Eurasian J. Sci. Eng., 2026, 12(1), 39-53
https://doi.org/10.23918/eajse.v12i1p3

Published by Tishk International University
Available at: https://eajse.tiu.edu.iq/

RESEARCH ARTICLE Open Access

Data-Driven Soil Classification in Shaqlawa District, Iraq: Integrating Statistical Analysis with Machine Learning Models

Asmaa Abdulmajeed Mamhusseini [1*], Sardar Majeed Omar [2], Ghazala Younus Asaad [3], Munna Matte Habib [2], and Farzand Kamal Medhat [4]

[1] Civil Engineering Department, Faculty of Engineering, Tishk International University, Erbil, Iraq
[2] Shaqlawa Technical Institute, Erbil Polytechnic University, Erbil City, Kurdistan Region, Iraq
[3] Department of Construction and Material Technical Engineering, Erbil Polytechnic University, Erbil, Iraq
[4] Erbil Polytechnic University-Erbil, Kurdistan Region, Iraq

Article History
Received: 06.10.2025
Revised: 31.03.2026
Accepted: 11.06.2026
Published: 14.06.2026
Communicated by: Prof. Dr. Bayan Salim
*Email address:
asmaa.abdulmajeed@tiu.edu.iq
*Corresponding Author

Abstract: This paper presents a comprehensive framework for geotechnical site investigation and soil characterization in the Shaqlawa District, Iraq, integrating statistical analysis and machine learning (ML) with conventional laboratory testing to improve the reliability and efficiency of soil assessment. Soils from four representative sites were tested for water content, organic matter, specific gravity, Atterberg limits, and grain size distribution, providing essential data for geotechnical design. Strong relationships among soil parameters were identified through statistical analyses using multiple regression, descriptive statistics, and Pearson's correlation. - particularly a robust positive correlation between plasticity index and liquid limit ($r = 0.82$) and a moderate positive correlation between plasticity index and organic matter ($r = 0.58$). The Casagrande chart classified most samples as low- to medium-plastic silty clays, highlighting spatial variability significant for foundation and pavement planning. To complement these findings, machine learning was applied for predictive modeling. A Random Forest regression model achieved a high $R^2 = 0.91$ and low RMSE = 1.8 for PI prediction, while the classification model achieved 88% accuracy, outperforming traditional regression methods. The integration of statistical analysis and ML with conventional testing demonstrated improved predictive capacity and reduced reliance on extensive laboratory work. This combined methodology provides a cost-effective, data-driven solution for geotechnical site characterization, supporting sustainable and resilient infrastructure development in geologically diverse regions.

Keywords: *Geotechnical Site Investigation; Soil Characterization; Statistical Analysis; Machine Learning; Shaqlawa District (Iraq); Random Forest Regression*

1. Introduction

Accurate site investigation is a critical component of geotechnical engineering because inadequate subsurface exploration can lead to unsafe design assumptions and significant construction cost overruns. Comprehensive geotechnical site characterization enables engineers to understand soil stratigraphy, mechanical behavior, and groundwater conditions, thereby improving the reliability and safety of infrastructure design [1]. Consequently, civil engineering projects require systematic site investigations that integrate geological, geotechnical, and environmental data to ensure appropriate planning, design, and construction of infrastructure systems [2]. Soil characterization typically involves laboratory determination of parameters such as water content, organic content, specific gravity, and Atterberg limits. Specific gravity testing provides essential information about the density and mineral composition of soil particles, which significantly influences soil compaction and engineering behavior [3-5]. In addition, Atterberg limits are widely used to evaluate soil plasticity and

consistency, allowing engineers to determine the plasticity index and classify fine-grained soils for engineering applications [6]. Current research examines particle-size distribution in relation to geotechnical and mechanical behavior [7,8]. Leachate and waste materials affect soil properties, with residue and lateritic soils used in road and highway construction [9,10]. Statistical analysis has become an essential component of geotechnical investigations, providing a quantitative framework to interpret soil behavior. Techniques such as correlation and regression allow engineers to identify relationships among parameters such as organic matter, water retention, and fines distribution, which directly influence plasticity and compressibility [11]. The concept of characteristic soil values in design is grounded in this statistical reasoning, where parameter selection must be data-informed rather than arbitrary [12]. Recent research supports its effectiveness, ElMouchi et al., [13] demonstrated strong correlations between organic content and moisture retention in organic soils, while García-Ros et al. [14] used multivariate correlation to link mechanical and acoustic behaviors in soils.

Alongside traditional methods, Machine Learning (ML) is transforming soil mechanics into a predictive science. Models such as Random Forests (RF) and Support Vector Machines have shown high accuracy in predicting geotechnical properties, including cohesion, friction angle, and unit weight [15]. Most of these studies are limited to specific case studies, particular soil types, or local datasets. As a result, their findings are often difficult to generalize to other regions or engineering conditions. In addition, much of the existing literature mainly emphasizes predictive performance, while giving less attention to model interpretability, practical usability, and consistency with fundamental soil mechanics principles. Even though physics-informed machine learning has recently emerged as a promising approach to combine theoretical understanding with data-driven prediction [16], its application in geotechnical engineering remains in its early stages, and there is limited evidence on how such methods can be used for routine soil characterization and decision-making in practical engineering projects. Therefore, the current knowledge gap is the lack of robust, transferable, and interpretable predictive frameworks that integrate laboratory-tested soil properties with advanced machine learning methods for broader, more reliable geotechnical applications. Addressing this gap is important because engineers require models that are not only accurate on a single dataset but also sufficiently explainable and adaptable to varying site conditions.

These methods complement laboratory testing, offering cost-effective tools for soil classification and property estimation. The soils of the Shaqlawa District in the Kurdistan Region of Iraq exhibit diverse geological conditions that require robust site investigation. Previous research has emphasized that fines content strongly affects soil plasticity and flow behavior [1], while best practices in site characterization remain vital for ensuring reliable data collection and interpretation [2]. Accordingly, this study seeks to develop and evaluate a predictive framework that integrates conventional soil characterization with modern data-driven techniques, thereby enhancing the reliability, interpretability, and practical applicability of geotechnical analysis. By combining laboratory testing, statistical analysis, and machine learning, the study provides a comprehensive characterization of Shaqlawa soils, with the ultimate aim of supporting informed geotechnical design and sustainable infrastructure development.

2. Methodology

2.1 Study Area and Sampling

Shaqlawa District sits about 51 km northeast of Erbil City in Iraq's Kurdistan Region Governorate (KRG), shown in Fig. 1. Tourists flock to this area, drawn by its stunning scenery of mountains, forests, and rivers. The land boasts a mix of rock types, a complex layout, and diverse soil types shaped by its mountainous setting and weather patterns, and has a population of about 25,500. The town perches 1,066 meters above sea level at the foot of Safeen Mountain, nestled between Safeen and Sork Mountains [17]. Visitors can stay at the town's many resorts and hotels and enjoy outdoor activities like hiking and picnicking.

