Received:22 September 2025 | Revised: 17 October 2025 | Accepted: 25 November 2025 | Vol.1 | Issue 1| December 2025
ISSN: Pending

Civil Engineering Science and Technology Journal

Predictive Modeling of Soil Compaction Parameters Using Multiple Linear Regression and Support Vector Machines

Jinan Abdulkareem1, Ammar Salman Dawood2, Ihsan Al-abboodi 3,
Civil Department, College of engineering, University of Basrah, Basrah, Iraq 1,2
, Department of Civil Engineering, Shatt Al-Arab University College, Basrah, Iraq 3

Correspondence
Jinan Abdulkareem
Civil engineering department, College of Engineering, Basrah University
Email: pgs.jinan.ali@uobasrah.edu.iq

Abstract

Field dry density is a soil compaction characteristic that is useful for geotechnical engineering design. Laboratory methods are laborious and time-consuming. The purpose of this paper is to determine and compare the effectiveness of MLR and SVM models in predicting this essential parameter from the fundamental soil property index level. A dataset of 86 soil samples with various geotechnical qualities was used, containing data such as gravel, sand, fines, liquid limit, and plastic limit. The dataset was split into 80% training and 20% testing. Using R^2, RMSE, and MSE, the performance of the built MLR and SVM prediction models was thoroughly examined. With an R2 value of 0.988 (on the test set), the SVM model outperforms the MLR model in terms of prediction accuracy for FDD (R2 = 0.814). Compaction behavior and soil property index properties have a complicated relationship, as seen by the performance gap. According to feature importance analysis, the SVM model's predictions heavily relied on the fines content. According to this study, SVM is a useful method that geotechnical engineers can employ to quickly and affordably estimate compaction parameters in the early phases of site investigations and design optimization.

Keywords

Field Dry Density (FDD), Soil compaction, Support Vector Machine (SVM), Multiple Linear Regression (MLR), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Soil index properties, Geotechnical engineering

1. INTRODUCTION

Compaction of soil is one of the main operations in civil and geotechnical engineering and is considered one of the oldest and most successful methods in ground improvement. It is defined as "a process by which soil is compacted by the imparting of mechanical energy to the soil, thus reducing the air voids present in the soil mass and bringing the soil particles close to one another [1]. The ultimate goals of soil compaction are to modify the engineering properties of soil in such a way that it will satisfy the requirements of an intended engineered use [2]. These improvements consist of improving the shear strength and bearing capacity of the foundations, moderating the unwanted settlement of the structures, restricting the volumetric changes, diminishing the permeability of the soil (hydraulic conductivity), and improving the stability of slopes and embankments [4, 3]. It is not only a procedure but also an important decision-making process on the long-term performance, sustainability, safety, and durability of the construction, like highways, railway tracks, airport pavements, earth dams, and landfills. Poor compaction can have catastrophic consequences that include excess settlement, structural failure, and slope stability, and these failures result in high maintenance costs and pose serious threats to safety [5]. There are two principal indexes that evaluate the efficiency of the compaction process: Maximum Dry Density (MDD) and Optimum Moisture Content (OMC). MDD is the highest density that a soil can attain under a given compactive effort, and OMC refers to such water content at which this maximum density could be obtained [3]. Inherent with these factors are the interrelationships among the above-mentioned variables; the soil is stiffer at moisture contents less than the OMC, and interparticle friction opposes densification. As water is added, it acts as a lubricant, allowing particles to slide over one another and pack more tightly. Beyond the OMC, however, excess water begins to occupy space that could otherwise be filled by soil solids, leading to a decrease in dry density. Therefore, MDD and OMC are not just abstract values but critical indicators of a soil's ability to support structural loads and its overall stability. They serve as the benchmark for quality control in the field, where engineers strive to achieve a specified percentage of the laboratory-determined MDD within a narrow range of the OMC to ensure the constructed earthwork performs as designed [6]. Achieving these parameters is essential for the longevity and safety of civil engineering projects, as deviations can lead to unforeseen settlements or failures. Consequently, rigorous testing and monitoring during construction become paramount to maintain the integrity of the soil structure. The conventional method for determining MDD and OMC in the laboratory is the Proctor Compaction Test, first developed by R.R. Proctor in the 1930s. The test exists in two primary forms: the Standard Proctor Test [7] and the Modified Proctor Test [8], with the latter applying a higher compactive effort to simulate heavier modern compaction equipment [18]. The process consists of ramming a soil sample at different moisture contents into a mold of a known volume with a standard hammer and a specific number of blows. The dry densities and their respective moisture contents are plotted to prepare the compaction curve of a soil, from which maximum dry density (MDD) and optimum moisture content (OMC) may be determined [10]. Although the Proctor test has been widely adopted and practiced for decades, it is plagued by serious practical problems. For a start, the test is notoriously hard work and time-consuming. A single test involved the preparation of several samples at various moisture contents, compacting them, and weighing and oven-drying the compacted soil samples, a process that is observed to take more than 24 hours for one soil type [11]. Highway projects involving multiple soil types from multiple borrow sites, for example, can require significant aggregate time and effort and tend to "tie up" project schedules. Second, the test is resource-demanding; skilled laboratory personnel are needed, and so is a large amount of soil per test point. In such situations, under some test conditions, such as preliminary site investigations or situations of material scarcity, it is quite difficult or impractical to get the large quantity of sample needed for the test [12]. Finally, the costs suffered, which can be considerable (such as the cost of workers, machines, and laboratory time), especially for small- to medium-sized projects and for the very limited budget. These limitations create a strong demand for other approaches that could offer an easy, fast, reliable, and economical estimation of soil compactness parameters, in order to meet the requirements of design and construction [11, 12].

