Hindawi
Advances in Civil Engineering
Volume 2023, Article ID 8198648, 11 pages
https://doi.org/10.1155/2023/8198648

Research Article

Use of Machine Learning to Predict California Bearing Ratio of Soils

Semachew Molla Kassa [1,2] **and Betelhem Zewdu Wubineh** [3,4]

[1]*Faculty of Civil and Water Resource Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia*
[2]*Department of Civil Engineering, College of Engineering and Technology, Wachemo University, Hosaena, Ethiopia*
[3]*Faculty of Information and Communication Technology, Wroclaw University of Science and Technology, Wroclaw, Poland*
[4]*Department of Information Technology, College of Engineering and Technology, Wachemo University, Hosaena, Ethiopia*

Correspondence should be addressed to Semachew Molla Kassa; smakmolla23@gmail.com

Received 15 November 2022; Revised 13 December 2022; Accepted 16 January 2023; Published 25 January 2023

Academic Editor: Romulus Costache

CBR is a crucial metric used to assess the durability of base course materials and subgrade soils in various types of pavements. In this research, the machine learning (ML) approach has been implemented using random forest (RF), decision tree (DT), linear regression (LR), and artificial neural network (ANN) models to estimate CBR (California bearing ratio) values of the soil based on seven predictors such as maximum dry density, soil classification, optimum moisture content, liquid limit, plastic limit, plastic index, and swell, which can be easily determined from the laboratory. AASHTO M 145 was used to categorize 252 soil samples that formed the basis of an experimental data set. In this model study, the data were split into 20% test data and 80% training data. Standard statistical measures including coefficient of determination, correlations, and errors were used to assess the effectiveness of the models such as MSE (mean squared error), MAE (mean absolute error), and RMSE (root mean square error). From these evaluation metrics, the random forest algorithm gets a smaller error and larger relative error (R^2) value to compare with other algorithms. Therefore, it can be concluded that a random forest algorithm based on the analysis findings can accurately forecast the soil's CBR.

1. Introduction

To determine the strength of the subgrade material of the pavement structure of roads, airfields, and railways, the California bearing ratio (CBR) is a parameter of increased importance in civil engineering, particularly in construction material and geotechnical engineering. CBR can be measured both in situ and in a laboratory. The field CBR test method comprises driving a piston into the soil mass and subgrade material at the test site using a loading jack to determine the strength of in situ soils and base course material for pavement design. Field CBR equipment is expensive and cumbersome to transport to various areas. In order to determine the CBR of soil and subgrade material, laboratory procedures are typically used. CBR is measured in the laboratory by putting a standard-diameter plunger into a sample of compacted soil that has been prepared at the ideal moisture content at a pace of 1.3 mm/min. Any soil sample's CBR values can be calculated both with and without wetting the soil. The CBR values of soil samples that have been soaked are typically lower than those of unsoaked ones [1].

As a result, the CBR values of soaking samples are typically used to estimate the quality of subpar materials. Since the procedure of determining the soil CBR is time-consuming, the process significantly affects the construction time delay. However, the soil samples prepared at the optimum moisture content (OMC) need to be kept in water saturation conditions for 4 days, which is considered the worst case if the rainfall is expected to continue continuously for 4 days. CBR must typically be calculated for many samples, which is expensive and takes time [2].

CBR testing needs trained laboratory technicians. This process delays the completion of the project. One solution to this problem is to know or predict the CBR value to avoid wasting time using a machine learning technique. Machine learning is a field of artificial intelligence that study about computer algorithms that can learn to do tasks better based on prior experience without the program being explicitly [3]. In machine learning, there are different types of learning. Those are supervised, unsupervised, reinforcement, and semisupervised learning. This paper focuses on a regression-supervised learning technique to predict the CBR. Regression is a method for determining how independent features or variables relate to a dependent feature or result. Recently, researchers have combined real-world geotechnical engineering problems with machine learning techniques, such as an artificial neural network (ANN), a multilayer perceptron neural network (MLP), gene expression programming (GEP), a support vector machine (SVM), and the multigroup method of data handling, to predict the desired output data. The list of works that have been done previously is listed in the following Table 1.

