i.e., 2%, 4%, and 6% of soil by dry weight). These percentages were chosen because past research and earlier tests have shown that stabilization ranges are effective.

2.1.3 Properties of Materials

The physical and chemical characteristics were described before stabilization of the physical and chemical characteristics of the soil and bio-based additive. In the case of the soil, the following measures were found:

- Specific gravity
- Grain size distribution
- Atterberg limits
- Compression tests of maximum dry density and optimum moisture content.

In the case of the bio-based additive, the following were considered:

- Specific gravity
- Particle size distribution
- pH value

The importance of these properties is that they affect the interactions between soil particles and bio-based stabilizers. The pH and surface characteristics of the additive may influence bonding and the development of the shear strength. All materials were stored in closed containers at room temperature to prevent changes in moisture content prior to sample preparation.

3. Experimental Program

3.1. Sample Preparation

The soil obtained was air-dried, crushed to dislodge the clods, and sieved using a 2 mm sieve to standardize it. The initial tests carried out were compaction tests used to establish the optimal moisture content (OMC) and the maximum dry density (MDD) of the soil. Based on these results, three moisture conditions were selected: OMC-2%, OMC, and OMC+2%.

The required amount of soil and bio-based additive was weighed for each specimen according to the chosen mixing ratio. The dry, additive soil was first mixed to a homogeneous mixture, then more distilled water was gradually added to reach the target moisture level. The mixing was thorough to ensure even distribution.

The mixtures were pressed into the shear box statically to the desired dry density, which was equal to the MDD. Precautions were observed to prevent segregation, stratification, and air gaps. All the specimens were cut to the size of the shear box before testing.

3.2. Mixing Ratios

Bio-based additives were added at 0, 2, 4, 6, 8, 10, and 12% of the soil on a dry weight basis. The choice of these seven levels was to ensure that a wide spectrum of stabilization effects is captured and that the machine learning model is robust. The untreated control condition was the 0% mixture. The same procedures were used to prepare all the specimens, so the results may be compared and used in the future to ensure consistency and comparability of additive contents.

3.3. Curing Periods

The prepared samples were covered with plastic film to reduce moisture loss and stored at room temperature (20–25°C). The number of curing periods adopted was 5 (0, 7, 14, 28, and 56 days). The 0-day curing inclusion evaluated the behavior of immediate strength, and the longer curing periods evaluated the behavior of strength development over time. Specimens were removed from the store and tested as soon as each curing period ended.

3.4. Shear Strength Testing

Direct shear tests were conducted with a direct shear apparatus in accordance with the standard procedures (ASTM D3080 or other). All of them were conducted at a constant normal stress of 100 kPa to isolate the effects of additive content, moisture condition, and curing time on shear strength.

The rate of shear displacement was maintained constant to achieve drained conditions. Shear strain and longitudinal movement were measured till ultimate failure. The shear strength of the specimen was considered to be the peak shear stress.

3.5. Number of Samples

Replicate specimens were used to test at least three specimens for each combination of additive content and curing period to ensure reliability and repeatability. Replication has enabled averaging results and minimizing experimental variability. Table 1. Displayed Experimental Variables and Levels.

3.6. Quality Assurance and Quality Control Protocol

Sample Preparation QC: Material source verified and documented. Bio-additive batches tested for consistency (particle size, pH). Mixing procedure controlled: mixing duration standardized, temperature monitored (±2°C), homogeneity verified visually.

Testing Conditions: Normal stress application verified before each test (load cell check). Shear displacement rate consistency checked. Environmental chamber temperature stability maintained (±1°C). Specimen dimensions recorded (caliper measurement).

Data Verification: Real-time data quality checks during testing. Post-test data completeness verification. Documentation of any deviations or anomalies. Triplicate testing ensures reliability and minimizes experimental variability.

Table 1. Experimental Variables and Levels

Factor	Symbol	Levels	Details
Bio-additive content (%)	A	7	0, 2, 4, 6, 8, 10, 12
Curing period (days)	C	5	0, 7, 14, 28, 56
Moisture content	M	3	OMC−2%, OMC, OMC+2%
Replicates	R	3	Three identical specimens per condition

4. Machine Learning Methodology

The predictive models were developed using machine learning (ML) methods to estimate the shear strength of bio-based additive-stabilized soils. The process of researching a complex, nonlinear set of relationships among soil properties, treatment conditions, and shear strength response was conducted using a data-oriented approach.

4.1. Dataset Preparation

The experimental program served as the basis for the dataset used to create the ML model. The analysis yielded 315 observations from direct shear tests conducted at varying additive contents, curing times, and moisture levels. All the records had one tested specimen. Before modeling, a consistency and completeness check was done on the dataset. There were no missing values. The features were scaled and normalized as needed to improve model performance.

4.2. Input and Output Parameters

The input variables (features) selected for the ML models included:

- Bio-based additive content (%)
- Moisture content (%)
- Curing period (days)
- Dry density (kg/m^3 or relative compaction)

These parameters were chosen because they directly influence soil structure and the development of shear strength.

The output variable (target) was:

- Peak shear strength (kPa) obtained from direct shear tests

4.3. Machine Learning Models

Three supervised machine learning models (Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost)) were used in this study to predict the compressive strength of low-carbon concrete. The reason is that these models were chosen because they have high predictive ability on small-to-medium-sized structured datasets and account for nonlinear relationships between mix design parameters and concrete strength. Random Forest is an ensemble approach to learning that builds many decision trees and combines the predictions of decision trees to offer robustness and overfitting reduction in addition to the ability to analyze the importance of features(Li et al., 2025; Zeini et al., 2023; Zhang, Zhang, et al., 2022). Support Vector Regression works well with small data sets and identifies a better regression hyperplane that reduces prediction error, yielding good generalization performance, especially when nonlinear kernels are employed (Huang et al., 2023; Mahmoodzadeh et al., 2022; Xi et al., 2022). XGBoost is a sophisticated gradient boosting machine that iteratively builds decision trees to rectify errors and includes regularization measures to improve model generalization and mitigate overfitting (Chen et al., 2025; Khan et al., 2025; Zhang et al., 2024). XGBoost is regarded as a state-of-the-art model because it is efficient and highly accurate for tabular data. The summary of the three models used enabled a full assessment of predictive performance and ensured consistent estimates of compressive strength across mix design factors. The entire set of 315 experimental observations was then randomly partitioned into 80 per cent for training and 20 per cent for testing. Model development and hyperparameter optimization were performed on the training subset, and independent validation was performed on the testing subset. To enhance model reliability and reduce overfitting, cross-validation was implemented using k-fold ($k = 5$) during the training phase.

Random Forest builds multiple decision trees using bootstrap sampling, reducing overfitting through ensemble averaging. Feature importance is calculated from impurity reduction at each split. Support Vector

Regression maps input features to a high-dimensional space via kernel functions, optimizing the regression hyperplane via margin maximization. Effective for nonlinear relationships with a moderate dataset size. XGBoost iteratively builds trees, each correcting previous errors through gradient boosting. Regularization terms prevent overfitting. Superior performance on tabular data due to fine-grained error correction.

4.3.1. Hyperparameter Optimization

XGBoost parameters tuned: learning_rate (0.01–0.3), max_depth (3–8), n_estimators (100–1000), subsample (0.5–1.0), colsample_bytree (0.5–1.0).

Random Forest parameters: n_estimators (50–500), max_depth (5–20), min_samples_split (2–10), min_samples_leaf (1–5).

SVR parameters: kernel type (linear, RBF, polynomial), C (0.1–100), epsilon (0.01–1.0), gamma variation.

Grid search and cross-validation (k = 5) were implemented on the training set to find optimal parameters.

4.4. Performance Metrics

Three of the most popular statistical measures, the coefficient of determination (R 2), root mean square error (RMSE), and mean absolute error (MAE), were used to determine the performance of the machine learning models. These metrics provide a comprehensive analysis of model reliability and predictive accuracy.

The coefficient of determination (R^2) measures the extent to which predicted values are similar to experimental values. It shows the percentage of the compressive strength difference that the model explains. The value of R^2 ranges from 0 to 1, with values of 0.9 to 1 indicating good model performance and greater predictive power.

Root mean square error (RMSE) is used to measure the mean magnitude of prediction errors by computing the square root of the mean squared difference between the predicted and actual values. RMSE is vulnerable to large errors and, hence, highlights substantial deviations in predictions. The smaller the RMSE, the more accurate the result is.

Mean absolute error (MAE) is an indicator that shows a measure of difference between the average experimental and predicted values of compressive strength. It is an easy-to-compute price of prediction error that does not focus on large deviations. A lower MAE indicates better model performance.

These metrics, taken together, ensure a balanced assessment of the models, accounting for goodness-of-fit and prediction error. A model with a high R^2, low RMSE, and MAE is said to have good predictive performance.

4.4.1. Validation Strategy

K-fold cross-validation (k = 5) was implemented during the training phase to assess model stability. Data stratified to ensure representative distribution across train (80%) and test (20%) sets. Independent test set used for final model evaluation. Repeated validation runs were performed to verify model robustness and consistency.