Due to the obvious constraints of laboratory testing, the geotechnical community has long desired to predict compaction parameters from readily available soil characteristics. The earliest efforts were concerned with modelling empirical-study results and statistical relations primarily through regression analysis. Those models were intended to connect MDD and OMC to basic properties of soil index, which are simple and low-cost to measure. These characteristics are Atterberg limits (liquid and plastic limits), which indicate the soil behavior to moisture content, and grain-size parameters, gravel, sand, and fines (silt and clays) percentiles [14, 13]. A rich body of literature documents this historical progression. Early researchers like [23] and [15] proposed correlations using parameters like specific gravity, shrinkage limit, and liquid limit [18, 21]. Subsequent studies refined these approaches; for instance, [16] incorporated compaction energy into their predictive equations, while others developed multiple linear regression (MLR) models that combined several index properties to improve accuracy [18]. However, these traditional regression models, while useful for preliminary estimates, suffer from fundamental limitations. They are often based on linear assumptions, yet the relationships governing soil behavior are known to be highly complex and nonlinear. Furthermore, many of these empirical equations are site-specific or soil-type-specific, exhibiting low correlation coefficients (R^2) and poor generalization ability when applied to datasets outside of their development context [4, 24]. The advent of artificial intelligence (AI) and, more specifically, machine learning (ML), has marked a paradigm shift in geotechnical predictive modeling. Unlike traditional statistical methods that require predefined mathematical relationships, ML algorithms can learn complex, nonlinear patterns directly from data. The interest in implementing machine learning in geotechnical engineering springs from the fact that materials react to different factors. For instance, a number of ML algorithms have been utilized in geotechnical disciplines for the previous two decades; the machines predicted performance is significantly better than traditional methods. [21], as well as [22], used artificial neural networks to describe complicated interactions between soil variables and compaction parameters. Moreover, more ensemble construction methods, such as Random Forest (RF) and Extreme Gradient Boosting (XG Boost), which merge several weak learners to form a powerful predictive model, are performing extraordinarily well in predicting undrained shear strength and ground settlement [4, 17].Additionally, Genetic Expression Programming (GEP) and Support Vector Machines (SVM) have been successfully used to elucidate the compaction characteristics of different soil types, even the problematic expansive soils [19, 18].

1.1 Literature Review on MLR and SVM for Geotechnical Prediction.

In the field of machine learning, Multiple Linear Regression (MLR) and Support Vector Machines (SVM) can be seen as applying two different paradigms, thus providing an excellent comparison of their respective strengths and weaknesses in terms of interpretability and predictive power. MLR is a robust statistical baseline that generalizes simple linear regression in a way that it can be used to understand a dependent variable aesthetically with the help of several independent variables. The application of MLR in geotechnical engineering to create easy-to-use empirical models for predicting soil properties is very common. For example, [3] applied MLR along with other models to predict compaction characteristics, which denote its high outcome of transparent equations that could be field applied easily. Likewise, [18] have shown in their literature survey that the multi-linear regression method was often used to predict OMC and MDD through the combined effects of the liquid limit and plastic limit. Nevertheless, MLR, which is greatly cherished for its ease of use and the direct interpretability of its coefficients, suffers from the fundamental flaw of assuming linearity. This assumption may not always account for the complex, non-linear interrelationships between soil parameters that usually exist, which, in turn, results in the constructed models being less accurate (as shown, for example, by the R^2 values reported in several studies being relatively low [1]). Thus, based on statistical learning theory, SVM has turned out to be an outstanding resource for geotechnics both in the classification of specific data and in regression. The formula for regression, Support Vector Regression (SVR), is the most powerful. The key concept is that of "kernel trick." This technique introduces the idea of implicitly embedding the data into a higher-dimensional feature space where complex, non-linear relationships can be seen as linear [18]. Therefore, it can stress that SVM is more than capable of addressing the non-linearity that is intrinsic to soil mechanics while remaining highly robust, particularly with small to medium datasets where the use of deep neural net models could lead to overfitting. A lot of research has shown that this method has been successfully used in the fields mentioned [18]. Applied the SVR to foretell soil compaction parameters; they reached extremely high R-squared values (0.86 for OMC and 0.91 for MDD) and outperformed the simpler regression models, thus proving the efficacy of the method [4]. juxtaposed it with SVM, among other methods, and after analyses, the SVM method proved to be quite good, especially when the Gaussian Radial Basis Function (RBF) was used as a non-linear kernel. The findings derived from these investigations, in unison, highlight that the reliability and accuracy of predictions no longer stay within the bounds of SVM alone, and it can also be a good alternative to the classical regression and other complex ML models. To provide a broader context for the current study, Table 1 summarizes the key findings and characteristics of several relevant research papers on the prediction of soil compaction parameters. This table facilitates a quick comparison of methodologies, datasets, and outcomes across the field.

Table 1: Summary of Key Findings from Relevant Literature

Author &Year	Model Investigated	Key Input Variables	Dataset Size	Key Performance (Test Set)	Main Conclusion
Li et al. (2024)	SVM, ANN, RF, XGBoost	FC, S, SG, LL, PL	168	XGBoost (MDD R^2=0.91), SVM (MDD R^2=0.88)	Ensemble models (XGBoost) outperform single algorithms. LL and PL are the most important features.
Ali et al. (2024)	ANN, NLR, LR, MLR	G, S, F, LL, PL, PI	2162	ANN (OMC R^2=0.92), MLR/LR (MDD R^2≈0.76)	ANN is superior for OMC prediction, but simpler linear models are reliable for MDD. PL is most influential for OMC, G for MDD.
Hasnat et al. (2019)	Support Vector Regression (SVR)	LL, PL, PI	40 + 5 validation	MDD R^2=0.91, OMC R^2=0.86	SVR provides highly accurate predictions, especially when combining LL and PL.
Tenpe and Kaur (2015)	Artificial Neural Network (ANN)	LL, PI, %finer/coarser than 75µm	210	MDD R^2=0.81, OMC R^2=0.76	ANN models are satisfactory for predicting compaction parameters from basic index properties.
Sinha and Wang (2008)	Artificial Neural Network (ANN)	Gradation (Fm, D10, U), Plasticity (LL, PL)	55	MDD R^2=0.978, OMC R^2=0.920	ANN models can predict compaction and permeability with sufficient accuracy and can be upgraded as more data becomes available.

1.2 Research Gap, Objectives, and Contribution

While a growing body of research has explored the application of various ML models for predicting soil compaction parameters, a

focused, in-depth comparison between the transparent, linear MLR model and the powerful, non-linear SVM model using a consistent dataset is essential. Many studies either focus on a single advanced model or compare a wide array of complex models without a clear baseline, leaving a gap in understanding the specific trade-offs between the simplicity and interpretability of MLR versus the predictive accuracy of SVM for this particular geotechnical problem. This study aims to fill that gap by providing a direct, rigorous comparison of these two foundational modeling techniques. The primary objectives of this research are therefore defined as follows: To develop predictive models for Field Dry Density (FDD) using both Multiple Linear Regression (MLR) and Support Vector Machine (SVM) based on a unified dataset of fundamental soil index properties. To conduct a rigorous comparative performance evaluation of the developed models using standard statistical metrics, including the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE). To perform a sensitivity analysis to identify and rank the most influential input features for predicting each compaction parameter within the context of the better-performing model.