From the above table, the accuracy of the prediction model is different. The model of the California bearing ratio depends on the sampling size of index properties of the soil and types of predictors and the workmanship during the laboratory test. In this literature, increasing the sampling size is not the factor in the quality of prediction accuracy. According to [5] in the SVM model, the sampling size is 49 but the prediction of the accuracy is 98 percent, and according to [10], also the sampling size is 389, but the prediction accuracy is not good. In this research, the sample size is 252, and the accuracy is 84 percent. This implies that this prediction works in this specific area because the soil type is different from the previous literature. In addition to this, the most important feature that predicts the CBR is discussed.

The objective of this study is to predict the CBR, i.e., the variation in one variable which is dependent based on the independent variable.

2. Methodology

The materials and procedures used in our study are discussed in this section.

2.1. Study Area. The study area is situated at the highway project in the Ethiopian province of Amhara region, along a 48.92 km-long route between Mekane Eyesus and Simada town. This project helps two towns, Mekane Eyesus and Simada, to communicate economically and socially and also minimize the duration of the time taken from rural area villages to hospitals during the delivery time for pregnant women. For laboratory testing, soil samples were gathered from the highway section situated along the route. The testing of 252 samples took place between 10/13/2020 and 10/7/2021. Grain size analysis, tests for figuring out the liquid limit (LL), testing for moisture-density relationships, and tests for the CBR were among the tests conducted.

Table 1: The related works used the ML model to predict the CBR of soil.

Reference	Algorithm	Accuracy	Number of soil samples
[4]	MLR, ANN	$R^2 = 0.928$ $R^2 = 0.92$	128
[5]	SVM ANN	$R^2 = 0.98$ $R^2 = 0.86$	49
[6]	RF	$R^2 = 0.98$	214
[7]	RSS-ET	$R^2 = 0.98$	214
[8]	MARS-L	$R^2 = 0.98$	312
[9]	RF	$R^2 = 0.98$	290
[10]	SVM	$R^2 = 0.77$	389

2.2. Index Properties of Soils. By eliminating air from soil particles using mechanical force, the process of soil compaction densifies the soil, resulting in good strength characteristics that lessen the permeability of the soil. With compacting effort and the amount of water given to the soil, the densification of soil varies. The compaction curve or the moisture-density curve is used to describe this relationship. The methods for determining the moisture density curve equation have been codified, and they are often established through conventional Proctor, modified Proctor, and AASHTO (American Association of State Highway and Transportation Officials) tests.

There is a standard procedure for determining this curve equation and is typically determined by tests for the CBR (AASHTO T 193) [11], swell (AASHTO T 258) [12], MDD (AASHTO T 180) [13], OMC (AASHTO T 180) [13], LL (AASHTO T 89) [14], PL (AASHTO T 89) [14], PI (AASHTO T 90) [14], and soil classification (AASHTO M 145) [15]. Both coarse- and fine-grained soils are eligible for these tests. A crucial part of a geotechnical survey is figuring out the soil's capacity to swell in the pavement. As part of the study, soil samples are often collected at shallow depths beneath the proposed pavement elevation, and their ability to swell can be assessed using a variety of methods.

When determining a soil's swelling potential and measuring the crucial moisture level of fine-grained soil, Atterberg's limits are frequently used. The shrinkage limits and/or plastic limitations will often be carried out in a laboratory. The soils' PI, LL, and PL are calculated, in accordance with AASHTO T 89 and 90. When the moisture content, or LL, rises, plastic soil will behave more like a liquid. The moisture content is known as the plastic limit, and as it rises, semisolid soil will turn plastic. The plastic index (PI) is the difference between the liquid limit (LL) and the plastic limit (PL), (PI = LL – PL).