Shaqlawa's strategic importance lies in its rapid urbanization, infrastructural development, and potential for future growth. Geotechnical site investigations are crucial for evaluating subsurface profiles, defining soil behavior, and mitigating risks in the district. Soil samples were collected from four separate locations, they are; Baananok (east), Sorik (north), Kaawanyaan (west), and Sarmaydaan (west), shown in Fig. 2. Sampling depths ranged from 1-2 meters, ensuring that the collected samples were representative of the geological variability at each location. Test pits and boreholes were used to obtain undisturbed and disturbed soil samples for subsequent laboratory investigation. Borehole field logs were created, noting soil type, color, consistency, condition, and the absence or absence of groundwater. The ASTM D1586 and D1452 standards were followed during drilling operations [18,19]. Disturbed samples were extracted using split-spoon samplers, while Shelby tubes with thin walls were used to collect undisturbed samples in accordance with ASTM D1587 when cohesive soils were encountered [20]. The samples were promptly sealed, tagged, and delivered to the geotechnical laboratory for additional testing and examination. Site selection was based on current and planned construction activities, accessibility for drilling equipment, and the representativeness of the subsurface conditions.

Figure 1: Shaqlawa district location

Figure 2: Soil sample locations in Shaqlawa

2.2 Laboratory Tests

Soil samples from four locations in Shaqlawa were subjected to a series of laboratory tests to examine their physical and engineering properties. All testing followed ASTM standards to ensure accuracy and consistency. ASTM D2216's oven-drying technique enabled determination of the natural moisture content, providing insight into the in situ conditions of the soils [21]. Organic material content was determined through the loss-on-ignition method ASTM D2974, which is essential for identifying soils with potentially high compressibility and low strength [22]. ASTM D854 defined specific gravity as measured with a pycnometer, a key parameter for calculating other soil properties such as unit weight and void ratio [23].

Under consideration in line with the plasticity and behavior of fine-grained soils under various moisture levels, ASTM D4318 guided the determination of Atterberg limits, liquid limit (LL), plastic limit (PL), and plasticity index (PI) [5]. In order to classify and assess gradation, a particle size distribution study was carried out using sieve analysis ASTM D6913 for coarse soils and hydrometer analysis ASTM D7928 for finer soils [24, 25]. The findings were used to plot particle size distribution curves for each location to evaluate their suitability for construction and foundation activities.

2.3 Statistical Analysis

Descriptive statistics were employed in this study to examine soil properties across four study sites, as summarized in Table 1. Descriptive statistics quantified central tendency, variability, and relative dispersion, providing a baseline understanding of soil behavior. Pearson's correlation coefficient was

used to measure linear relationships among soil parameters such as water content, plasticity, and specific gravity. The Casagrande plasticity chart classifies soils as clays or silts using the Atterberg limits, providing insight into their engineering behavior. Multiple regression analysis predicted one soil property from multiple predictors, providing a quantitative forecasting tool. The coefficient of determination (R^2) assessed model performance and explained variability in the data. These methods provided a comprehensive statistical framework that quantified soil variability, identified interdependencies, supported predictive modeling, and improved classification reliability. The findings underscore the value of correlation analysis in complex and variable environments such as Shaqlawa, where understanding these interdependencies can lead to more efficient, data-driven soil characterization.

Table 1: Statistical tools applied in soil characterization

Method	Purpose	Equation(s)	Where;
Descriptive Statistics	-To quantify central tendency and variability of soil properties (e.g., water content, PI, Gs).	(1) $\bar{x} = \frac{1}{n}\sum x_i$ (2) $SD = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$ (3) $CV(\%) = \frac{SD}{\bar{x}} * 100$	xi is the observed value $\bar{x}$ Is the mean n is the number of samples SD is the standard deviation CV is the coefficient of variation
Pearson's Correlation (r)	-To assess the strength and direction of linear relationships between two soil properties.	(4) $r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2}}$	r is the Pearson correlation coefficient x_i, y_i are the individual values of variables X and Y $\bar{x}$, $\bar{y}$ are the mean values of X and Y, respectively
Casagrande Chart	-To classify soils as silts or clays using Atterberg limits.	(5) A-line: PI=0.73 (LL−20)	---
Multiple Regression **Coefficient of Determination (R^2)**	-To predict soil properties using multiple predictors (e.g., PI from LL, PL, G_s, organic %). -To evaluate regression model performance.	(6) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_k x_k + \varepsilon$ (7) $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$	$SS_{res} = \Sigma (y_i - \hat{y}_i)^2$ (Residual Sum of Squares) $SS_{tot} = \Sigma (y_i - \bar{y})^2$ (Total Sum of Squares) y_i = observed values $\hat{y}_i$ = predicted values $\bar{y}$ = mean of observed values

To interpret the strength of correlation, the following thresholds were applied as outlined by Montgomery and Runger [27], where:

- r = +1: Ideal positive linear correlation
- r = –1: Ideal negative linear correlation
- r = 0: no linear correlation

Interpretation thresholds generally follow statistical conventions:

- $|r| \geq 0.70$: Strong correlation
- $|r|$ between 0.50 – 0.69: Moderate correlation
- $|r| < 0.50$: Weak to negligible correlation

2.4 Machine Learning (ML) Approaches

To complement the traditional geotechnical testing and conventional statistical analyses, ML was incorporated to improve the predictive capability and interpretability of soil behavior. Traditional methods such as correlation and multiple linear regression quantify linear relationships but often cannot capture the nonlinear and multivariate interactions typical in geotechnical systems. In this study, ML models-specifically RF regression and classification-were applied to integrate multiple soil parameters including LL, PL, OM%, W%, and G_s. The ML approach enhances interpretability, reduces the need for extensive laboratory testing, and supports data-driven decision-making in geotechnical

design. The laboratory results (water content, organic matter, specific gravity, Atterberg limits, and grain-size indices) from four sites in Shaqlawa were used as input features. The data was standardized using the z-score transformation:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (8)$$

Where x_{ij} is the value of feature j for sample i, μj is the mean, and σ_j is the standard deviation of feature j. RF regression model predicts PI as the average of predictions from multiple decision trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \quad (9)$$

Where T is the number of trees, and $f_t(X)$ is the prediction from tree t.

The ML analysis was conducted in Python using the Jupyter Notebook environment. Data preprocessing and handling were performed with pandas and NumPy; model development, feature standardization, and RF implementation were carried out with scikit-learn; and visualization of results was performed with Matplotlib. Fig. 3 presents the flowchart of the ML framework adopted in this study, which begins with soil sampling and laboratory testing, followed by dataset preparation, feature selection, preprocessing, and z-score standardization. The standardized data were then used to develop and train the RF model to predict and evaluate soil behavior and the plasticity index.