To discuss the practical implications, advantages, and limitations of each model, providing clear guidance for their potential application in geotechnical engineering practice. This paper publication makes a contribution to the field by providing a direct, side-by-side comparison of MLR and SVM performance, thereby showing the extent of nonlinear modeling applicability for the prediction of soil compaction issues. The results obtained from this study will assist the geotechnical engineers and researchers considerably in determining a suitable model having the degree of the desired balance between the accuracy of the model and the ease of the interpretability.

2. Methodology

2.1 Research Framework

The study makes use of a systematic and structured framework that ensures the reliable construction and strict assessment of the predictive models. Initial steps carried out in the procedure include collecting and preparing data and creating two separate models for prediction (MLR and SVM). The procedure ends with a detailed performance evaluation and sensitivity analysis. This framework is the means of a straight line, reproducible comparison between a classical linear statistical method and a very strong technique that is machine learning.

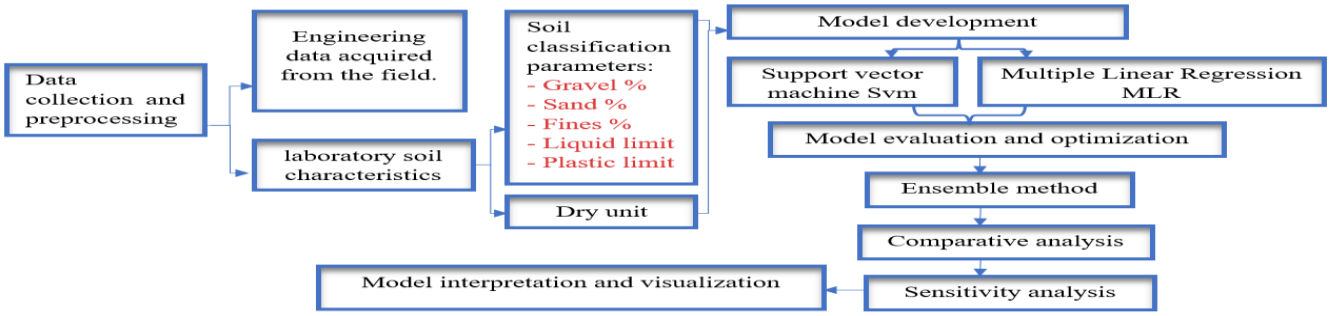

Figure (1): shows the conceptual representation of the research methodology framework that is composed of the sequential steps from data collection to the analysis and conclusion.

2.2 Dataset and Variables

The foundation of this study is a comprehensive dataset compiled from laboratory test results. The dataset consists of 86 unique soil samples from Basrah governorate field works, providing a diverse range of geotechnical properties suitable for developing and testing predictive models. Each sample in the dataset is characterized by a set of input variables (predictors) and corresponding output variables (targets).

Input Variables (Predictors): The selection of input variables was based on established geotechnical principles and a review of existing literature, which consistently identifies soil gradation and plasticity as primary factors influencing compaction. The five input features used are:

Gravel Content (G%): The percentage of soil particles larger than 4.75 mm.

Sand Content (S%): The percentage of soil particles between 0.075 mm and 4.75 mm.

Fines Content (F%): The percentage of soil particles smaller than 0.075 mm (silt and clay).

Liquid Limit (LL%): The water content at which soil transitions from a plastic to a liquid state.

Plastic Limit (PL%): The water content at which soil transitions from a semi-solid to a plastic state.

Output Variables (Targets): The models were developed to predict the primary compaction parameter. For the purpose of this study, the target variable is the Field Dry Density (FDD), which serves as a proxy for the Maximum Dry Density (MDD) that would be determined in a laboratory setting.

Field Dry Density (FDD g/cm³): This serves as the target variable representing the compacted dry density of the soil.

An initial statistical analysis was performed to understand the distribution and inter-relationships of the variables. Table 2 presents the descriptive statistics for the entire dataset, summarizing the central tendency and dispersion of each parameter. Figure 2, show the frequency distribution of the sample's minimum and maximum ranges for inputs and output. These provides an overview of the data ranges covered in the study.

Table 2: Descriptive statistics of the collected database (N=86)

NO.	Parameters	Units	Types	Minimum	Maximum	Mean	Median	standard division

1	Gravel G	Percentage	Input	2.70	45.80	25.91	26.35	10.25
2	Sand S	Percentage	Input	45.60	91.10	61.63	59.80	9.35
3	Fines F	Percentage	Input	5.80	28.00	12.34	11.00	4.47
4	Liquid Limit LL	Percentage	Input	17.00	35.80	24.04	23.32	4.05
5	Plastic Limit PL	Percentage	Input	12.65	22.62	19.09	19.93	2.51
6	Fielddry Density FDD	g/cm3	Output	1.92	2.10	2.03	2.04	0.04