2.3. Dataset

2.3.1. Determination of CBR Value in the Laboratory. To determine the properties of the subgrade and subbase's cohesive soil particles of pavement layers on a road, the California Department of Highways developed the CBR test method in the late 1920s. The American provided the means State Highway and Transportation Association Administration figures and the American Society for Testing, and Materials.

When building roads, airports, parking lots, and other pavement in the United States, certain organizations, including the Federal Highway Administration (FHWA), the Federal Aviation Administration (FAA), and AASHTO, have employed CBR values. Researchers identified several engineering soil metrics, including resilient modulus and CBR, and established an empirical relationship between them. The use of CBR in mechanistic and mechanistic-empirical design procedures is unsuitable because it is not a fundamental property of materials. However, it has a lengthy history in the construction of pavements, is relatively simple and affordable to conduct, and exhibits fair correlations with more fundamental characteristics like robust modulus. As a result, it is still in use in reality.

The subgrade materials are frequently described in terms of their strength and stiffness. The three primary subgrade stiffness/strength characterizations that are often used in the United States are CBR, elastic (resilient) modulus, and modulus of subgrade reaction (k). Although there are other factors to take into account when assessing subgrade materials (such as swell when materials contain clay), stiffness is the most prevalent one. Furthermore, the homogeneity of the subgrade affects pavement performance. It is difficult to obtain a perfect subgrade because of the inherent variety of the soil and the impacts of water, temperature, and construction operations. According to [16]'s research, if the subgrade strength is less than a CBR value of 10, the subbase layer will deflect under traffic loadings in the same way as the subgrade in the United States. Deflection has an impact on the pavement, initially on flexible pavements but eventually also on hard pavements.

Table 2 shows the statistical significance of the study's data.

The correlation between the parameters is depicted in the following Figure 1.

2.3.2. Influencing Factors (Input Parameters). Seven influential factors were taken into account in this study: soil classification (i.e., *A*-7–5, *A*-2–4, *A*-2–7, *A*-2–7, *A*-2–6, *A*-7–6), and liquid limit, optimum moisture content, plastic limit, maximum dry density, and plastic index dependent variables, i.e., California bearing ratio and swell for the estimation of CBR using random forest model. A thorough grasp of soil properties, adequate grading, qualified laboratory personnel, experienced geotechnical engineers, and contemporary quality control testing are needed to obtain a subgrade material of high quality. However, the relevance of the structure, its size, its lifespan, and the cost of projects should all be taken into consideration when determining the criteria for pavement design and the level of engineering work. Therefore, fundamental engineering knowledge of subgrade soil properties is necessary for the design. These include the type of soil, its density, its coefficient of lateral earth pressure, its permeability, its internal angle, its cohesive characteristics, and its estimated CBR or robust modulus.

The American Concrete Pavement Association, Asphalt Pavement Association, State of Ohio, State of Iowa, and Rolling's provide examples of typical CBR values for various soil types [17]. The value of the CBR is influenced by the soil's texture, dry density, and moisture content. The moisture content that is typically achieved for the CBR test in the laboratory is different from what is anticipated to be attained in the field. Finding the maximum dry density requires determining the stable moisture content. The worst scenario is typically taken into account in many other nations; hence, when computing CBR values, the 4-day soaking CBR samples are used.

For this experimental research, the author collects the data from 10/13/2020 and 10/7/2021. The author gets 252 records and 8 attributes of the CBR data. Before going to model development, first, it needs to preprocess the data to get a good result. Data preprocessing may include data cleaning, data transformation, data integration, and attribute selection [18]. Because of the data that get from the DANA consulting laboratory, there were no any missing values that need to be filled. However, it needs some transformation of the data from categorical value to numerical value that is suitable for the selected technique and algorithms. Therefore, the author transforms soil classification from categorical values to their corresponding numerical value. The list of attributes/parameters that are used for the study is depicted in the following Table 3.

The following Figure 2 shows the distribution of attributes in the study.