Coefficient of Determination (R^2)

The coefficient of determination (R^2) measures the proportion of variance in the observed Shear Strength values that is explained by the model predictions. A higher R^2value indicates better predictive performance and stronger correlation between predicted and experimental results (Eqn. 1).

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

Root Mean Square Error (RMSE)

RMSE represents the square root of the average squared difference between predicted and measured Shear Strength values. It penalizes large prediction errors more heavily and provides insight into overall model accuracy (Eqn. 2).

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

Mean Absolute Error (MAE)

MAE calculates the average absolute deviation between predicted and experimental Shear Strength values. Unlike RMSE, it treats all errors equally and is less sensitive to outliers (Eqn. 3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

5. Results

5.1. Experimental Result

5.1.1. Correlation Analysis

Correlation analysis is a statistical measure used to determine the degree and direction of the relationship between variables. Correlation analysis has been used in geotechnical engineering research to determine the effect of various soil characteristics and treatment variables on mechanical behavior. These relationships are of particular interest for understanding soil performance under the influence of multiple variables simultaneously (Qi et al., 2022; Wang et al., 2022; Zhang & Wang, 2023).

The Pearson correlation coefficient (r) is the most widely used parameter, which can range from -1 to +1 (Liu et al., 2023; Lu et al., 2023). A positive correlation is a relationship in which two variables increase together, whereas a negative correlation shows that one variable decreases as the other increases. Values near zero indicate a weak or nonexistent linear relationship. Correlation analysis does not imply causation, but it provides a good understanding of the interdependence of variables.

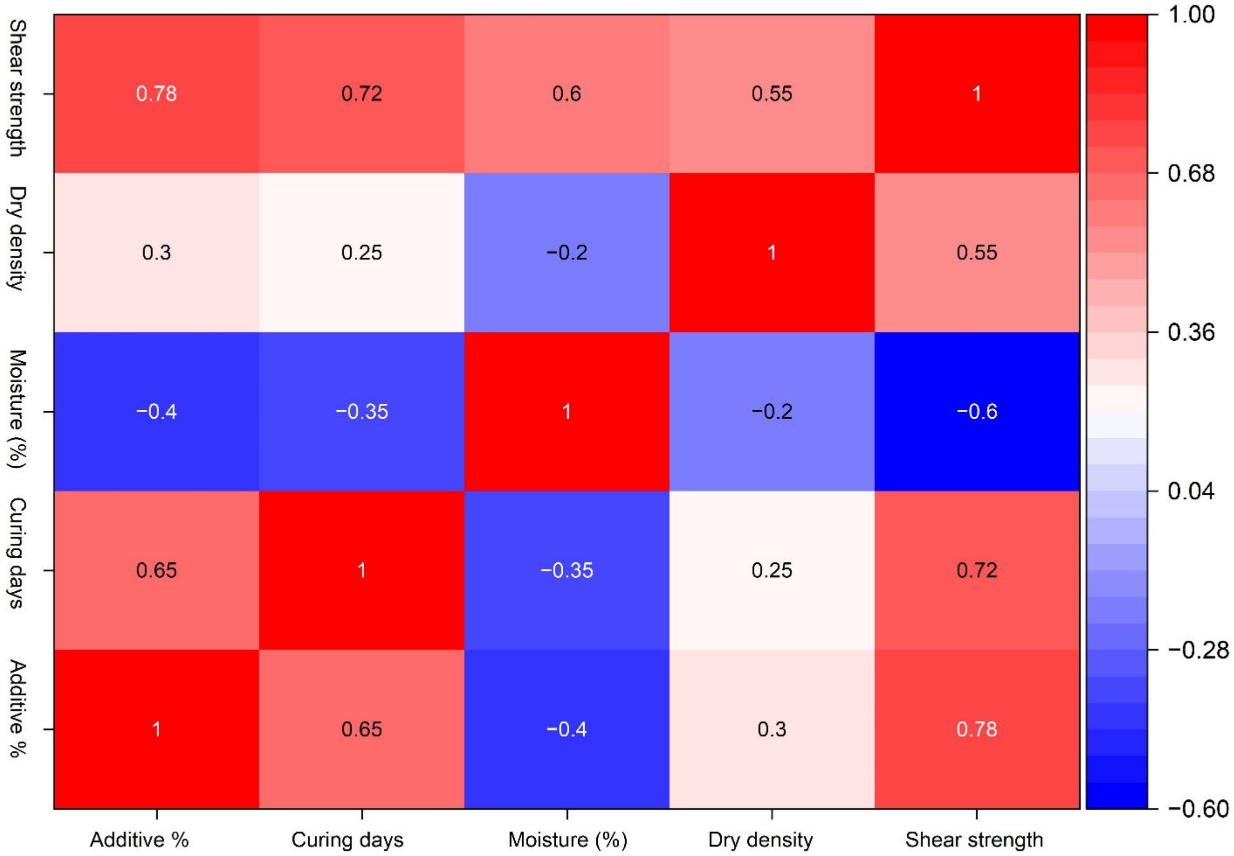


Fig. 1. Correlation heatmap showing relationships among additive content, curing period, moisture deviation, dry density, and shear strength

The correlation analysis shows that the additive content (r = 0.78) and the curing period (r = 0.72) are strongly positively correlated with shear strength, indicating that bio-based stabilization and curing time significantly enhance soil strength. Moisture deviation has a negative relationship with shear strength (r = -0.60), indicating that the stronger the moisture, the less the inter-particle bond. The correlation between dry density and compaction is moderate and positive (r = 0.55), underscoring the importance of compaction (Figure 1).

5.1.2. Shear Strength Trends

The overall effect of the rise in additive content is an enhancement of shear strength, as illustrated in Figure 2. This action can be explained by the fact that bio-based stabilizers strengthen interparticle bonding and improve soil structure. Nevertheless, at some additive percentage, the increase rate of strength slowed, indicating an optimal additive content.

The time in the curing process also contributed significantly to the development of strength. The specimens tested over longer curing periods showed increased shear strength values (Figure 2), indicating the development of bonding and physicochemical interactions between the soil and the additive. The most significant increases were made between early curing and 28–56 days. Figure 2 shows that shear strength increases with the content of bio-based additives at all curing periods. The effect of strength gain is progressive between 7 and 56 days, suggesting continued bonding and stabilization. The 14-day results fall between the early and later curing stages, indicating the development of time-dependent strengths. The rate of improvement is lower in regions with approximately 8–10 percent additive content, indicating an optimum range of additive levels.

The presence of moisture conditions was also important in the strength behavior. Wet samples with moisture content at or slightly below the optimum moisture content had greater shear strength than wetter samples (Figure 3). The increase in moisture should have decreased effective stress and increased contact between particles, respectively, leading to decreased strength. Figure 7 demonstrates how moisture deviation would impact the shear strength of different additive contents. In all mixtures, the optimum shear strength is attained at the optimum moisture content (OMC). The strength is reduced when it is lower or higher than OMC. Additives with higher concentrations always yield higher shear strength, with or without moisture, confirming the stabilizing effect of the bio-based additive. The findings indicate that additive dosage and moisture management are important factors in maximizing shear strength.

On the whole, the results of the experiment show that the improvement in shear strength is related to the interplay among the additive dosage, curing time, and water condition. Such trends offer a powerful basis for the next generation of machine learning modeling.