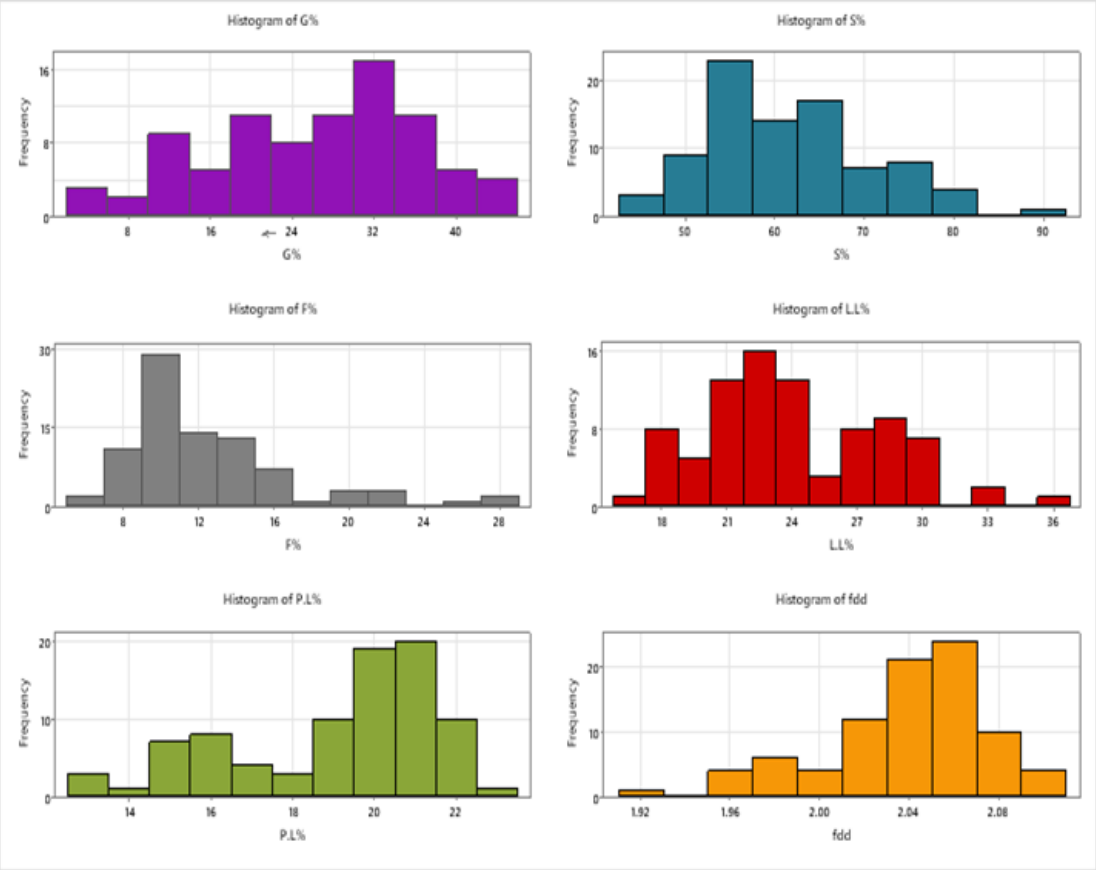

Figure (2): The graphs (A, B, C, D, E, and F) depict the frequency distribution of the sample's minimum and maximum ranges for gravel (G%), sand (S%), fines (F%), liquid limit (LL%), plastic limit (PL%), and field dry density (FDD).

2.3 Data Preprocessing

Proper data preprocessing is a critical step in the machine learning pipeline, ensuring that the data is in a suitable format for model training and helps to improve model performance and stability. The preprocessing stage in this study involved two key steps: data splitting and data normalization.

Data Splitting: The entire dataset of 86 samples was randomly partitioned into a training set and a testing set. Following common practice in machine learning to ensure a robust evaluation, 80% of the data (69 samples) was allocated for training the models, while the remaining 20% (17 samples) was held out as an unseen test set. The training set is used to "teach" the models the underlying patterns between the input and output variables. The dataset, which the model is not trained on, is employed in the assessment of the generalization performance of the model unbiased, that is its capacity to correctly make the predictions on the new data [4]. The aims, test and training data collaborating with statistics for statistical features of the test and train sets can be seen in tables (3) and (4) respectively.

Table 3: Statistical parameters of the test database.

NO.	Parameters	Units	Types	Minimum	Maximum	Mean	Median	standard division
1	Gravel G	Percentage	Input	20.00	45.80	2.20	28.6	9.09

2	Sand S	Percentage	Input	45.6	69.6	57.52	58.10	8.21
3	Fines F	Percentage	Input	8.20	21.40	11.747	10.00	3.528
4	Liquid Limit LL	Percentage	Input	18.25	35.80	24.37	23.73	4.84
5	Plastic Limit PL	Percentage	Input	14.26	22.31	18.399	19.14	2.806
6	Fielddry Density FDD	g/cm3	Output	2.01	2.09	2.0406	2.03	2.09

Table 4: Statistical parameters of the train database

NO.	Parameters	Units	Types	Minimum	Maximum	Mean	Median	standard division
1	Gravel G	Percentage	Input	2.70	41.50	24.72	26.00	10.32
2	Sand S	Percentage	Input	46.80	91.1	62.65	60.60	9.39
3	Fines F	Percentage	Input	5.80	28.00	12.504	11.00	4.677
4	Liquid Limit LL	Percentage	Input	17.00	33.14	23.961	23.00	3.87
5	Plastic Limit PL	Percentage	Input	12.65	22.62	19.262	20.03	2.422
6	Field dry Density FDD	g/cm^3	Output	1.92	2.10	2.0325	2.04	0.0385

Data Normalization: Statistical machine learning algorithms, especially distance-based or those using the gradient descent optimization method, can be impacted considerably by the scale of input features. More significantly, the factors having larger values can give an unjust weightage to the model, make the model learn slower, and decrease the model's performance. In order to counter this, data normalization was used. For this purpose, all input and output variables were rescaled to a common range in this study. This mechanism helps to ensure that equal attention is given to each of the features in the training phase, which eventually leads to a more stable, and, in most cases, a more accurate model [22]. A well-known method like Min-Max scaling converts each feature to a specific range, usually [0, 1] or [-1, 1], through the equation:

$$X_{normalized} = (X_i - X_{min}) / (X_{max} - X_{min})$$

This standardized dataset was then used for all subsequent model development and evaluation steps.

2.4 Development of Predictive Models

Two different modeling methods were used for the prediction of the soil compaction parameter (FDD): Multiple Linear Regression (MLR) and Support Vector Machine (SVM). The decision to use these two techniques provides the opportunity to make a direct comparison of their performance, a classic, interpretable linear model, and a powerful, non-linear machine learning algorithm.

2.4.1 Model 1: Multiple Linear Regression (MLR)

Theoretical Foundation: A statistical technique called MLR, which stands for multiple linear regression, is a modeling tool developed to describe the linear relationship between one dependent variable and one or more independent variables. Data fitting should be done with a linear equation that is specific to the dataset. An MLR equation is represented in the generic form as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where Y is the dependent (target) variable (FDD), $X_1 \ldots X_p$ are the independent (predictor) variables (G%, S%, F%, LL, PL), β_0 is the intercept, $\beta_1 \ldots \beta_p$ are the regression coefficients representing the change in Y for a one-unit change in the respective X, and ε is the model error term.