The above figure shows the density plot of each attribute in the dataset which is used to show the distribution with smooth curves. From this, LL, PL, MDD, and OMC have no skews, which means that the mean is less than the median. On the other hand, CBR is right-skewed, and the mean value is greater than the median. In addition to this, it can possible to describe the distribution of the dataset using a box plot diagram which is depicted in the following Figure 3. It is used to visualize the range and other characteristics of the data such as minimum value, maximum, median, and some quartile values.

From the above box plot of the attribute's distribution, it can be observed that there is an outlier in PL, MDD, OMC, and CBR which means that the values are greater than the middle value. The green lines are the median value of the parameters. The following Figure 4 shows the proportion of CBR in the data set in which the majority of values range from 0.5 to 5.

2.4. Methods Used. In this experimental research, to develop the prediction model, a Python programming language is used with a Jupyter notebook environment. 80 percent of the data set was used for training, while the remaining 20 percent was used for testing. Based on the linearity of the data, sample size and number of parameters random forest, decision tree and linear regression machine learning supervised algorithms, and deep learning techniques like artificial neural network (ANN) are selected to predict the CBR. This technique uses the collected laboratory soil sample of the study area. By comparing the accuracy of those techniques, the one that has good prediction accuracy is selected. Supervised learning is a function that connects

Table 2: Statistical value of the data in the study.

	SC (%)	LL (%)	PL (%)	PI (%)	MDD (kg/m^3)	OMC (%)	Swell (%)	CBR (%)
Max	36	83.1	46.29	45.67	18	39.5	12.23	124
Min	35	37	21.78	9.63	1.23	2.6	0.15	0.24
Avg	35.57	61.1	31.55	29.54	1.6	22.03	4.16	15.17
Mean	35.57	61.1	31.55	29.54	1.6	22.03	4.16	15.17
Skew	−0.27	−0.08	0.94	−0.12	0.15	0.15	0.23	1.83
Kurt	−1.94	−0.69	4.94	−0.95	243.55	2.82	−1.23	3.9
Var	0.25	96.37	9.64	65.12	1.09	19.55	9.17	394.03
Std	0.5	9.82	3.1	8.07	1.05	4.42	3.03	19.85

where SC = soil classification, LL = liquid limit, PL = plastic limit, PI = plasticity index, MDD = maximum dry density, OMC = optimum moisture content, CBR = California bearing ratio.