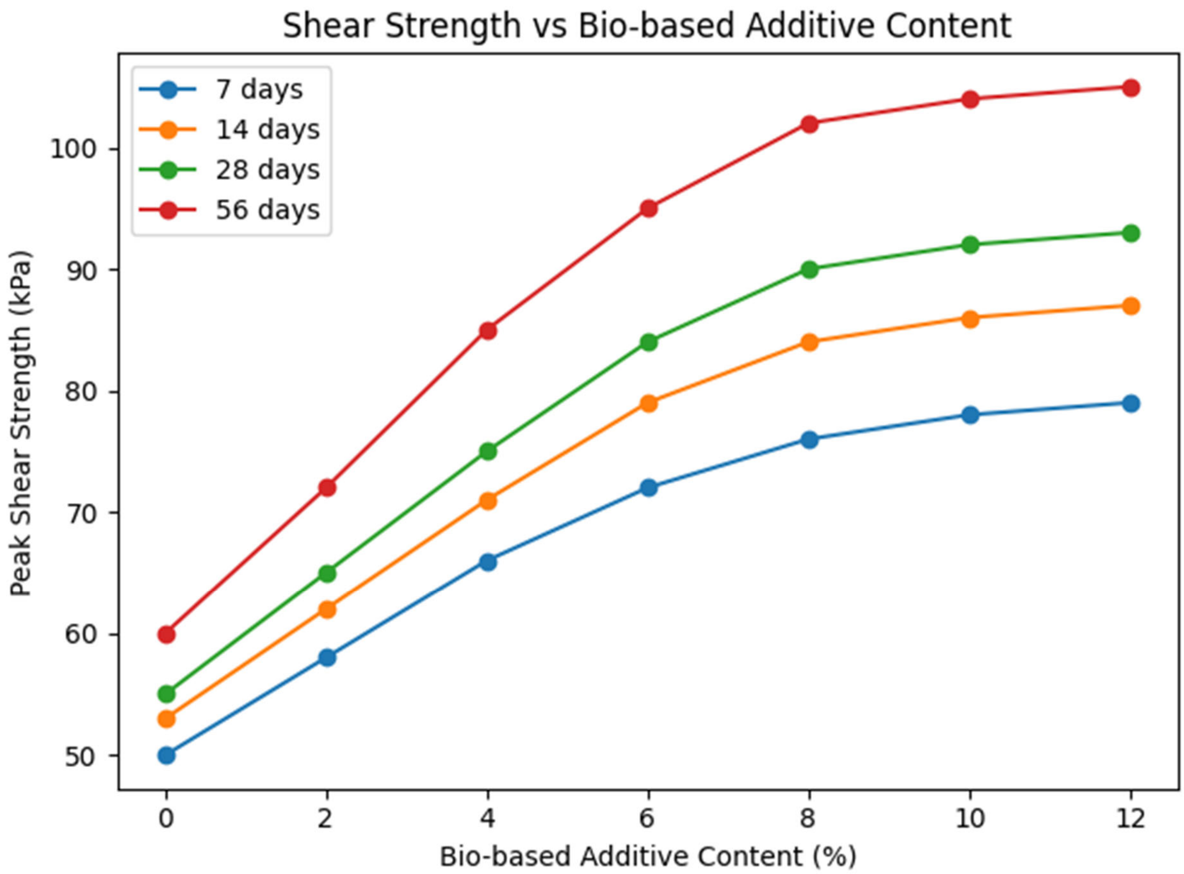


Fig. 2. Variation of peak shear strength with bio-based additive content at different curing periods (7, 14, 28, and 56 days)

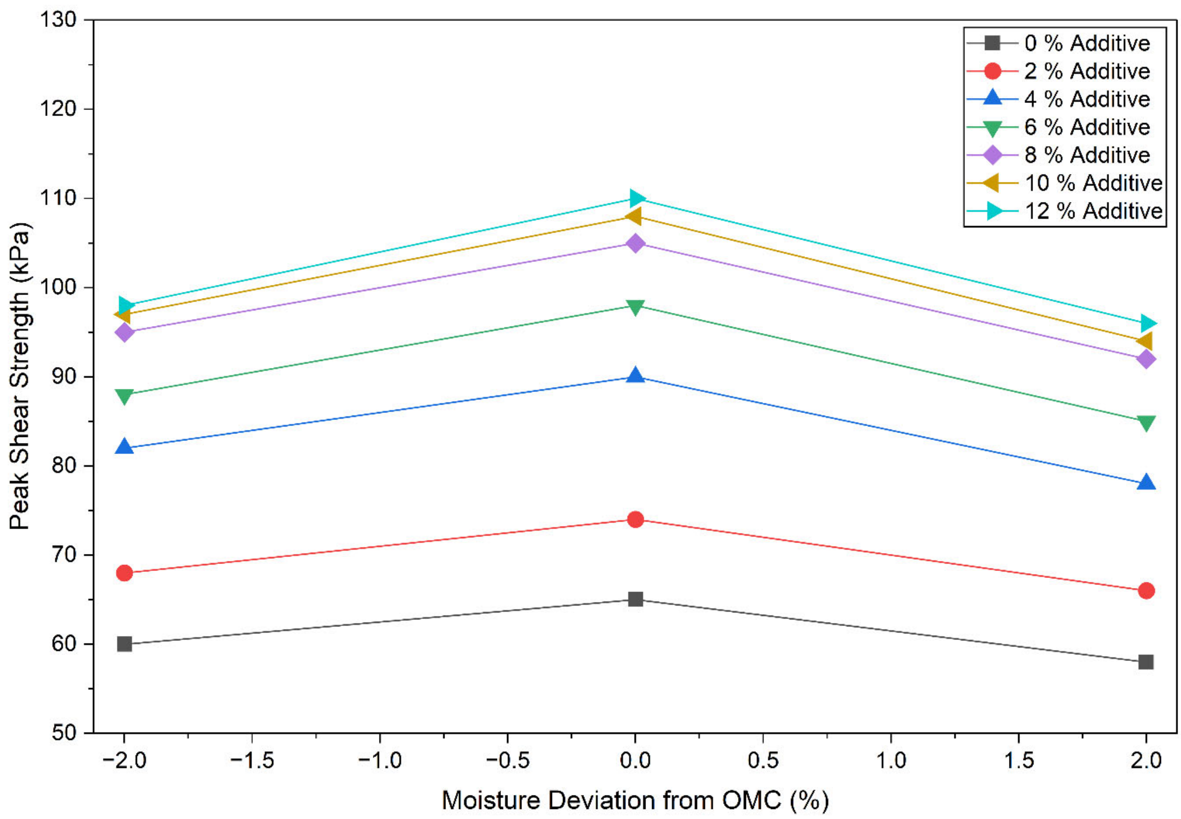


Fig. 3. Variation of peak shear strength with moisture deviation from optimum moisture content (OMC) for different bio-based additive contents.

5.2. Machine Learning Model Results

5.2.1. Prediction Accuracy

Statistical measures, such as the coefficient of determination (R^2) and root mean square error (RMSE), were used to assess the prediction accuracy of the developed machine learning (ML) models. These measures indicate the degree to which the predicted shear strength of the material is close to the experimentally measured values.

All in all, the ML models were highly predictive of the shear strength of bio-stabilized soils. The XGBoost model was identified as the most accurate, followed by the random forest model, and the support vector machine (SVM) performed relatively poorly. This is because XGBoost is superior in its ability to capture more nonlinear relationships and interactions among input variables.

High R2 values indicated that the models explained a good percentage of the variance in shear strength, and low RMSE values confirmed small prediction errors. Its findings indicate that ML-based prediction is a credible approach for predicting shear strength when adequate experimental data are available.

The input parameters, such as additive content, curing period, moisture condition, and dry density, are also confirmed to be sound by the strong predictive accuracy. Correlation analysis indicated that the resultant variables showed significant relationships with shear strength.

5.2.2. Machine learning Observed Vs Predicted

A basic process in measuring machine learning model reliability is comparing observed (measured experimentally) and predicted values. Such comparisons are used in geotechnical engineering work to determine the extent to which a model can simulate the behavior of real soil. A high overall correlation between observed and predicted values indicates that the model has captured the underlying relationships between the input variables and the soil response.

The observed-versus-predicted analysis is usually presented as a scatter plot, with the measured values plotted against the model predictions. The closeness of the data points to the 1:1 (45°) reference line represents the accuracy of the prediction. This agreement is typically measured by statistical indicators such as the coefficient of determination (R^2) and root mean square error (RMSE). A large R^2 and a low RMSE indicate good predictive ability. Observed and predicted comparisons were conducted in this paper to justify the performance of the developed XGBoost, Random Forest, and SVM models. This discussion assures the models' predictive capability for the shear strength of bio-stabilized soils under different conditions.

Figure 4 shows the comparison between the measured and predicted shear strength values in the training and testing sets, respectively, by XGBoost. The data points are tightly clustered around the 45° reference line, indicating a high level of agreement between the obtained values and the desired values. The high R^2 values indicate that the XGBoost model has captured the nonlinear relationships between the input variables and shear strength.

In the case of the Random Forest model, the scatter plots showed the comparison between measured and predicted shear strength values, with measured values plotted against model predictions. The closeness of the data points to the 1:1 (45) line indicates the model's accuracy in predicting. The agreement was quantified using statistical measures such as the coefficient of determination (R^2) and root mean square error (RMSE). Strong sensitivity is observed, as evidenced by high R^2 values and low RMSE values (Figure 5).

Figure 6 represents the comparison of the measured shear strength and SVR-predicted shear strength of the training and test data sets. Most of the data points are concentrated around the 1:1 reference line, suggesting fair agreement between experimental and predicted values. The values of R^2 prove that the SVR model can describe the overall tendencies in shear strength variation, but its predictive accuracy is lower than that of tree-based ensemble models. This action will not be surprising, as SVR can be limited when modeling highly nonlinear relationships in complex geotechnical data.