Model Implementation: The MLR model was trained using the 69 samples in the training dataset. The process involves using the least-squares method to estimate the coefficients (β values) that minimize the sum of the squared differences between the actual FDD values and the values predicted by the linear equation. The final output is a single, interpretable equation that can be used to predict FDD based on the five soil index properties.

2.4.2 Model 2: Support Vector Machine (SVM)

Theoretical Foundation (Support Vector Regression—SVR): SVM, when adapted for regression tasks, is known as Support Vector Regression (SVR). Unlike traditional regression, which aims to minimize error, SVR works on the principle of defining a margin of tolerance, known as the ε-insensitive tube.

The algorithm aims to fit as many data points as possible within this tube while balancing model complexity and prediction error. Data points outside this tube are penalized, but errors for points inside the tube are ignored. This approach makes SVR robust to outliers and effective in capturing the underlying data structure.

The Kernel Trick: A key feature of SVM is its ability to model non-linear relationships through the "kernel trick." This technique allows the algorithm to operate in a high-dimensional feature

space without explicitly computing the coordinates of the data in that space. Instead, it uses a kernel function to compute the dot products between the images of all pairs of data in the feature space. This study will compare two common kernel functions:

Linear Kernel: $K(x, z) = x^T z$. This results in a linear model, providing a baseline for SVM performance comparable to MLR.

Gaussian (RBF) Kernel: $K(x, z) = exp(-\gamma ||x - z||^2)$. This is a powerful non-linear kernel capable of capturing complex relationships. The parameter γ defines the influence of a single training example.

2.5 Model Evaluation and Validation Strategy

To quantitatively assess and compare the performance of the developed MLR and SVM models, a set of standard statistical metrics was employed. These metrics provide objective measures of predictive accuracy and are widely used in the machine learning literature. The evaluation was performed on the unseen test set to provide an unbiased estimate of how the models would perform on new, real-world data. The following performance metrics were used:

Coefficient of Determination (R^2): This metric indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 value ranges from 0 to 1, where 1 indicates a perfect fit and 0 indicates that the model provides no better prediction than the mean of the target variable. The formula is:

$$R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2]$$

Root Mean Squared Error (RMSE): RMSE is the square root of the average of the squared differences between predicted and actual values. It measures the standard deviation of the residuals. Because the errors are squared before they are averaged, RMSE gives a relatively high weight to large errors. A lower RMSE value indicates a better model fit. The formula is:

$$RMSE = \sqrt{[\Sigma (y_i - \hat{y}_i)^2 / n]}$$

Mean Squared Error (MSE): (MSE) is a metric that provides information about the average size of the prediction errors, ignoring their signs. It is the mean of the absolute value of the deviations of the forecast from the actual observations in the test sample. MSE is less affected by the presence of large errors (outliers) than RMSE. A better model fit corresponds to a smaller value of the MAE. The expression of MSE is

$$MSE = [\Sigma(y_i - \hat{y}_i)^2] / n$$

In these formulas, the actual value is represented by y_i, the predicted value is represented by $\hat{y}_i$, the mean of the actual values is represented by $\bar{y}$, and the number of samples is represented by n. The final validation process is performed by training the models on the 80%training data and further using the trained models for making predictions on the 20% test data. The performance metrics that are derived from this test set are used as the definitive measure for each model's predictive capability.

3. Results

The findings of the Multiple Linear Regression (MLR) and Support Vector Machine (SVM) models are disclosed in this section. These are statistical models that are designed specifically to predict FDD, or field dry density. These models were first exposed to 80% of the dataset to see how they would increase their accuracy in weight rerouting as of and subsequently assessed on 20% of the dataset for returning their generalization capabilities. Their performance is mentioned in terms of the coefficient of determination (R^2), root mean squared error (RMSE), and mean squared error (MSE).

3.1 Performance of the Multiple Linear Regression (MLR) Model

The MLR model was created to find out a direct linear mark between the five input soil properties (G%, S%, F%, LL, PL) and the target variable (FDD). The obtained is an equation, which is clear and easy to understand users who are interested in developing a tool for prediction equation. The linear equation FDD based on the training dataset is as follows:

$$FDD\ (g/cm^3) = 1.9651 + 0.00156G - 0.00004S - 0.00492F - 0.0004LL + 0.00531PL$$

The performance of this model on both the training and testing datasets is summarized in Table 5. The model achieved a high R^2 value of 0.956 on the training data, indicating a strong linear fit to the data it was trained on. However, when applied to the unseen test data, the performance dropped, yielding an R^2 of 0.814. This decrease suggests that while a linear relationship captures a significant portion of the variance, it may not fully account for the more complex interactions present in the data, leading to some degree of overfitting or a lack of generalization.

Table 5: Performance Results of the MLR Model

Dataset	R^2	RMSE (g/cm^3)	MSE (g/cm^3) ^{2}
Training	0.956	0.00836	6.99e-05
Testing	0.814	0.01150	1.32e-04

The scatter plot in Figure 4 illustrates the correlation between the FDD values predicted by the MLR model and the actual measured values for the test set. While many points cluster around the ideal 1:1 line, there is noticeable scatter, particularly for higher density values, which visually confirms the R^2 value of 0.814 and the presence of prediction errors.

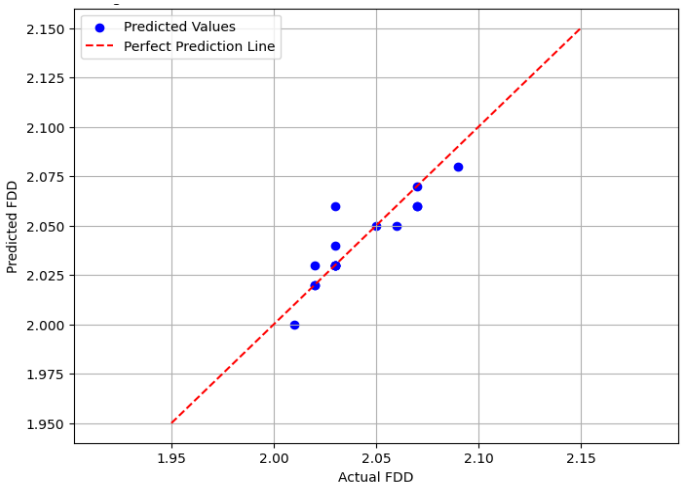

Figure 4: Predicted vs. Actual FDD for the MLR Model on the Test Set.