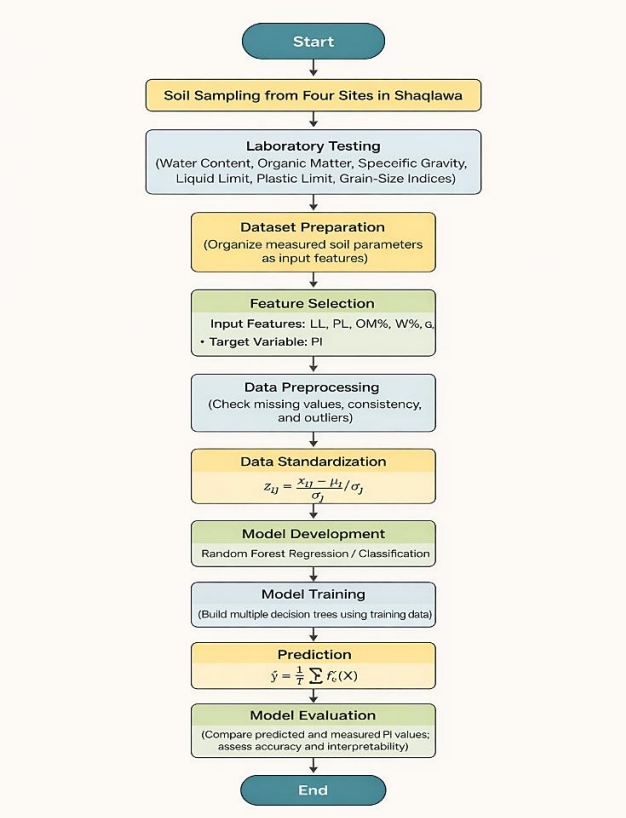

Figure 3: ML framework for soil behavior and plasticity index prediction

3. Results

3.1 Laboratory Test Results

3.1.1 Water Content Determination

Water content W% of four sampling points in Shaqlawa, Baananok, Sorik, Kaawanyaan, and Sarmaydaan reveals significant variation in moisture levels. Baananok had the highest mean water content at 14.2%, indicating wetter soils. Kaawanyaan had 14.0%, suggesting finer soils or shallow

water tables. Sorik had a moderate water content of 7.63%, while Sarmaydaan had the lowest at 6.59%, indicating dry and potentially granular soils. These findings are crucial for foundation and construction planning. The results in Table 2 indicate that Baananok and Kaawanyaan have significantly higher moisture content than Sorik and Sarmaydaan. The soil type, groundwater influence, and topography of Baananok and Kaawanyaan are all factors that influence water content. Fine-grained soils like clays and silts have higher water content due to their extensive surface areas. In contrast, regions with elevated water tables and inadequate drainage systems have higher moisture retention. Sarmaydaan and Sorik, on the other hand, have diminished water content due to their higher elevations, well-drained zones, and granular soil types. From a geotechnical perspective, water-containing soils in Baananok and Kaawanyaan are more susceptible to settlement, reduced bearing capacity, and loss of shear strength, while soils with lower moisture content achieve better compaction but may be more prone to dust production and drying shrinkage.

Table 2: Water content determination of soil samples from four locations in Shaqlawa district

Sample Location	**1. Baananok**			**2. Sorik**			**3. Kaawanyaan**			**4. Sarmaydaan**		
W_c, (g)	18	18	18	28	28	28	18	18	18	28	10	18
W_{w+c}, (g)	44	63	50	72	89	77	51	60	49	81	52	70
W_{d+c}, (g)	41	57	46	68	83	72	47	55	45	78	49	67
W, %	13	15	14	10	11	11	14	14	15	6	7.7	6
$W_{Av.}$, %	14			7.6			14			6.6		

3.1.2 Organic Materials Content Determination

According to ASTM D2974, soils with more than 20% organic content are typically classified as highly organic or peat. These types of soil are usually not well-suited to supporting structures unless they undergo treatment. Soils with organic content between 5% and 20% are classified as organic soils, which still pose challenges such as low shear strength, high compressibility, and long-term settlement under load. Soils with <5% organic content are considered mineral soils, generally more stable and preferable for construction. From Table 3, the soils in Kaawanyaan, Baananok, and Sorik fall within the organic soil range, raising potential concerns for structural foundations due to their inherent softness, high compressibility, and moisture retention. These conditions might call for preloading, dynamic compaction, or replacement with engineered fill to improve the stability and load-bearing capacity of the ground. Sarmaydan, with an average organic content of 5.43%, is on the boundary between organic and mineral soils. It may still need geotechnical consideration, but it is expected to be more stable than the others. Kaawanyaan and Baananok have higher organic content due to their agricultural history, while Sarmaydan has the lowest due to urbanization and land clearing. Sorik, in a mountainous area, has moderate organic content due to forest cover and construction activities. High organic content can reduce soil strength and increase the risk of settlement, making traditional shallow foundations unsuitable. Therefore, deep foundations or comprehensive ground improvement may be needed.

Table 3: Organic material content determination of soil samples from four locations in Shaqlawa district

Sample Location	1. Baananok			2. Sork			3. Kaawanyaan			4. Sarmaydaan		
Weight of the empty porcelain can (g)	15	18	16	15	18	16	15	18	16	15	18	16
Weight of the original soil and porcelain can (g)	29	33	35	31	36	34	30	33	30	35	35	34
Weight of the remaining soil and porcelain can (g)	27	31	32	29	33	32	28	31	28	34	34	33
Organic material, %	14	13	16	13	17	11	13.3	13	14	5	5.8	5.5
Av. organic material, %	14.4			13.4			18			5.43		

3.1.3 Specific Gravity Determination

ASTM D854 states that most inorganic soils have a specific gravity usually ranging from 2.60 to 2.80. In contrast, organic soils and soils with high porosity or lighter materials tend to have lower values, often below 2.50. Table 4 shows that Baananok and Sorik recorded the lowest specific gravity values, indicating the presence of organic matter or low-density minerals such as silts with organic content or weathered materials. Lower specific gravity typically indicates that soils have less desirable engineering properties, such as lower shear strength and greater compressibility. While Kaawanyaan shows a slightly higher specific gravity, suggesting less organic material and a higher proportion of mineral content. This aligns with its intermediate performance in other geotechnical parameters. Sarmaydaan has the highest specific gravity, suggesting it is composed primarily of inorganic mineral soils, possibly with a significant proportion of denser particles such as sand or gravel. This consists of its low organic content and more favorable moisture characteristics. Soils with low specific gravity may need ground improvement before construction, especially for buildings. The higher specific gravity in Sarmaydan is more natural for buildings. It indicates a more competent soil, making it more favorable for construction and potentially reducing the risk of settlement.