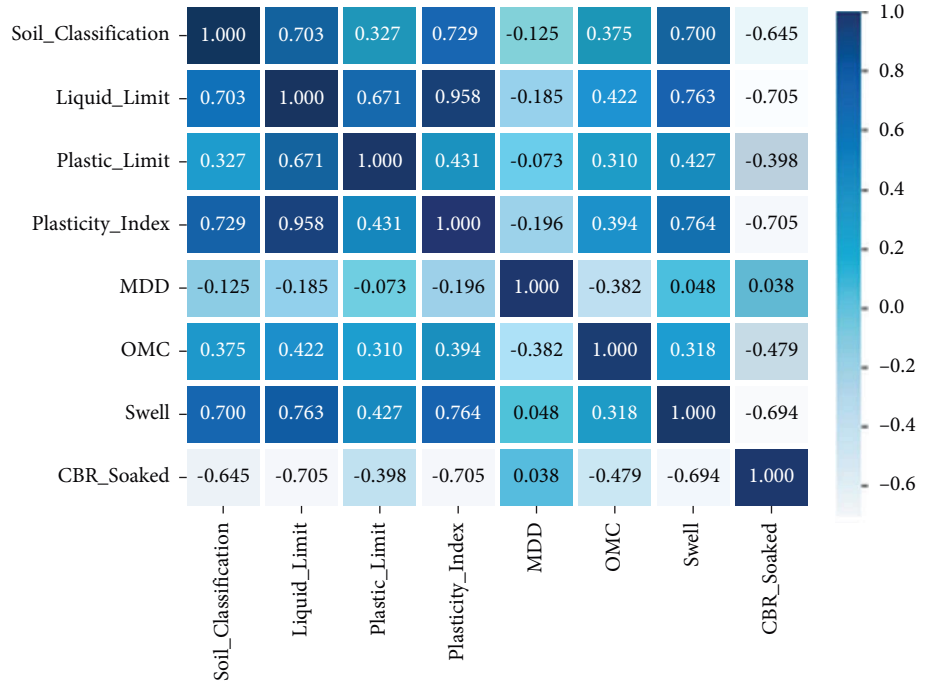

Figure 1: The correlation of parameters.

Table 3: List of parameters in the study.

No	Parameter	Description
1	Soil classification	Is the separation of soil into classes or groups based on different criteria?
2	Liquid limit	The water content where the soil starts to behave as a liquid
3	Plastic limit	Changes from semisolid to plastic
4	Plasticity index	The difference between LL and PL
5	MDD	The dry density of the soil corresponding to optimum moisture content
6	OMC	The water content at which the soil attains maximum dry density
7	Swell	Soil containing montmorillonite clay minerals or others
8	CBR soaked	Simulate the worst condition: the subgrade material gain moisture

inputs with desired outcomes [19]. Regression is one of the tasks in supervised learning. Finally, mean squared error (MSR), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) are used to assess the performance of the method. The study's flowchart is shown in the following Figure 5.

2.4.1. Random Forest ML Model. Using a randomized variant of the tree induction mechanism, a set of techniques called random forests can be used to build an ensemble of decision trees. Different from conventional decision trees, random forests approaches add random perturbations to the induction process.

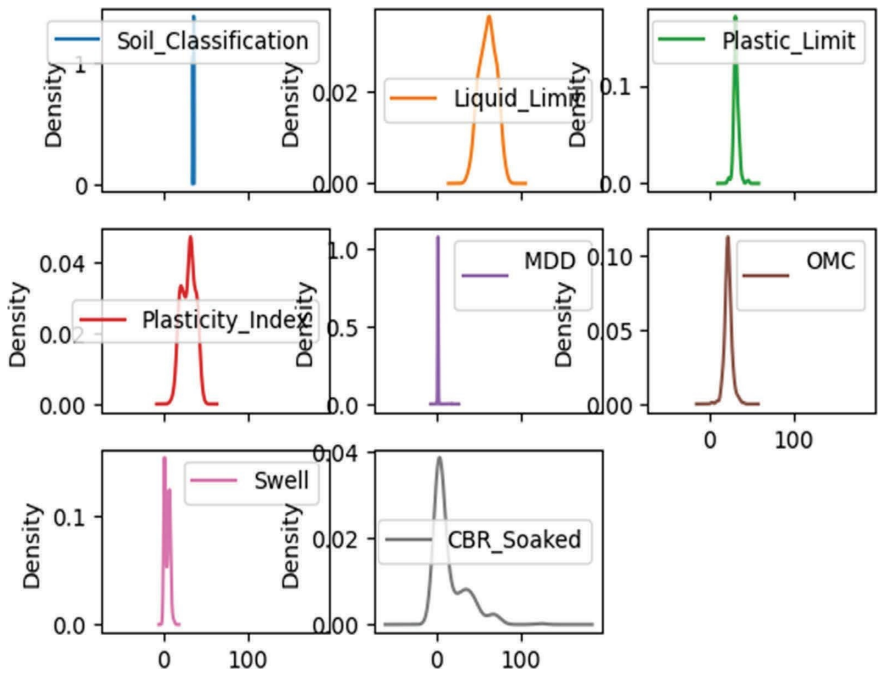

Figure 2: The distribution of attributes using density plot.