3.2 Performance of the Support Vector Machine (SVM) Model

The SVM model was developed to explore both linear and non-linear relationships. The performance of the SVM with a linear kernel was found to be comparable to the MLR model. However, the introduction of the non-linear Gaussian Radial Basis Function (RBF) kernel yielded a significant improvement in predictive accuracy, indicating that the underlying relationships between the soil properties and compaction are indeed non-linear. After hyperparameter tuning, the optimized SVM-RBF model was selected for final evaluation. The performance of the optimized SVM model is detailed in Table 6. The model achieved an exceptionally high R^2 of 0.967 on the training set and, crucially, maintained this high level of performance on the test set with an R^2 of 0.988. This demonstrates the model's excellent ability to generalize to new, unseen data with minimal performance degradation. The RMSE on the test set was 0.0080 g/cm³, substantially lower than that of the MLR model, indicating smaller average prediction errors.

Table 6: Performance Results of the Optimized SVM Model (RBF Kernel)

Dataset	R^2	RMSE (g/cm³)	MSE (g/cm³) ²
Training	0.967	0.00190	3.80e-06
Testing	0.988	0.00800	6.46e-05

Figure 5 presents the scatter plot for the SVM model's predictions on the test set. The data points are tightly clustered around the line of perfect equality, visually confirming the high R^2 value and the superior predictive power of the SVM model compared to the MLR model. The minimal scatter indicates a very strong correlation between the predicted and actual FDD values.

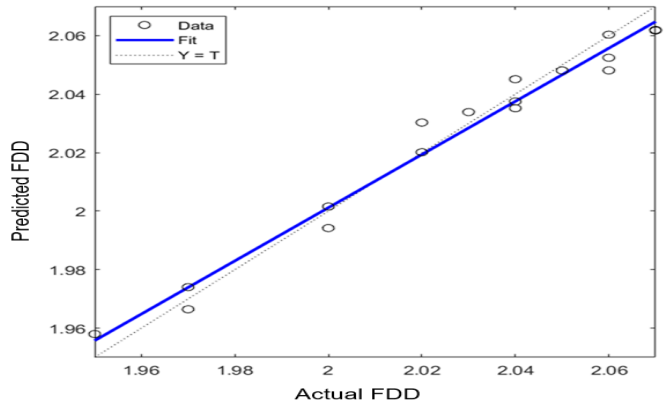

Figure 5: Predicted vs. Actual FDD for the SVM Model on the Test Set.

3.3 Comparative Analysis of Models

A direct comparison of the MLR and SVM models on the test dataset underscores the significant advantage of using a non-linear approach for this prediction task. Table 7 and Figure 6 consolidate the key performance metrics, providing a clear head-to-head evaluation. The SVM model outperformed the MLR model across all evaluation metrics. The R^2 value for the SVM model (0.988) was approximately 21% higher than that of the MLR model (0.814), indicating that the SVM model could explain a much larger proportion of the variance in the FDD. Correspondingly, the RMSE of the SVM model (0.0080 g/cm³) was about 30% lower than the MLR model's RMSE (0.0115 g/cm³), signifying substantially more accurate predictions on average. This stark difference in performance strongly suggests that the relationships between soil index properties and compaction parameters are

inherently non-linear, and models capable of capturing this complexity, such as SVM with an RBF kernel, are far more effective.

Table 7: Comparative Performance of MLR and SVM Models on the Test Set

Model	R^2	RMSE (g/cm³)	MSE (g/cm³) ²
Multiple Linear Regression (MLR)	0.814	0.0115	1.32e-04
Support Vector Machine (SVM)	0.988	0.0080	6.46e-05

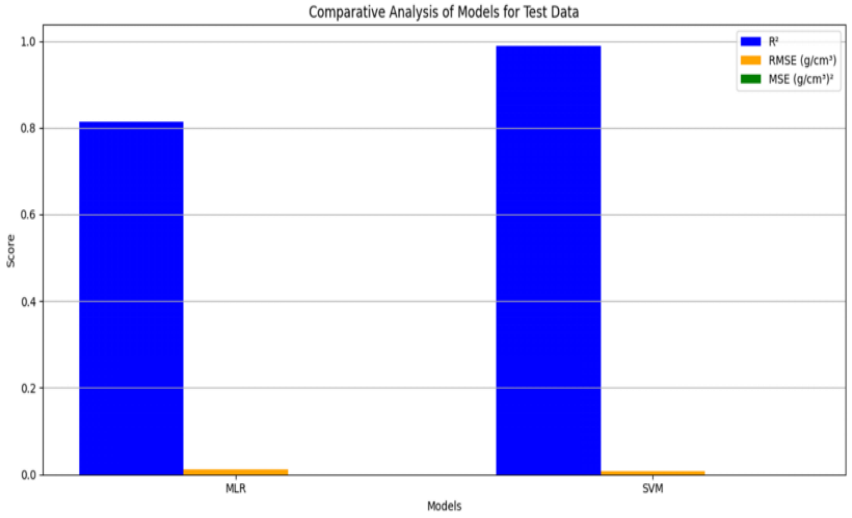

Figure 6: Comparison of R^2 and RMSE for MLR and SVM Models on the Test Set.

3.4 Sensitivity and Feature Importance Analysis

To understand which input variables had the most significant impact on the predictions of the superior SVM model, a feature importance analysis was conducted. This analysis helps to open the "black box" of the model by quantifying the contribution of each feature to the predictive outcome. The results, presented in Table 8 and Figure 7, rank the features based on their influence. According to the outcome of the study, it was found that the Fines content (F%) is the SVM model's principal determining feature that reflects its importance the most, which is closely followed by Gravel content (G%) and Liquid Limit (LL%). Sand content (S%) and plastic limit (PL%) had a comparatively smaller effect on the model's performance. This is a reasonable finding because the proportion of fine-grained material (silt and clay) is the main factor that impacts the soil's structure, water capacity, and overall response during compaction. The deficient fines content is in conformity with the results of previous studies, which is also the case in the one conducted by Karimpour-Fard et al. (2019), where it was also emphasized that its effect is profound. The considerable effect of gravel content also makes sense since the presence of larger particles determines the arrangement and the total mass of the soil matrix. These observations give new clues for the accurate prediction of soil grading, noting that especially the fines and gravel content are essential for reliable compacting.