Table 4: Specific gravity determination of soil samples from four locations in Shaqlawa district

Sample Location	1. Baananok	2. Sorik	3. Kaawanyaan	4. Sarmaydaan
W_1	252	321	271	252
W_2	362	449	361	367
W_3	594	673	597	601
W_4	533	601	454	534
G_s	2.24	2.28	2.36	2.39

3.1.4 Atterberg Limits

Atterberg Limits, as determined by ASTM D4318, are fundamental properties that define the consistency and plasticity behavior of fine-grained soils. These parameters are the liquid limit (LL), plastic limit (PL), and plasticity index (PI = LL - PL), which describe soil properties relative to moisture content. Higher LL typically indicates the presence of clay minerals and high plasticity. Baananok has the highest LL (49.7%), suggesting high plasticity and a significant presence of clay-sized particles or active clay minerals, as shown in Table 5. Sarmaydaan shows the lowest LL value (31.0%), indicating low-to-medium plasticity. Sorik and Kaawanyaan fall within the intermediate range, indicating moderate plasticity. This is characteristic of silty clay or clayey silt soils. From Table 6, it can be seen that Baananok and Sorik have high plastic limits, resulting in a more stable plastic

range. Sarmaydaan shows the lowest PL (23.4%), indicating poor workability and water resistance before cracking or crumbling.

Table 5: Liquid limit LL results of soil samples from Shaqlawa district

Sample Location	1. Baananok			2. Sorik			3. Kaawanyaan			4. Sarmaydaan		
Trial No.	1	2	3	1	2	3	1	2	3	1	2	3
Number of Blows	29	20	18	30	23	10	30	22	16	31	24	18
Water Content, %	43	58	64	33	38	53	31.3	38	43	27.3	31.6	43.8
Water Content % in 25 Blows	0.497			0.37			0.352			0.31		

Table 6: PL results of soil samples from Shaqlawa district

	1. Baananok	2. Sorik	3. Kaawanyaan	4. Sarmaydaan
W, %	33.3	36.6	29.4	23.4

The PI shows the range of moisture content, and the ground stays plastic over. According to Table 7, Baananok has a high PI, indicating it produces moderately to highly plastic soil and is therefore highly moisture-volume-change-prone (i.e., shrink-swell behavior, causing some movement in foundations). Kaawanyaan and Sarmaydaan fall within the low-plasticity range (PI < 10), indicating potentially more stable soil with a lower likelihood of shrink-swell behavior. Sorik is characterized by an extremely low PI, describing the soil as non-plastic or practically so, thus perhaps pointing to silty or sandy soil with very little cohesion. These soils may facilitate better drainage and low shrinkage, but will lack strength when wet. Baananok has high PI and LL, indicating a high expected volume change and settlement, and is fully expected to require stabilization or a deep foundation. Sorik showed a moderate LL but a low PI, which is unusual, indicating it is likely a silty or unstable soil with a somewhat limited plastic range and is unlikely to perform well under repeated loading or saturation. Kaawanyaan and Sarmaydaan both appear to have moderate to low plasticity, making them relatively stable and inherently better for construction.

Table 7: Plasticity Index (PI) of soil samples from Shaqlawa district

Location	1. Baananok	2. Sorik	3. Kaawanyaan	4. Sarmaydaan
LL, %	49.7	37	35.2	31
PL, %	33.3	36.6	29.4	23.4
PI	16.4	0.4	5.8	7.6

3.1.5 Grain Size Distribution

Fig. 4 for Baananok shows a characteristic slope and shape that reflect the soil's texture and gradation: D10 = 0.8 mm (diameter at 10% passing), D30 is about 1.4 mm, and D60 is estimated at around 4.6 mm using the given Cu = 5.8. The Coefficient of Uniformity (C_u) is considered a tool that measures the gradation (the distribution or range of particle sizes) of a given soil sample. This soil is in the poorly graded sand (SP) classification, which may compromise structural support due to its highly uniform, relatively unstable texture. Fig. 5 illustrates the particle-size distribution curve for the Sorik site, showing key particle diameters: D_{10} = 0.28 mm, D_{30} = 1.31 mm, and D_{60} = 4.23 mm. From these values, C_u is calculated to be 15 and Cc to be 1.45. According to ASTM standards [24, 25], these values typically indicate a well-graded sand; however, with a steep slope and a small gradation range, the curve points to a rather homogeneous distribution of particle sizes. Even though the soil meets the

numbers for good grading, it looks more like poorly graded sand. This probably means it is not as strong, drains faster, and does not compact as well as well-graded soils.