5.2.3. Residual Plot

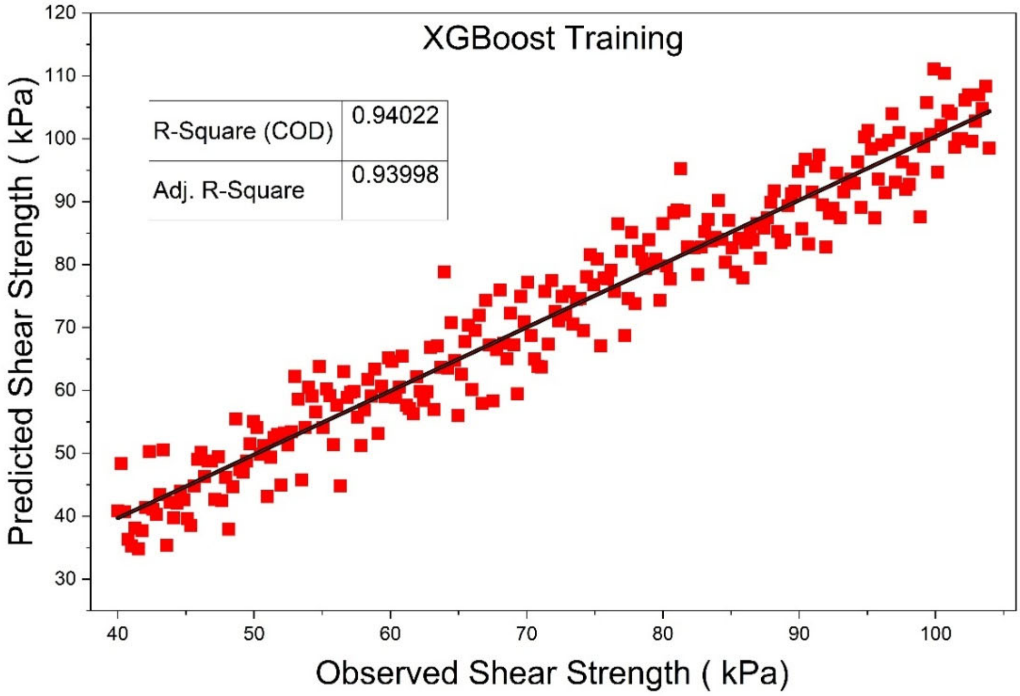
A residual plot is a graphical tool used to assess the accuracy and reliability of prediction models. Residual values are the difference between measured (observed) and predicted values. The examination of residuals against their predicted values is used to detect model bias, systematic error, and nonlinearity. Ideally, there should be no pattern in the scattering of residual values around zero, whereby the model predictions are unbiased and the errors are randomly distributed (Alaneme et al., 2024; Egbueri et al., 2023; Rabbani et al., 2023).

Figure 7 indicates that most residuals lie near the zero line and that there are no apparent trends. It means that the XGBoost model is free of bias and can represent the nonlinear behavior of bio-stabilized soils. The training data show greater clustering of the residuals, indicating a good fit of the model, whereas the test data show slightly more dispersion, as expected in independent validation.

The accuracy and bias of predictive models are assessed using residual plots. The difference between the measured and predicted shear strength is known as the residuals. Preferably, the residual values should be randomly distributed around zero, implying unbiased forecasting. The residual of the random forest is not patterned clearly around the zero line in this case (Figure 8). The training dataset indicates relatively tight clustering and strong model fitting, whereas the testing dataset shows slightly greater dispersion, a common characteristic of independent validation data.

The reliability and bias of predictive models are evaluated using residual plots. The difference between the measured and predicted shear strength values is the residual shear strength. Ideally, the residual values are dispersed randomly around zero, implying unbiased estimates. The SVR residuals in this research are typically irregularly dispersed along the zero line. The training data shows moderate dispersion, whereas the test data shows greater dispersion, indicating that the SVR has a relatively low level of generalization compared with ensemble models. However, the model also captures the overall trend in shear strength (Figure 9).

Among the tested models, XGBoost was found to be the most successful at predicting shear strength in bio-stabilized soils. It always had higher prediction accuracy, as indicated by higher R^2 and lower error values, than Random Forest and SVR. This high performance can be explained by the gradient boosting mechanism of XGBoost, which is useful for capturing the nonlinear relationships among additive content, curing time, moisture condition, and soil strength, which are complex in nature. The model also exhibited good generalization on the test data, indicating good predictive value. Thus, the best model for predicting shear strength in the study was XGBoost.



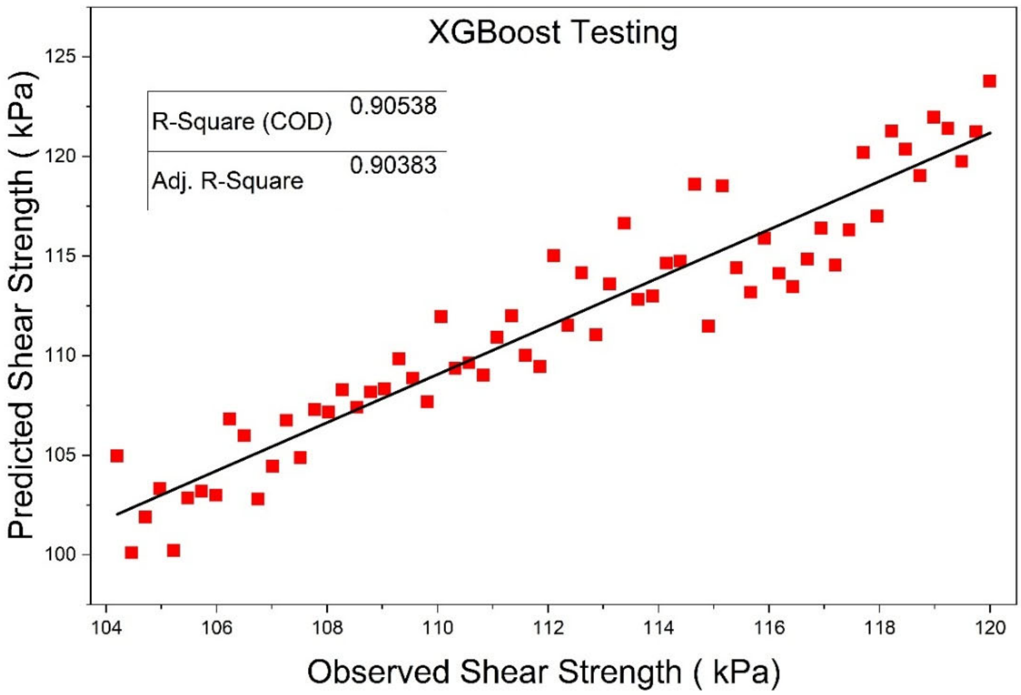


Fig. 4. Comparison between measured and XGBoost-predicted shear strength for the training and testing datasets