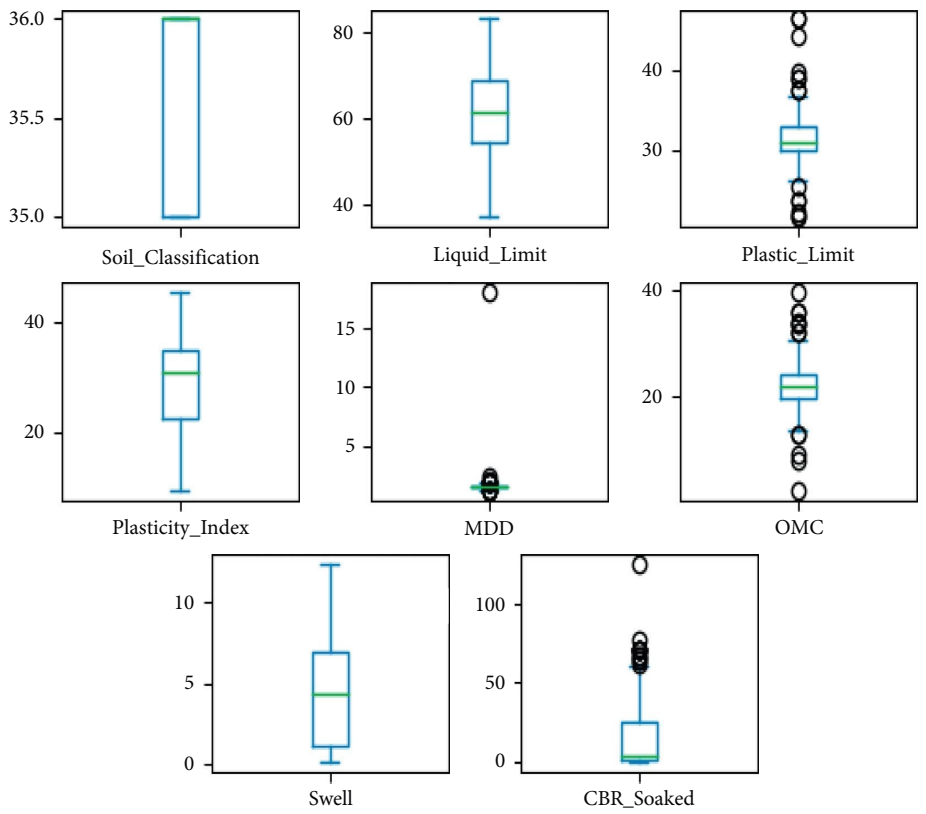

Figure 3: The distribution of the dataset using a box plot.

It is challenging to minimize (x) and maintain a minimal bias while introducing randomization into any decision tree. Kwok and Carter were the first to introduce the ensemble of decision trees [20]. Averaging many decision trees with various structure types frequently yields better results than any one of the ensemble's component parts. This approach, however, was neither completely automatic nor random. Rather, decision trees were constructed by manually