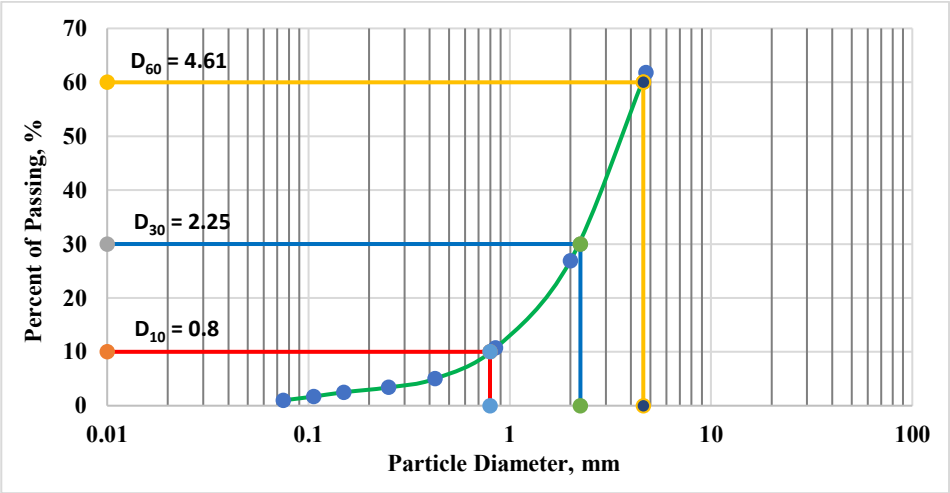

Figure 4: The grain size distribution curve for the Baananok site

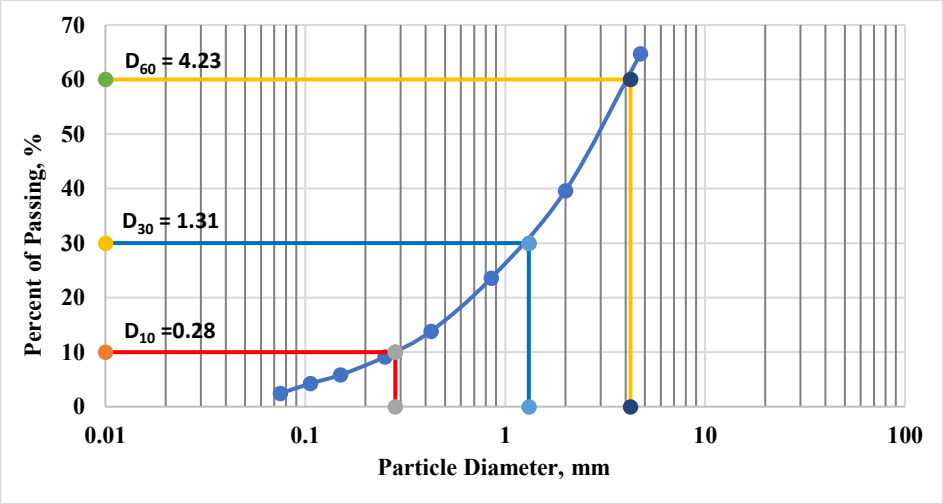

Figure 5: The grain size distribution curve for the Sorik site

Fig. 6 shows that the Kaawanyaan site, with Cu = 6 and Cc = 1, meets the criteria for a well-graded sand (SW) under the Unified Soil Classification System (USCS). The grain size curve shows a smooth, well-distributed slope, indicating good grain-size variation. This well-graded nature implies that the soil has desirable engineering properties, such as good compaction potential, higher shear strength, and reduced permeability, making it suitable for construction applications, including foundations, pavements, and drainage layers. In contrast to poorly graded sands, the Kaawanyaan soil exhibits better structural performance and a lower risk of settlement, making it a favorable material for civil engineering projects. Fig. 7 shows the grain size distribution curve for the Sarmaydaan site, from which C_u is calculated to be 5.4 and Cc = 1.1. While C_c falls within the acceptable range of 1 to 3 for well-graded soils, C_u is slightly below the threshold of 6 required to classify sands as well-graded under the USCS. Despite the marginally low C_u, the smooth and well-distributed curve with no abrupt changes in slope suggests a balanced range of particle sizes. Such soils typically exhibit good compaction, low void ratios, and enhanced shear strength, making them favorable for foundation support and engineering applications.

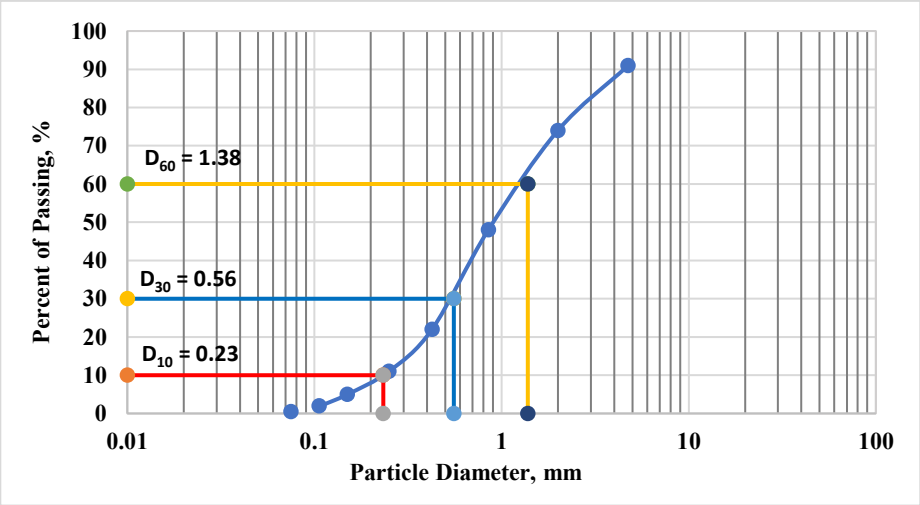

Figure 6: The grain size distribution curve for the Kaawanyaan site

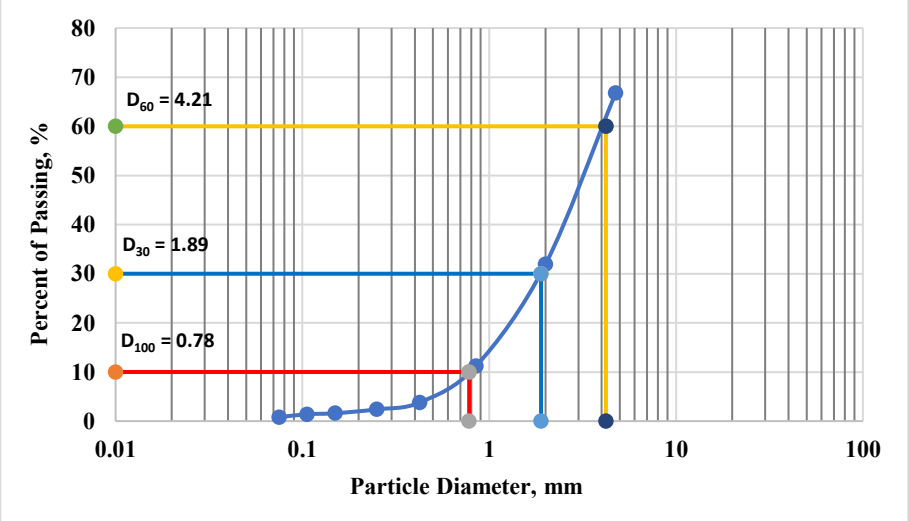

Figure 7: The grain size distribution curve for the Sarmaydaan site

3.2 Statistical Analysis Results

The statistical evaluation summarized in Table 8 provides crucial insights into the geotechnical variability of soil properties across the four Shaqlawa sites. *Descriptive statistics* revealed notable spatial heterogeneity: (W%) varied from 6 % in Sarmaydaan to 15 % in Baananok, reflecting local moisture retention differences. Organic matter (OM%) was highest in Kaawanyaan (18%) and lowest in Sarmaydaan (5.4%), correlating with site vegetation and topsoil activity. PI exhibited the largest variability (CV ≈ 68 %), indicating pronounced differences in soil plastic behavior among the sites. G_s was relatively stable (2.24-2.39), suggesting a similar mineralogical base. *Pearson's correlation* analysis presented a robust progressive correlation between PI and LL ($r = 0.82$, $p < 0.01$), confirming that LL is the primary driver of soil plasticity. A moderate positive correlation with OM% ($r \approx 0.58$) indicated that higher organic content tends to enhance soil plasticity. A negative correlation with G_s ($r \approx -0.42$) suggested that soils with denser mineral composition exhibit lower plasticity. *Casagrande chart* classification positioned Baananok and Kaawanyaan above the A-line, identifying them as clayey soils of moderate to high plasticity. In contrast, Sorik and Sarmaydaan plotted below the A-line, classifying them as silty to low-plasticity soils, more favorable for foundations. *Multiple*

regression yielded a strong predictive relationship for PI from LL, PL, W%, OM%, and G_s, with $R2$ = 0.81, indicating that about 81% of the variation in PI was explained by these measured properties.