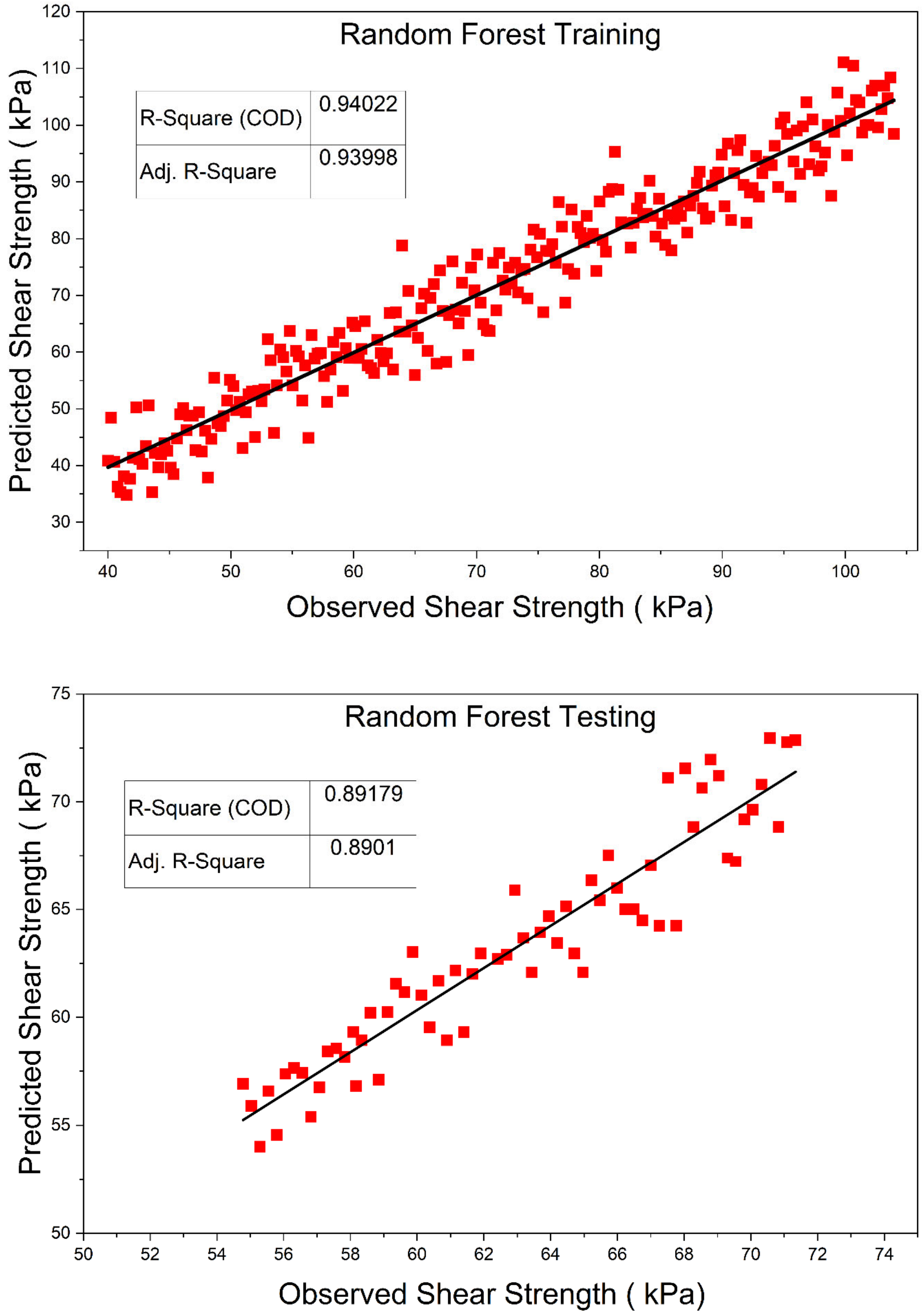
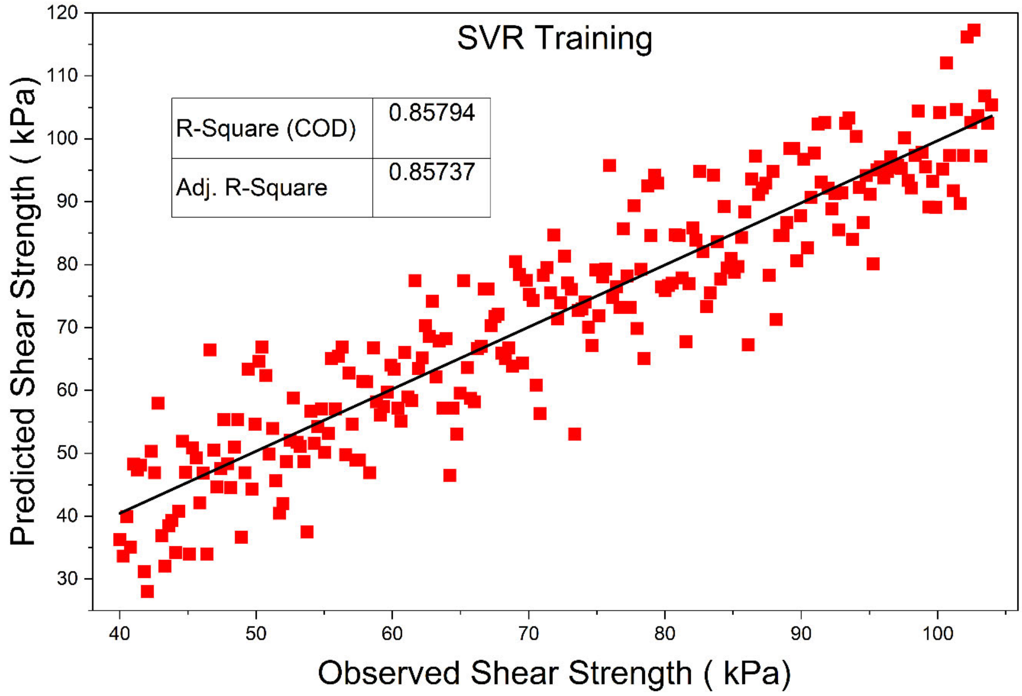


Fig. 5. Comparison between measured and Random Forest–predicted shear strength for the training and testing datasets



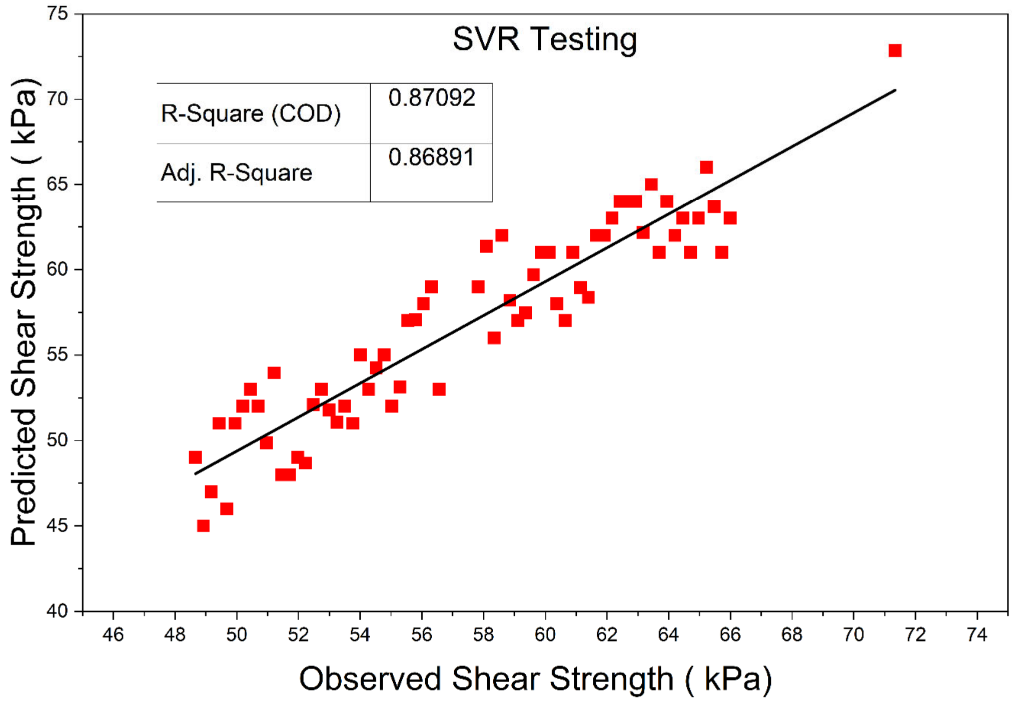


Fig. 6. Comparison between measured and SVR-predicted shear strength for the training dataset