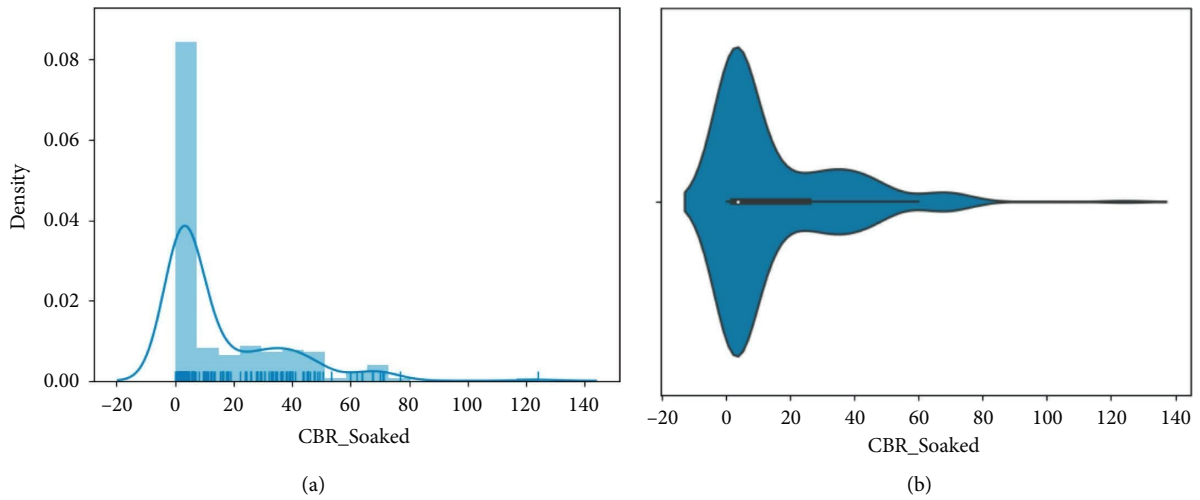

Figure 4: Proportion of CBR in the dataset.

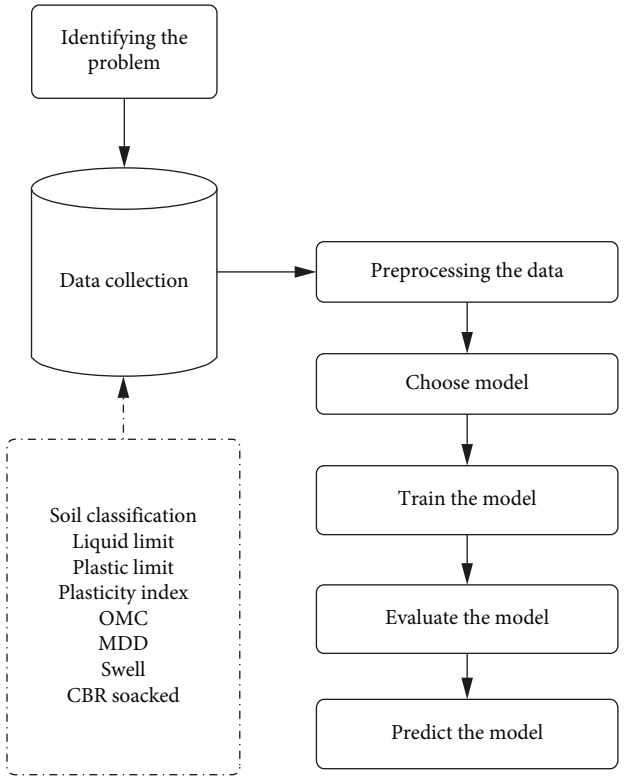

Figure 5: Flowchart of the study.

selecting splits towards the top of the tree that was almost as good as the ideal splits, enlarging them, and then using the ID3 induction method as usual.

One of the first to demonstrate, mathematically and practically, that aggregating various iterations of a predictor into an ensemble may lead to appreciable gains in reliability was Breiman, a formerly technical study. He notices and shows that the expected generalization error is lower for the mean model for the model L Lm (for $m = 1, \ldots, M$) of the training set L [21]. The "Lm" form consists of a collection of L copies, each of which has N randomly selected examples (x, y), with replacements drawn from L. Despite the fact that $|L| = |\mathrm{Lm}| = N$, the bootstrap replication reveals that on average, 37 percent of the pairs (x, y) from L are missing. In fact, there is a substantial possibility of never being chosen after N drawings with replacements.

However, when the training set L is small, subsampling 67 percent of the objects may result in an increase in bias (for example, because a model's accuracy is decreased) that is too big to be offset by a decrease in variance, which will result in poorer overall performance.

Bagging is a helpful technique in a variety of situations since it has the advantage that it can be used to improve any type of model, not only decision trees. In Breiman's foundational random forests (RF) study, each node includes bagging in addition to random variable selection. Combining both methods and adjusting randomness results in one of the most efficient off-the-shelf machine learning algorithms that perform surprisingly well for almost any task. Boosting and arcing algorithms, which are likewise intended to eliminate bias, are proven competitive with random forests, although forests place more emphasis on decreasing the error.