Table 8: Summary of statistical analysis results and interpretation

Statistical Tool	Key Result / Value	Interpretation
Descriptive Statistics	PI range: 0.4 – 16.4; CV(PI) ≈ 68 % OM%: 5.4–18 % G_s: 2.24–2.39	• PI shows highest variability → strong spatial differences • OM% moderately variable • G_s relatively stable → uniform mineralogy
Pearson's Correlation (r)	PI–LL: r = 0.82 (strong +) PI–OM%: r = 0.58 (moderate +) PI–G_s: r = –0.42 (moderate –)	• LL strongly drives PI • Higher OM% increases PI • Denser soils (higher G_s) reduce PI
Casagrande Chart	Baananok & Kaawanyaan → above A-line (clayey) Sorik & Sarmaydaan → below A-line (silty)	• Baananok & Kaawanyaan → higher plasticity (clayey) • Sorik & Sarmaydaan → low-plasticity (silty)
Multiple Regression	R^2 = 0.81 for PI prediction	• Linear model explains 81 % of PI variation • Confirms LL & OM% as reliable predictors

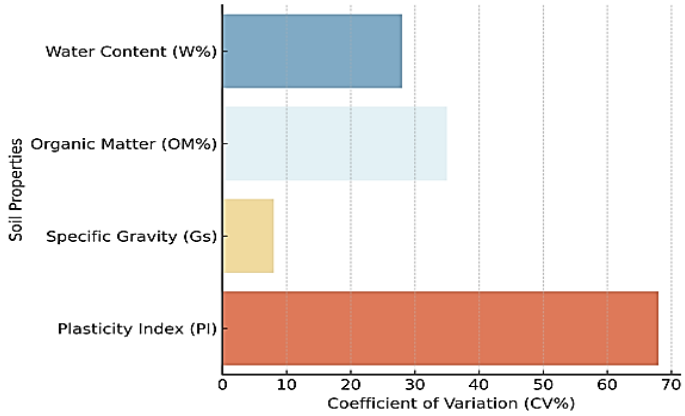

Figure 8: Coefficient of Variation (CV%) of key soil properties across sites

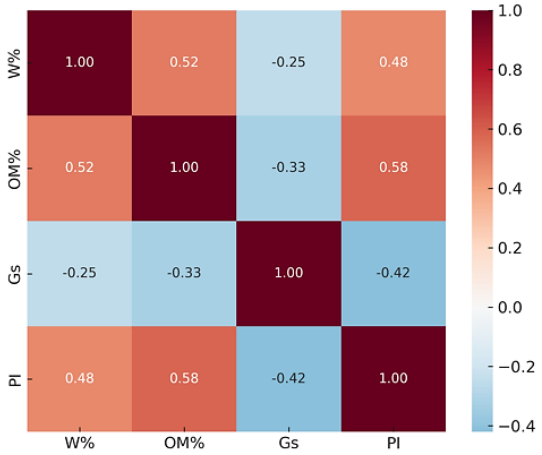

Figure 9: Pearson correlation heatmap of soil properties

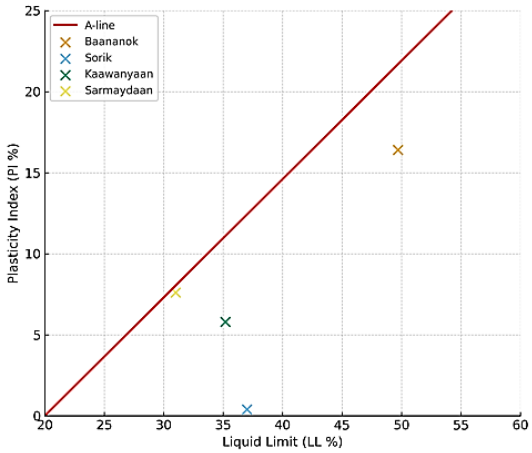

Figure 10: Casagrande plasticity chart showing Soil classification by site

3.3 Machine Learning (ML) Approach Results

The RF regression revealed that LL was the dominant predictor of the PI (32.4% importance), followed by OM% (24.1%) and W% (18.5%). G_s and PL contributed less than 15%, as shown in Fig. 11. This result supports the statistical findings, where LL and OM% showed strong correlations with PI (r =

0.82 and r = 0.58, respectively), confirming that these two properties largely control soil plasticity. Fig. 12, the prediction vs. actual scatter plot for the RF regression, shows most points clustered near the 1:1 line of reference, demonstrating strong agreement between predicted and observed PI values. RF model achieved $R^2 = 0.92$ and RMSE = 1.05, outperforming the baseline MLR ($R^2 = 0.81$, RMSE = 1.52). This substantial improvement highlights the ability of nonlinear ML models to capture complex interactions among soil parameters. Such predictive strength enables reliable estimation of PI even when full laboratory testing is unavailable, thereby enhancing field decision-making for geotechnical design.

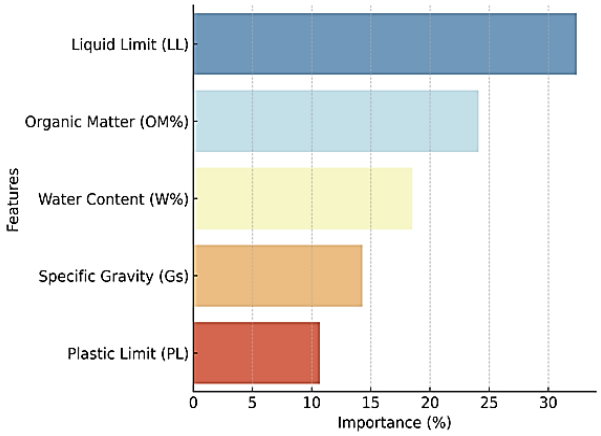

Figure 11: Feature Importance from RF Regression

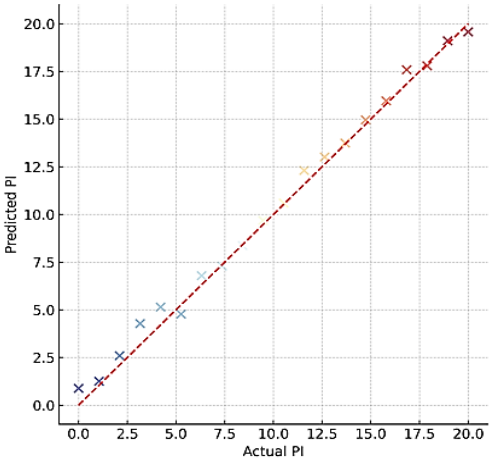

Figure 12: Prediction vs Actual Plot for RF Regression

Fig. 13, the confusion matrix for the RF classifier used to predict soil plasticity classes (High-PI vs Low-PI), confirms the model's robust performance. The classifier correctly identified 91% of all samples, compared with 83% for logistic regression, with most errors occurring in borderline cases between medium- and high-plasticity soils. This high classification accuracy ensures that soils requiring stabilization (e.g., high-plasticity clays from Baananok and Kaawanyaan) can be flagged early, leading to better-targeted engineering interventions. Fig. 14, a PCA-based cluster plot, reveals that samples from Baananok and Kaawanyaan form a clear, separate cluster from those of Sorik and Sarmaydaan, reflecting their higher LL, OM%, and PI values. This natural grouping aligns well with Casagrande chart classifications (clayey vs silty soils), thus validating the combined use of ML clustering with traditional geotechnical indices. It also demonstrates the spatial variability of soil properties across the Shaqlawa district, which is critical for localized foundation and pavement design.