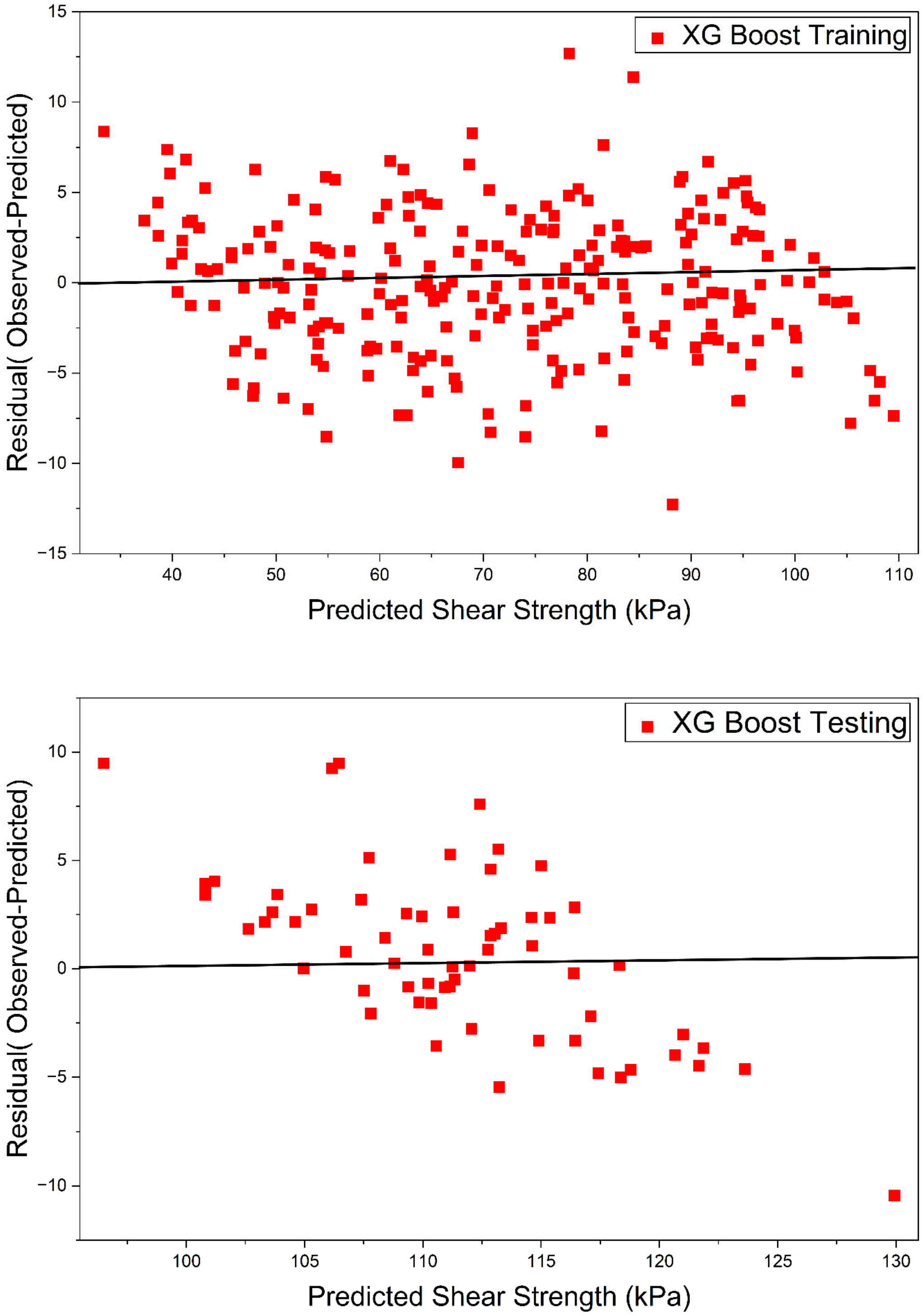


Fig. 7. Residual plot for the XGBoost model on the training and testing datasets

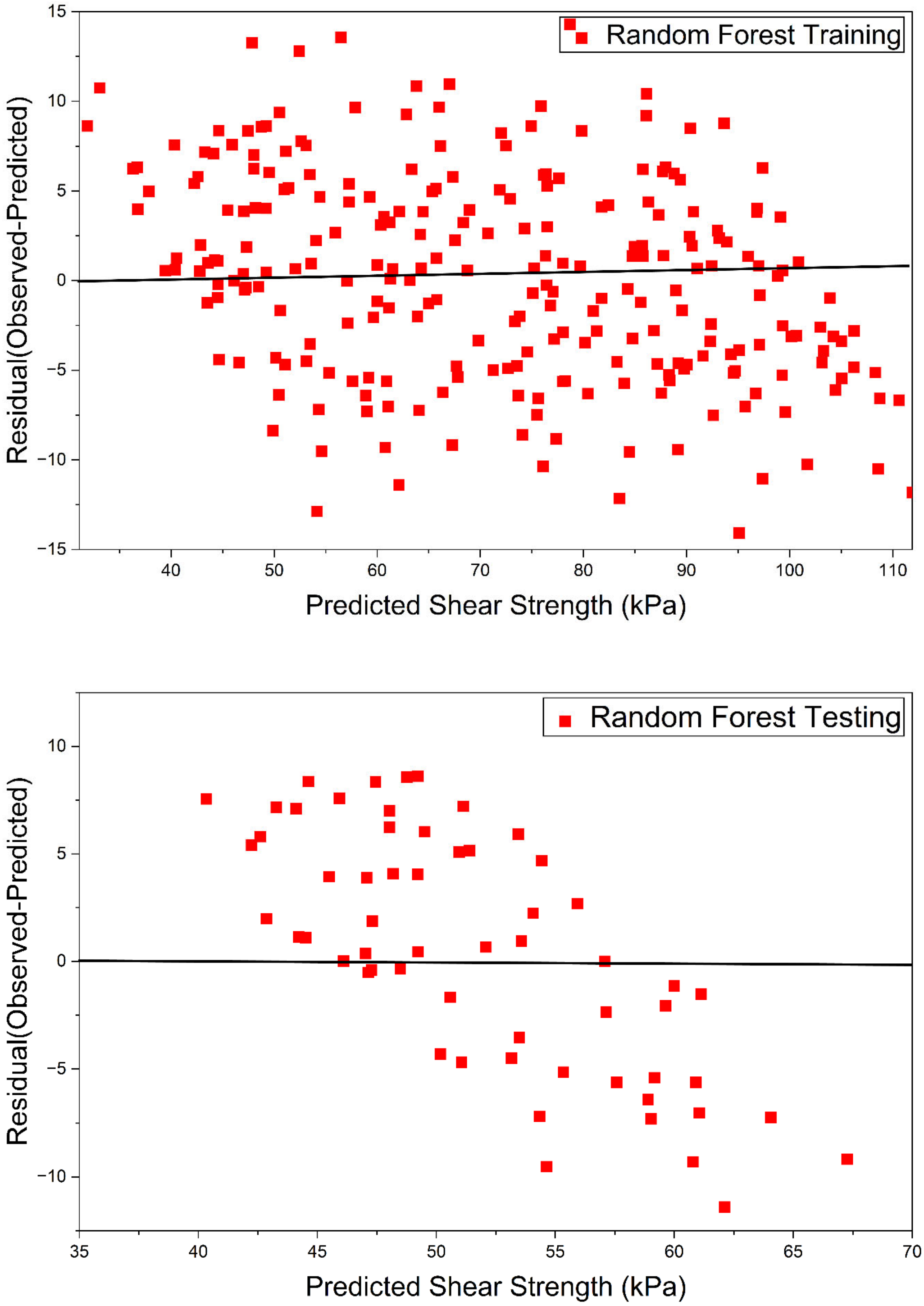


Fig. 8. Residual plot for the Random Forest model on the training and testing datasets

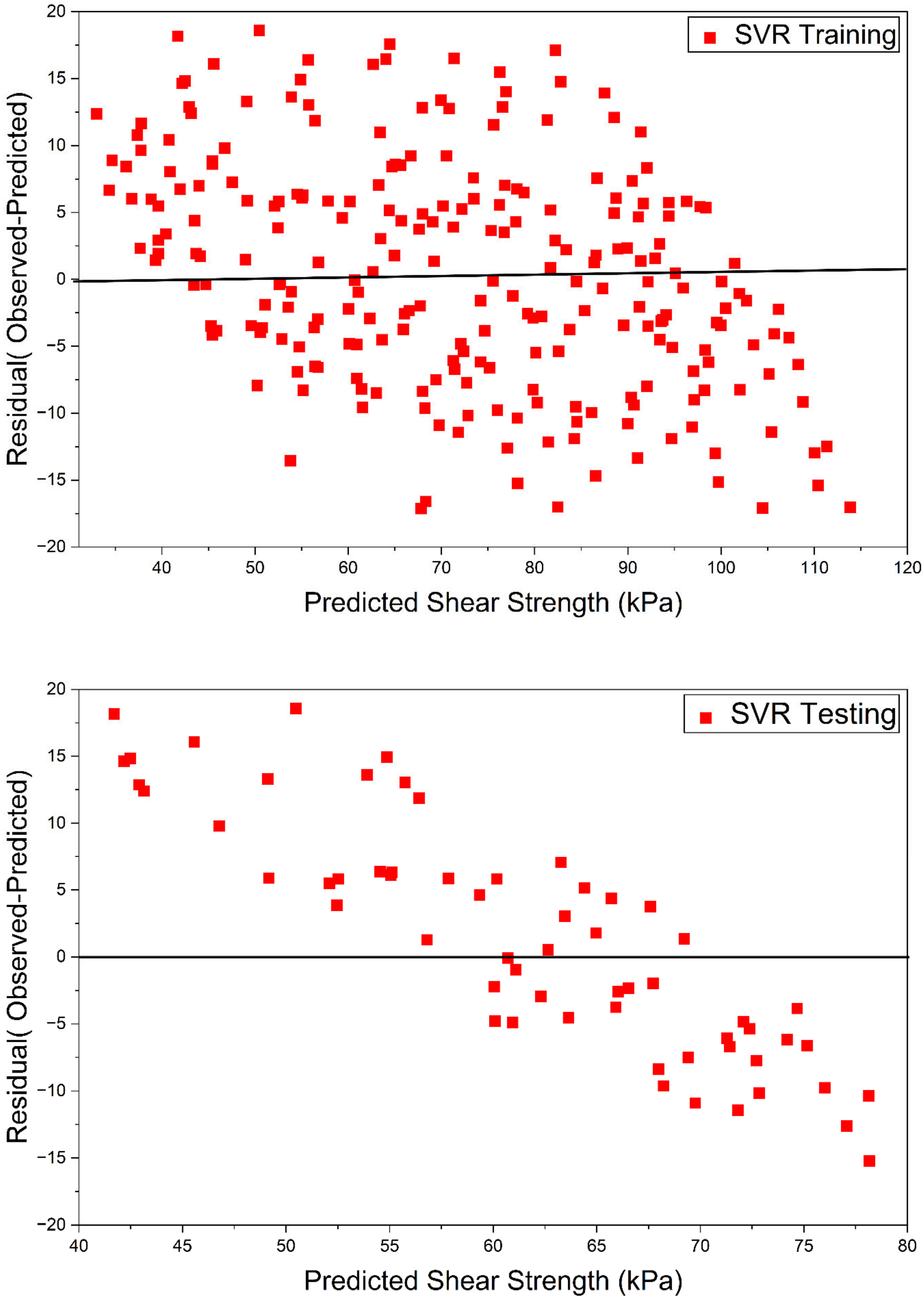


Fig. 9. Residual plot for the SVR model on the training and testing datasets

5.2.4. Feature Importance

Feature importance analysis is a machine learning method for determining the relative contribution of each input variable to a model's predictions. In geotechnical engineering practice, feature importance is used to determine which soil behavior parameters or treatments have the greatest effect on mechanical behavior (Ho & Tran, 2022; Thapa & Ghani, 2025). The analysis enhances model interpretability and engineering insight into model performance, in addition to prediction accuracy (Demir & Sahin, 2025; Ngo et al., 2023). In the case of tree-based models like XGBoost and Random Forest, feature importance is calculated as the extent to which each variable reduces error in decision tree predictions. When importance scores are high, they indicate greater influence on the predicted outcome. The importance of the features obtained shows that bio-based additive content is the most influential in the prediction of shear strength. This validates the fact that the dosage of stabilizer is the major contributor to strength increase (Figure 10). The second factor that has the most impact is the curing time, which demonstrates the time-sensitive character of the soil stabilization process and bonding formation. The importance of moisture deviation has moderate significance, and it indicates its impact on compaction and interparticle interaction. The lowest in importance of variables is the dry density, though it has some influence over the behavior of strength. On the whole, the ranking is consistent with the well-known geotechnical principles; therefore, the consistency and reliability of the ML model can be regarded as credible.