Table 8: Feature Importance for the SVM Model

Rank	Feature	Importance Score (Normalized)
1	Fines (F%)	0.06811
2	Gravel (G%)	0.02704
3	Liquid Limit (LL%)	-0.01287
4	Sand (S%)	0.00868
5	Plastic Limit (PL%)	0.00222

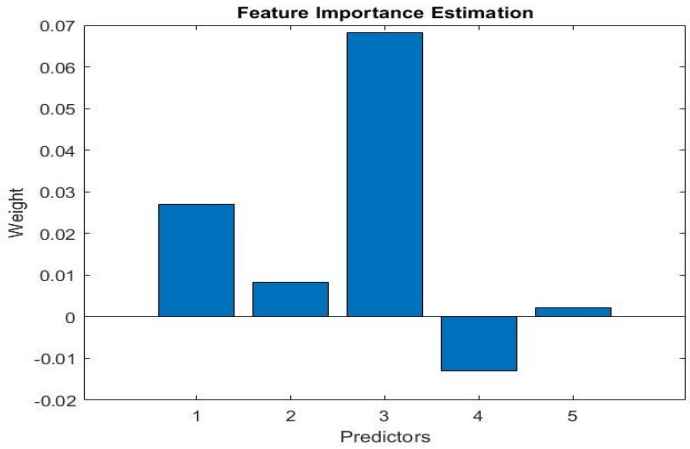

Figure 7: Feature importance analysis for SVM Model.

The significance scores are obtained from the internal metrics of the model; negative values may appear in some importance calculation methods and denote a complicated, non-monotonic relationship. The size represents the degree of the power exercised.

4. Discussion

4.1 Interpretation of Model Performance

The outcomes of this research create a laser narrative concerning the aspects of soil compaction behavior as well as the applicability of different modeling approaches for its prediction. The Multiple Linear Regression (MLR) model is the one to beat, with an R^2 of 0.814 on the test set, meaning that a linear combination of only seven other soil index properties can explain roughly 81% of the variance in Field Dry Density (FDD). The obtained equation being simple, transparent, and readily usable makes it a highly beneficial tool for engineering practice. According to the coefficients, for example, an increase in the plastic limit is the factor that protects density, while an increase in fines content has a negative linear effect. Nevertheless, the drop in performance from the training set (R^2=0.956) to the test set indicates the linear assumption's restriction that it must overcome. The model's weakness to generalize results in the misunderstanding of the intricate, non-linear properties of action between soil particles, water, and compactive energy that result in the final compacted status. The relative superiority of the Support Vector Machine (SVM) model, especially in the case of the non-linear RBF kernel, is the strongest evidence for this interpretation. Scoring an R^2 of 0.988 on the test data, the SVM model is successful in revealing its ability to appropriately model such complex, non-linear relations. A kernel trick, or the so-called SVM, allows it to work in a space that has more dimensions; hence, the intricate patterns become linearly separable, which makes the predictive model more accurate and robust. The considerable showing of SVM against MLR is not just a statistical artifact but serves to reflect the physical truth about the mechanics of soil. The behavior of soil is not typically linear; the like of particle shape, mineralogy, etc., plus water acting as both a lubricant and a void-filler contribute to a situation that linear equations can only estimate; hence, a complex system is created. The accomplishment of the SVM model endorses the theory that the machine learning algorithms that have the capability to learn these non-linearities are the core of the high-fidelity predictions in geotechnical engineering. The head-to-head comparison highlights a classic trade-off in machine learning: interpretability versus accuracy. The MLR model is fully transparent—the influence of each variable is explicitly defined by its coefficient. The SVM model, in contrast, operates as more of a "black box"; while its predictions are highly accurate, the exact mathematical relationship it has learned is not readily apparent. However, the substantial gain in accuracy (a ~30% reduction in RMSE) offered by the SVM model arguably justifies this loss of simple interpretability, especially in a context where predictive precision can lead to safer and more economical designs. This finding aligns with a broader trend in engineering, where data-driven models are increasingly accepted based on their validated performance, even if their internal workings are complex.

4.2 Contextualizing Findings with Existing Research

The performance of the models developed in this study is consistent with, and builds upon, the findings of previous research in this domain. The R^2 value of 0.988 achieved by our SVM model is highly competitive and, in some cases, superior to those reported in similar studies. For example, [18] reported R^2 values of 0.86 for OMC and 0.91 for MDD using an SVR model, while [18] achieved an R^2 of 0.9098 for MDD with an XGBoost model and 0.8831 with an SVM model. Our higher R^2 value may be attributable to the specific characteristics of our dataset or the thoroughness of the hyperparameter tuning process, but it firmly places our SVM model at the upper end of reported performance for this task. The finding that the non-linear SVM model significantly outperforms the linear MLR model aligns with the conclusions of [3], who also found that their ANN model (another non-linear approach) outperformed LR, NLR, and MLR models for predicting OMC. This recurring theme across different studies, using different datasets and non-linear models (ANN, SVM, XGBoost), solidifies the conclusion that non-linearity is a defining characteristic of the problem. The feature importance analysis also resonates with the existing literature. Our finding that fines content (F%) is the most influential predictor is supported by [20], who emphasized the significant impact of fine content on compaction characteristics. This is geotechnically intuitive, as the fine fraction controls the soil's plasticity and hydraulic properties. Our findings, nevertheless, slightly deviate from those of [4,3], who reported Atterberg limits (LL and PL) as the most influential predictors. Accordingly, this disparity is nothing but proof of the multifaceted nature of the system, not a counterargument. Plasticity's relative importance compared to gradation can fluctuate depending on the types of soil included in the dataset. So, in datasets with a predominance of fine materials, the plasticity parameter could be the leading factor. Conversely, in data with a greater variety of gradations, such as the one applied in the current research (fines content ranging from 5.8% to 28%), the particle size distribution would be more salient. This accentuates the necessity of constructing models based on a variety of datasets in order to encompass the entire range of soil behavior.