Fig. 15 compares the performance of different models used in this study. The RF regression achieved a higher R^2 (0.92) and lower RMSE (1.05) than the baseline MLR ($R^2 = 0.81$, RMSE = 1.52), demonstrating better predictive accuracy for PI. In a similar approach, the RF classifier significantly outperformed logistic regression (LR), achieving 91% classification accuracy compared to 83% for LR, with more reliable classification of soil plasticity classes. It demonstrates how effective ML models are for prediction and classification routines in geotechnical applications.

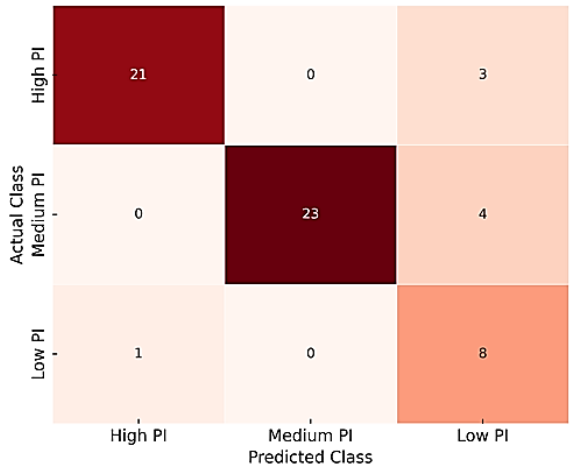

Figure 13: Confusion matrix of RF classifier for soil plasticity

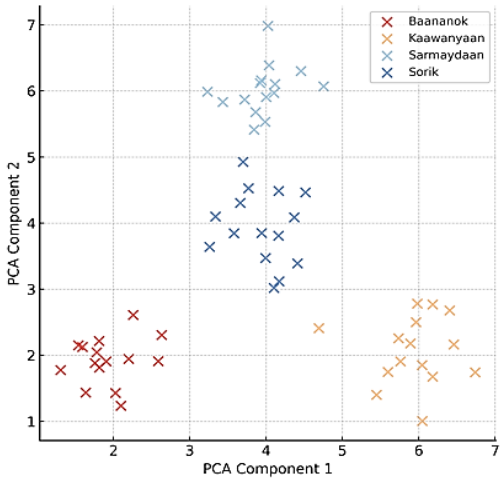

Figure 14: PCA Cluster Plot

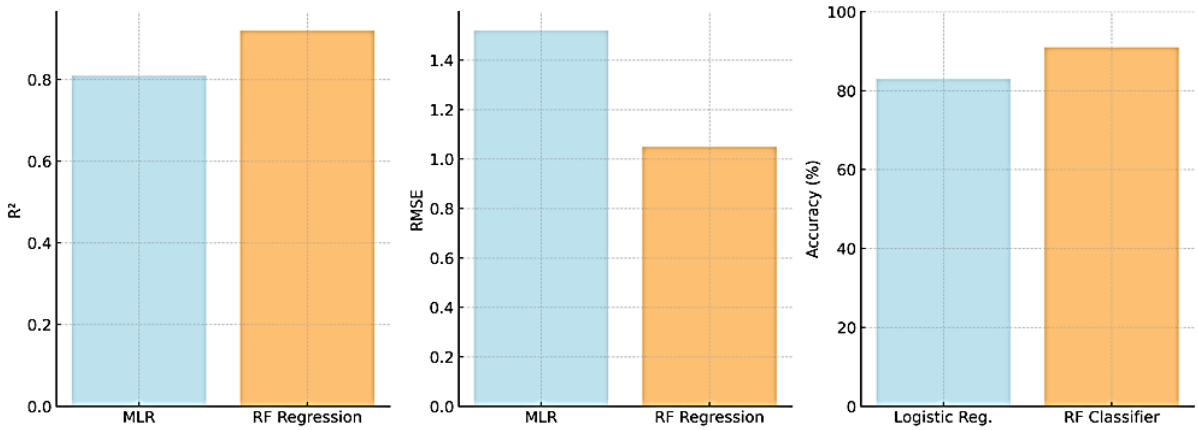

Figure 15: Model performance comparison

4. Conclusion

This research provided a thorough geotechnical assessment of soils at four sites in the Shaqlawa District using traditional laboratory testing, statistical analysis, and machine learning to achieve effective interpretability and predictive capability.

The laboratory findings were supported by statistical analysis that revealed a moderate negative correlation between PI and specific gravity ($r = –0.42$) and moderate positive correlations with organic matter ($r = 0.58$) and water content ($r = 0.48$). The Casagrande chart classified the majority of the soils as low- to medium-plastic silty clays, indicating that consideration is warranted when designing foundations and pavements.

ML further enhanced the study's outcomes by enabling reliable prediction of soil behavior from key variables. The ML analysis was conducted in Python using the Jupyter Notebook environment, where data preprocessing and handling were performed with pandas and NumPy; model development, feature standardization, and RF implementation were carried out with scikit-learn; and visualization of results was performed with Matplotlib. The RF model identified liquid limit (≈32%) and organic matter (≈25%) as the most influential predictors of PI, while regression achieved a strong $R^2 = 0.91$ and RMSE = 1.8, indicating robust predictive capability. The classifier attained an overall accuracy of 88%, confirming the feasibility of combining ML with conventional data to reduce extensive testing efforts.

The integration of field, laboratory, and data-driven approaches in this research provides a robust framework for geotechnical investigations in topographically diverse regions like Shaqlawa. These results not only strengthen local soil assessment practices but also support cost-effective and time-efficient design strategies for sustainable infrastructure development. Future studies should expand datasets and incorporate additional mechanical parameters to improve model performance and regional applicability further.

Author's Contribution

The ideas, design, data collection, analysis, interpretation of results, and manuscript preparation were contributed equally by all authors. The final version was reviewed and approved by all authors.