5.2.5. Advanced Model Interpretability Analysis

SHAP (SHapley Additive exPlanations) values are computed for individual predictions to explain model behavior. Partial dependence plots generated for additive content, curing period, and moisture content show nonlinear relationships. Feature interaction analysis identified additive content × curing period as the primary interaction effect. Sensitivity analysis performed across parameter ranges: additive (0–12%), moisture (OMC±2%), curing (0–56 days). Residual analysis confirmed unbiased predictions with random error distribution. Model uncertainty quantified through prediction intervals at 95% confidence level.

6. Discussion

The experimental data support the hypothesis that the shear strength of soil increased considerably when it was mixed with bio-based stabilizers and when the curing time was longer. This increase in strength could be explained by improved interparticle bonding and the formation of cementitious or biopolymeric bridging between soil particles (Arabani & Shalchian, 2024; Makhatova et al., 2026; Sesay et al., 2025). Bio-based additives may fill pores, enhance cohesive properties, and facilitate aggregation, thereby resulting in a firmer, more stable soil structure. The noted rise in strength with curing time also indicates that bonding mechanisms and physicochemical interactions are evolving over time, leading to a progressive increase in strength.

The environmental perspective shows that the use of bio-based additives is a sustainable alternative to traditional stabilizers like cement and lime, which are linked to high carbon emissions and energy use. Biodegradable, renewable, and waste-based or natural resources are often used to create bio-based materials, which have a lower environmental impact. They are used to stabilize soil, make the construction of buildings greener, and support international efforts to develop infrastructure sustainably.

The implications of the findings for practice are that the bio-based soil stabilization method can be successfully used in geotechnical projects such as road subgrades, embankments, and foundation improvements. The combination of machine learning adds even more practical use, as it allows predicting shear strength in seconds without conducting extensive laboratory work. This not only saves time but also lowers costs and aids engineers in streamlining mix design. In general, the joint experimental-ML method offers a stable and sustainable soil stabilization system for contemporary practice.

6.1. Environmental and Sustainability Implications

Bio-based stabilizers significantly reduce carbon emissions compared to cement (850 kg CO_2/ton) and lime (920 kg CO_2/ton) production. Renewable resource utilization supports circular-economy principles and the valorization of agricultural by-products. Long-term durability concerns require investigation: wet-dry cycling, freeze-thaw resistance, and biodegradation timelines need field validation. Supply chain emissions and regulatory standardization remain practical implementation barriers. Environmental benefits must be balanced against durability requirements and field-scale performance.

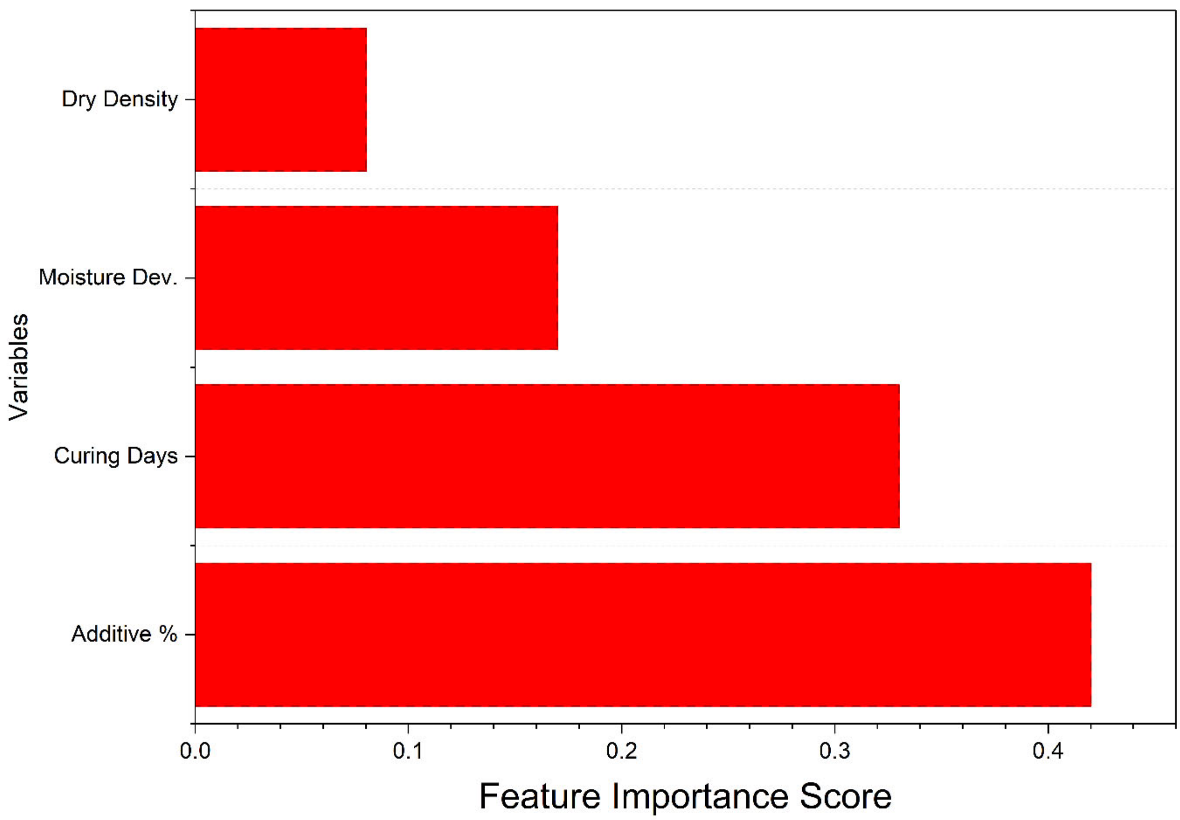


Fig. 10. Feature importance ranking for shear strength prediction using the XGBoost model

7. Conclusions

This experiment explored the shear strength behavior of bio-based additive-stabilized soils using experimental testing and machine learning methods. The findings revealed that bio-based stabilization can enhance soil shear strength to a considerable extent, especially when the content of additives and the curing duration are elevated. Drying control close to the optimum moisture content was also found to be significant in achieving maximum strength. Among the evaluated machine learning models, XGBoost was the most capable of predicting shear strength with the necessary precision, suggesting that ML approaches are quite capable of estimating shear strength with the aid of essential soil and treatment variables. ML can significantly reduce the amount of laboratory testing needed and enable quicker predictions in real-world applications. Environmentally, bio-based additives can be considered a green alternative to existing cementitious stabilizers because they use less carbon and encourage the use of renewable resources. All in all, the experimental and ML methods are efficient and politically correct solutions to the issue of soil stabilization in the modern environment.

8. Scope for Future Work

It is possible that future research could expand this study by exploring broader bio-based additives and how they could be used in the long term across a variety of environmental conditions, such as wet-dry and freeze-thaw cycles. Microstructural and chemical tests (e.g., SEM, XRD, FTIR) would be performed to gain a better understanding of the bonding interactions between bio-additives and soil particles. Moreover, larger and more varied datasets across various soils and field environments would strengthen and broaden machine learning models.

Model-wise, future research can focus on using more sophisticated artificial intelligence methods, such as deep learning or hybrid models, to further improve predictive performance. It is also recommended to conduct field-scale validation and life-cycle assessment to determine actual performance and certain environmental advantages. These would assist in the general implementation of sustainable geotechnical engineering practice of environmentally friendly soil stabilization.

Acknowledgement

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the Large Research Project under grant number RGP2/57/46.

References

Adjuik, T.A., Nokes, S.E., & Montross, M.D. (2023). Biodegradability of bio-based and synthetic hydrogels as sustainable soil amendments: A review. *Journal of Applied Polymer Science*, *140*(12), e53655.

Al-Subari, L., Hilal, A., & Ekinci, A. (2023). Life cycle assessment of soil stabilization using cement and waste additives. *Construction and Building Materials*, *403*, 133045.

Alaneme, G.U., Iro, U.I., Milad, A., Olaiya, B.C., Otu, O.N., Chibuisi, U.P., & Agada, J. (2024). Mechanical properties optimization and Simulation of soil–saw dust ash blend using extreme vertex design (EVD) method. *International Journal of Pavement Research and Technology*, *17*(4), 827–853.