5. Conclusion

This study set out to develop and compare predictive models for soil compaction parameters using Multiple Linear Regression (MLR) and Support Vector Machines (SVM). Based on a dataset of 86 soil samples characterized by their gradation and plasticity, we demonstrated that machine learning offers a highly effective means of estimating Field Dry Density (FDD). The key findings are threefold: First, the SVM model, equipped with a non-linear RBF kernel, significantly outperformed the traditional MLR model, achieving a test set R^2 of 0.988 compared to MLR's 0.814. This confirms that the relationship between soil index properties and compaction behavior is fundamentally nonlinear. Second, the feature importance analysis identified soil gradation, particularly the percentage of fines, as the most influential predictor in the dataset, followed by gravel content. Third, the results affirm that a well-trained SVM model can serve as a rapid, cost-effective, and highly accurate alternative to laborious laboratory testing for preliminary geotechnical assessments. This research provides compelling evidence for the adoption of machine learning techniques in routine geotechnical engineering practice. The demonstrated ability of the SVM model to accurately predict soil compaction parameters from simple index properties represents a significant step towards more efficient and data-driven site characterization. By leveraging such models, engineers can make faster, more informed decisions, optimize resource allocation, and ultimately enhance the safety and economy of civil engineering projects. While these models are tools to augment, not replace, engineering judgment and critical laboratory verification, they undoubtedly represent the future of predictive modeling in the field, where the power of data is harnessed to solve complex, real-world

References

1. Rathnam, U., & Prasad, K. (2020). ASSESSMENT OF COMPRESSION BEHAVIOUR OF COMPACTED SOILS. INTERNATIONAL JOURNAL OF CIVIL ENGINEERING AND TECHNOLOGY (IJCIET). https://doi.org/10.34218/ijciet.11.3.2020.001.
2. Hussain, S. (2017). Effect of Compaction Energy on Engineering Properties of Expansive Soil. Civil Engineering Journal, 3, 610-616. https://doi.org/10.28991/CEJ-030988.
3. Ali, H. F. H., Omer, B., Mohammed, A. S., & Faraj, R. H. (2024). Predicting the maximum dry density and optimum moisture content from soil index properties using efficient soft computing techniques. Neural Computing and Applications, 36(19), 11339-11369.
4. Li, B., You, Z., Ni, K., & Wang, Y. (2024). Prediction of Soil Compaction parameters using machine learning models. Applied Sciences, 14(7), 2716.
5. Nazari, Z., Tabarsa, A., & Latifi, N. (2021). Effect of compaction delay on the strength and consolidation properties of cement-stabilized subgrade soil. Transportation Geotechnics, 27, 100495.
6. Tilahun, Y., Qinghua, X., Ashango, A. A., & Han, X. (2024). Determination of Compaction Parameters of Cement-Lime Soils: Boosting-Based Ensemble Models. International Journal for Numerical and Analytical Methods in Geomechanics, 48(18), 4365-4382.
7. ASTM International. (2012). Standard Test Methods for Laboratory Compaction Characteristics of Soil Using Standard Effort (12 400 ft-lbf/ft3 (600 kN-m/m3)) (ASTM D698-12). ASTM International.
8. ASTM D1557 (2012) Standard test methods for laboratory compaction characteristics of soil using modified effort (56,000 ft-lbf/ft3 (2,700 kN-m/m3)). ASTM International.
9. Proctor, R. R. (1933). Description of field and laboratory methods. Engineering News-Record, 111(10), 286-289.
10. Zvonarić, M., Barišić, I., Galić, M., & Minažek, K. (2021). Influence of laboratory compaction method on compaction and strength characteristics of unbound and cement-bound mixtures. Applied Sciences, 11(11), 4750.
11. Cerni, G., Cardone, F., Virgili, A., & Camilli, S. (2012). Characterisation of permanent deformation behaviour of unbound granular materials under repeated triaxial loading. Construction and Building Materials, 28(1), 79-87.
12. Simatupang, E. C., Sari, P., & Gunawan, H. (2023). Proctor Application Benefits in Post Pandemic Online Test Implementation. Jurnal Sinestesia, 13(2), 995-1000.
13. Jamshidi Chenari, R., Tizpa, P., Ghorbani Rad, M. R., Machado, S. L., & Karimpour Fard, M. (2015). The use of index parameters to predict soil geotechnical properties. Arabian Journal of Geosciences, 8(7), 4907-4919.
14. Abdella, D., Abebe, T., & Quezon, E. T. (2017). Regression analysis of index properties of soil as strength determinant for California bearing ratio (CBR). Gsj, 5(6), 1.
15. Ramiah, B.K., Viswanath, V., Krishnamurthy, H.V., 1970. Interrelationship of compaction and index properties. In: Proceedings of the Second Southeast Asian Conference on Soil Engineering, Singapore,577, 577–587.
16. Gurtug, Y., & Sridharan, A. (2004). Compaction behaviour and prediction of its characteristics of fine grained soils with particular reference to compaction energy. Soils and foundations, 44(5), 27-36.
17. Almuaythir, S., Zaini, M. S. I., & Lodhi, R. H. (2025). Predicting soil compaction parameters in expansive soils using advanced machine learning models: a comparative study. Scientific Reports, 15(1), 24018.
18. Hasnat, A., Hasan, M. M., Islam, M. R., & Alim, M. A. (2019). Prediction of compaction parameters of soil using support vector regression. Curr. Trends Civ. Struct. Eng, 4(1), 1-7.
19. Jalal, F. E., Xu, Y., Iqbal, M., Jamhiri, B., & Javed, M. F. (2021). Predicting the compaction characteristics of expansive soils using two genetic programming-based algorithms. Transportation Geotechnics, 30, 100608.
20. Karimpour-Fard, M., Machado, S. L., Falamaki, A., Carvalho, M. F., & Tizpa, P. (2019). Prediction of compaction characteristics of soils from index test's results. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 43(Suppl 1), 231-248.
21. Sinha, S. K., & Wang, M. C. (2008). Artificial neural network prediction models for soil compaction and permeability. Geotechnical and Geological Engineering, 26(1), 47-64.
22. Tenpe, A., & Kaur, S. (2015). Artificial neural network modeling for predicting compaction parameters based on index properties of soil. International Journal of Science and Research, 4(7), 1198-1202.
23. Rowan, H. W., & Graham, W. W. (1948). Proper compaction eliminates curing period in construction fills. Civil Engineering, 18(7), 50-51.

24. Zhao, T., Shen, F., & Xu, L. (2024). Review and comparison of machine learning methods in developing optimal models for predicting geotechnical properties with consideration of feature selection. Soils and Foundations, 64(6), 101523.