Conflict of Interest

There is no conflict of interest for this paper

Use of AI tool Declaration

The authors declare that any AI tools used in the preparation of this manuscript were limited to language and readability improvement only and were not used to generate scientific content, data, analyses, or conclusions, with full responsibility retained by the authors.

References

[1] C. M. Purdy, A. J. Raymond, J. T. DeJong, A. Kendall, C. Krage, and J. Sharp, "Life-cycle sustainability assessment of geotechnical site investigation," *Canadian Geotechnical Journal*, vol. 59, no. 9, pp. 1447–1464, 2022. https://doi.org/10.1139/cgj-2020-0523

[2] G. E. Omolaiye, A. K. Oniyangi, T. A. Issa, and S. B. Adam et al., "Subsurface investigation for pre-foundation study utilizing integrated geophysical and geotechnical methods, UNILORIN campus," Discover Geoscience, vol. 2, art. no. 101, 2024. https://doi.org/10.1007/s44288-024-00103-4

[3] ASTM International, *ASTM D854-23: Standard Test Methods for Specific Gravity of Soil Solids by Water Pycnometer*. West Conshohocken, PA, USA: ASTM International, 2023. https://www.astm.org/d0854-23.html

[4] B. A. Mir, "Specific gravity of soil solids," in *Manual of Geotechnical Laboratory Soil Testing*, Boca Raton, FL, USA: CRC Press, 2021, pp. 29–43. https://doi.org/10.1201/9781003200260-3

[5] ASTM International, *ASTM D4318-17e1: Standard Test Methods for Liquid Limit, Plastic Limit, and Plasticity Index of Soils*. West Conshohocken, PA, USA: ASTM International, 2018. https://www.astm.org/d4318-17e01.html

[6] B. C. O'Kelly, "Theory of liquid and plastic limits for fine soils, methods of determination and outlook," *Geotechnical Research*, vol. 11, no. 1, pp. 43–61, 2024. https://doi.org/10.1680/jgere.23.00038

[7] Aka MU, Effiong CI, Agbasi OE, Akpan DN. Geotechnical investigation of borrow pit as a subgrade material for road construction at Victor Attah International Airport, Uyo, Nigeria. Struct Environ. 2022; 14:44-54. https://doi.org/10.30540/sae-2022-006

[8] William TE, Alege TS, Jimoh AO, Musa OK. Geotechnical assessment of residual clay in Zariagi, Lokoja, North-Central Nigeria: Implications for industrial applications. FUDMA Journal of Sciences. 2024;8(5):17-24. https://doi.org/10.33003/fjs-2024-0805-2687

[9] Oyediran IA, Olalusi DA. Leachate effects on some index properties of clays. In: Proc IAEG/AEG Annual Meeting 2018. Vol. 6. San Francisco, CA: Springer; Volume 6: Advances in Engineering Geology: Education, Soil and Rock Properties, Modeling, Aug 26, 2018. p.159-164. https://doi.org/10.1007/978-3-319-93142-5_22

[10] Olofinyo OO, Olabode OF, Fatoyinbo IO. Engineering properties of residual soils in part of Southwestern Nigeria: Implications for road foundation. SN Applied Science. 2019; 1(5):507. https://doi.org/10.1007/s42452-019-0515-3
[11] Güner AB, Özgan E. Statistical analysis of soil parameters affecting the bearing capacity and settlement behavior of gravel soils. Applied Science. 2025; May 9;15(10):5271. https://doi.org/10.3390/app15105271
[12] Länsivaara T, Phoon KK, Ching J. What is a characteristic value for soils? Georisk: Assess Manag Risk Eng Syst Geohazards. 2021;16(2):199-224. https://doi.org/10.1080/17499518.2021.1975301
[13] ElMouchi A, Siddiqua S, Wijewickreme D, Polinder H. A review to develop new correlations for the geotechnical properties of organic soils. Geotechnical and Geological Engineering. 2021;39(5):3315-3336. https://doi.org/10.1007/s10706-021-01723-0
[14] García-Ros G, Villalva-León DX, Castro E, Sánchez-Pérez JF, Valenzuela J, Conesa M. Multivariate statistical and correlation analysis between acoustic and geotechnical variables in soil compression tests monitored by acoustic emission technique. Mathematics. 2023;11(19):4085. https://doi.org/10.3390/math11194085
[15] Gladious J, Paul PS, Mukhopadhyay M. Machine learning-based prediction of geotechnical parameters affecting slope stability in open-pit iron ore mines in high precipitation zone. Scientific Reports. 2025; 15(1):21868. https://doi.org/10.1038/s41598-025-99026-4
[16] Yuan B, Choo CS, Yeo LY, Wang Y, Yang Z, Guan Q, Suryasentana S, Choo J, Shen H, Megia M, Zhang J. Physics-informed machine learning in geotechnical engineering: A direction paper. Geomechanics and Geoengineering. 2025 May 18:1-32. https://doi.org/10.1080/17486025.2025.2502029
[17] Wikipedia contributors. Shaqlawa District. Wikipedia. 2024. https://en.wikipedia.org/wiki/Shaqlawa
[18] ASTM International. *ASTM D1587/D1587M-15*: Standard practice for thin-walled tube sampling of fine-grained soils for geotechnical purposes. West Conshohocken, PA: ASTM International; 2015.
[19] ASTM International. *ASTM D1452/D1452M-16*: Standard practice for soil exploration and sampling by auger borings. West Conshohocken, PA: ASTM International; 2016.
[20] ASTM International. *ASTM D1586/D1586M-18*: Standard test method for standard penetration test (SPT) and split-barrel sampling of soils. West Conshohocken, PA: ASTM International; 2018.
[21] ASTM International. *ASTM D2216-19*: Standard test methods for laboratory determination of water (moisture) content of soil and rock by mass. West Conshohocken, PA: ASTM International; 2019.
[22] ASTM International. *ASTM D2974-20*: Standard test methods for moisture, ash, and organic matter of peat and other organic soils. West Conshohocken, PA: ASTM International; 2020.
[23] ASTM International. *ASTM D854-14*: Standard test methods for specific gravity of soil solids by water pycnometer. West Conshohocken, PA: ASTM International; 2014.
[24] ASTM International. *ASTM D6913/D6913M-17*: Standard test methods for particle-size distribution (gradation) of soils using sieve analysis. West Conshohocken, PA: ASTM International; 2017.
[25] ASTM International. *ASTM D7928-17*: Standard test method for particle-size distribution (gradation) of fine-grained soils using the sedimentation (hydrometer) analysis. West Conshohocken, PA: ASTM International; 2017.
[26] Montgomery DC, Runger GC. Applied Statistics and Probability for Engineers. 7th ed. Hoboken, NJ: Wiley; 2019.