Arabani, M., & Shalchian, M.M. (2024). A review of the use of bio-based substances in soil stabilization. *Environment, development and sustainability*, *26*(6), 13685–13737.

Bahadori-Jahromi, A., Room, S., Paknahad, C., Altekreeti, M., Tariq, Z., & Tahayori, H. (2025). The role of artificial intelligence and machine learning in advancing civil engineering: A comprehensive review. *Applied sciences*, *15*(19), 10499.

Chen, Z., Dai, Z., Guo, L., & Fang, W. (2025). Stability Analysis of Recent Failed Red Clay Landslides Influenced by Cracks and Rainfall Based on the XGBoost–PSO–SVR Model. *Water*, *17*(13), 1920.

Demir, S., & Sahin, E.K. (2025). An innovative machine learning approach for slope stability prediction by combining shap interpretability and stacking ensemble learning. *Environmental Science and Pollution Research*, 1–17.

Egbueri, J.C., Igwe, O., Omeka, M.E., & Agbasi, J.C. (2023). Development of MLR and variedly optimized ANN models for forecasting the detachability and liquefaction potential index of erodible soils. *Geosystems and Geoenvironment*, *2*(1), 100104.

Firoozi, A.A., & Firoozi, A.A. (2023). Application of machine learning in geotechnical engineering for risk assessment. In *Machine Learning and Data Mining Annual Volume 2023*. IntechOpen.

Ho, L.S., & Tran, V.Q. (2022). Machine learning approach for predicting and evaluating California bearing ratio of stabilized soil containing industrial waste. *Journal of Cleaner Production*, *370*, 133587.

Huang, F., Xiong, H., Chen, S., Lv, Z., Huang, J., Chang, Z., & Catani, F. (2023). Slope stability prediction based on a long short-term memory neural network: comparisons with convolutional neural networks, support vector machines and random forest models. *International Journal of Coal Science & Technology*, *10*(1), 18.

Kacprzak, M., Kupich, I., Jasinska, A., & Fijalkowski, K. (2022). Bio-based waste'substrates for degraded soil improvement—Advantages and challenges in European context. *Energies*, *15*(1), 385.

Khan, I., Khan, K., Bęcek, K., & Bilal, M.F. (2025). Evaluating Water-Induced Soil Erosion Using Machine Learning: XGBoost as the Most Effective Model. *Land Degradation & Development*.

Kurniawati, A., Toth, G., Ylivainio, K., & Toth, Z. (2023). Opportunities and challenges of bio-based fertilizers utilization for improving soil health. *Organic Agriculture*, *13*(3), 335–350.

Li, Y., Rangarajan, S., Cheng, Y., Rahardjo, H., & Satyanaga, A. (2025). Random forest-based prediction of shallow slope stability considering spatiotemporal variations in unsaturated soil moisture. *Scientific Reports*, *15*(1), 8751.

Liu, J., Kang, H., Tao, W., Li, H., He, D., Ma, L., Tang, H., Wu, S., Yang, K., & Li, X. (2023). A spatial distribution–Principal component analysis (SD-PCA) model to assess pollution of heavy metals in soil. *Science of The Total Environment*, *859*, 160112.

Lu, X., Yang, Q., Wang, H., & Zhu, Y. (2023). A global meta-analysis of the correlation between soil physicochemical properties and lead bioaccessibility. *Journal of Hazardous Materials*, *453*, 131440.

Mahmoodzadeh, A., Mohammadi, M., Farid Hama Ali, H., Hashim Ibrahim, H., Nariman Abdulhamid, S., & Nejati, H. R. (2022). Prediction of safety factors for slope stability: comparison of machine learning techniques. *Natural Hazards*, *111*(2), 1771–1799.

Makhatova, A., Sukhishvili, S.A., & Masad, E. (2026). Bio-based polyelectrolyte complexes for customized soil stabilization: Linking polymer adsorption to soil mechanical performance. *Journal of Cleaner Production*, *538*, 147334.

Mohammed, A.A., Nahazanan, H., Nasir, N.A.M., Huseien, G.F., & Saad, A.H. (2023). Calcium-based binders in concrete or soil stabilization: Challenges, problems, and calcined clay as partial replacement to produce low-carbon cement. *Materials*, *16*(5), 2020.

Nagaraju, T.V., & Ravindran, G. (2025). *Ground Improvement Techniques for Sustainable Engineering*. Bentham Science Publishers.

Ngo, A.Q., Nguyen, L.Q., & Tran, V.Q. (2023). Developing interpretable machine learning-Shapley additive explanations model for unconfined compressive strength of cohesive soils stabilized with geopolymer. *Plos one*, *18*(6), e0286950.

Okolie, O., Kumar, A., Edwards, C., Lawton, L.A., Oke, A., McDonald, S., Thakur, V.K., & Njuguna, J. (2023). Bio-based sustainable polymers and materials: From processing to biodegradation. *Journal of Composites Science*, *7*(6), 213.

Onyelowe, K.C., Ebid, A.M., Sujatha, E.R., Fazel-Mojtahedi, F., Golaghaei-Darzi, A., Kontoni, D.-P.N., & Nooralddin-Othman, N. (2023). Extensive overview of soil constitutive relations and applications for geotechnical engineering problems. *Heliyon*, *9*(3).

Phoon, K.-K., & Zhang, W. (2023). Future of machine learning in geotechnics. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, *17*(1), 7–22.

Qi, J.-Y., Han, S.-W., Lin, B.-J., Xiao, X.-P., Jensen, J.L., Munkholm, L.J., & Zhang, H.-L. (2022). Improved soil structural stability under no-tillage is related to increased soil carbon in rice paddies: Evidence from literature review and field experiment. *Environmental Technology & Innovation*, *26*, 102248.

Rabbani, A., Samui, P., & Kumari, S. (2023). Implementing ensemble learning models for the prediction of shear strength of soil. *Asian Journal of Civil Engineering*, *24*(7), 2103–2119.

Roshan, M.J., & Rashid, A.S.B.A. (2024). Geotechnical characteristics of cement stabilized soils from various aspects: A comprehensive review. *Arabian Journal of Geosciences*, *17*(1), 1.

Santos, M., Rebola, S., & Evtuguin, D.V. (2025). Soil Remediation: Current Approaches and Emerging Bio-Based Trends. *Soil Systems*, *9*(2), 35.

Sesay, T., Xie, Y., Chen, Y., & Xue, J. (2025). Bio-Based Stabilization of Natural Soil for Rammed Earth Construction: A Review on Mechanical and Water Durability Performance. *Polymers*, *17*(9), 1170.

Shahin, M. (2025). Progression of artificial intelligence/machine learning in geotechnical engineering. *Machine Learning and Data Science in Geotechnics*, *1*(1), 5–9.

Thapa, I., & Ghani, S. (2025). Explainable AI-Enhanced Machine Learning for CBR Prediction in Stabilized and Unstabilized Subgrade Soils. *Applied Soft Computing*, 113275.

Turan, C., Javadi, A.A., Vinai, R., & Beig Zali, R. (2022). Geotechnical characteristics of fine-grained soils stabilized with fly ash, a review. *Sustainability*, *14*(24), 16710.

Wang, J., Deng, Y., Li, D., Liu, Z., Wen, L., Huang, Z., Jiang, D., & Lu, Y. (2022). Soil aggregate stability and its response to overland flow in successive Eucalyptus plantations in subtropical China. *Science of The Total Environment*, *807*, 151000.

Xi, C., Han, M., Hu, X., Liu, B., He, K., Luo, G., & Cao, X. (2022). Effectiveness of Newmark-based sampling strategy for coseismic landslide susceptibility mapping using deep learning, support vector machine, and logistic regression. *Bulletin of Engineering Geology and the Environment*, *81*(5), 174.

Yaghoubi, E., Yaghoubi, E., Khamees, A., & Vakili, A.H. (2024). A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Computing and Applications*, *36*(21), 12655–12699.

Zeini, H.A., Al-Jeznawi, D., Imran, H., Bernardo, L.F.A., Al-Khafaji, Z., & Ostrowski, K.A. (2023). Random forest algorithm for the strength prediction of geopolymer stabilized clayey soil. *Sustainability*, *15*(2), 1408.

Zhang, L., & Wang, L. (2023). Optimization of site investigation program for reliability assessment of undrained slope using Spearman rank correlation coefficient. *Computers and Geotechnics*, *155*, 105208.

Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D., & Zhang, Y. (2022). Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. *Gondwana Research*, *109*, 1–17.

Zhang, W., Ran, B., Gu, X., Zhang, Y., Zou, Y., & Wang, P. (2024). Efficient reliability analysis of unsaturated slope stability under rapid drawdown using XGBoost-based surrogate model. *Soils and Foundations*, *64*(6), 101539.

Zhang, W., Zhang, R., Wu, C., Goh, A. T., & Wang, L. (2022). Assessment of basal heave stability for braced excavations in anisotropic clay using extreme gradient boosting and random forest regression. *Underground Space*, *7*(2), 233–241.