This technique is used to model the collection of soil test data from Mekane Eyesus to Simada town road section. Random forest technique is used to predict the California bearing ratio, which minimizes time and cost during the soil test parameter in this specific area. This prediction applies to other road construction in this study area.

2.4.2. Decision Tree. A decision tree, a hierarchical model for supervised learning, locates the local region through a sequence of recursive splits in fewer steps. Internal decision nodes and terminal leaves make up a decision tree. Both regression and classification are done using this technique. The construction of a regression tree resembles that of a classification tree almost exactly, with the exception that the measure of impurities used in classification is swapped out for a measure used in the regression.

Assume that given a node m, Xm is the subset of X that reaches node m or the set of all $x \in X$ that fulfills all the criteria in the decision nodes along the specified path from the root to node m.

$$b_m(x) = \begin{cases} 1, & \text{if } x \in x_m\text{: } x \text{ reaches node } m, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The mean square error from the estimated value determines if a tree is properly divided. Let gm be the regression's predicted value for node m.

$$\begin{aligned} E_m &= \frac{1}{N_m} \sum_t \left(r^t - g_m\right)^2 b_m\left(x^t\right), \\ N_m &= \left|x_m\right| = \sum_t b_m\left(x^t\right), \end{aligned} \tag{2}$$

E_m and m's variance are connected. A node uses the mean of the needed output samples that have arrived at the node.

$$g_m = \frac{\sum_t b_m\left(x^t\right) r^t}{\sum_t b_m\left(x^t\right)}. \tag{3}$$

A leaf node is generated and keeps the value of g_m if the error is tolerable for a node ($E_m < \theta_r$). In particular, leaf boundaries are used to build a piecewise constant approximation with discontinuities. If the error is unacceptable, the data that reach node m are further divided so that the total number of errors in the branches is kept to a minimum.

Table 4: Results of the study.

Algorithm	MAE	RMSE	MSE	R^2
RF	4.8	8.13	66.24	0.84
DT	5.3	11.92	143.3	0.66
LR	8.9	11.08	122.8	0.53
ANN	5.69	10.36	107.46	0.67

2.4.3. Linear Regression. For linear regression to work, the model's regression parameters must be linear. Regression analysis is a method for figuring out the relationship between the predictors (also known as independent variables, explanatory variables, control variables, or regressors, and typically denoted by $x1$, $x2$, ..., xp) and the response variables (also known as dependent variables, explained variables, predicted variables, or regressors, and typically denoted by y).

Regression comes in three different forms. The first is simple linear regression. Simple linear regression is used to model the linear relationship between two variables. Two of them are the independent variable x and the dependent variable y. The second type of regression is called many linear regressions, which is a linear regression model with one dependent variable and many independent variables. Multiple linear regressions make the assumption that the response variable is a linear function of the model's parameters and that there are many independent variables. The regression parameters for the third type of regression assume that the connection between the dependent variable and the independent variable is not linear. [22].

2.4.4. Validation Indicators. The developed random forest model's statistical performance was assessed using three analytical standard evaluation indicators, including the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE).

$$\begin{aligned} R^2 &= 1 - \frac{\sum_1^n \left(x_i - y_i\right)^2}{\sum_1^n \left(\overline{y}_i - y_i\right)^2}, \\ &\text{(worst value} = -\infty\text{; best value} = +1), \\ \text{MAE} &= \frac{1}{n} \sum_1^n \left|x_i - y_i\right|, \\ &\text{(best value} = 0\text{; worst value} = +\infty), \\ \text{MSE} &= \frac{1}{n} \sum_{i=1}^n \left(Y_i - \widehat{Y}\right)^2, \\ \text{RMSE} &= \sqrt{\frac{1}{n} \sum_1^n \left(x_i - y_i\right)^2}, \\ &\text{(best value} = 0\text{; worest value} = +\infty). \end{aligned} \tag{4}$$

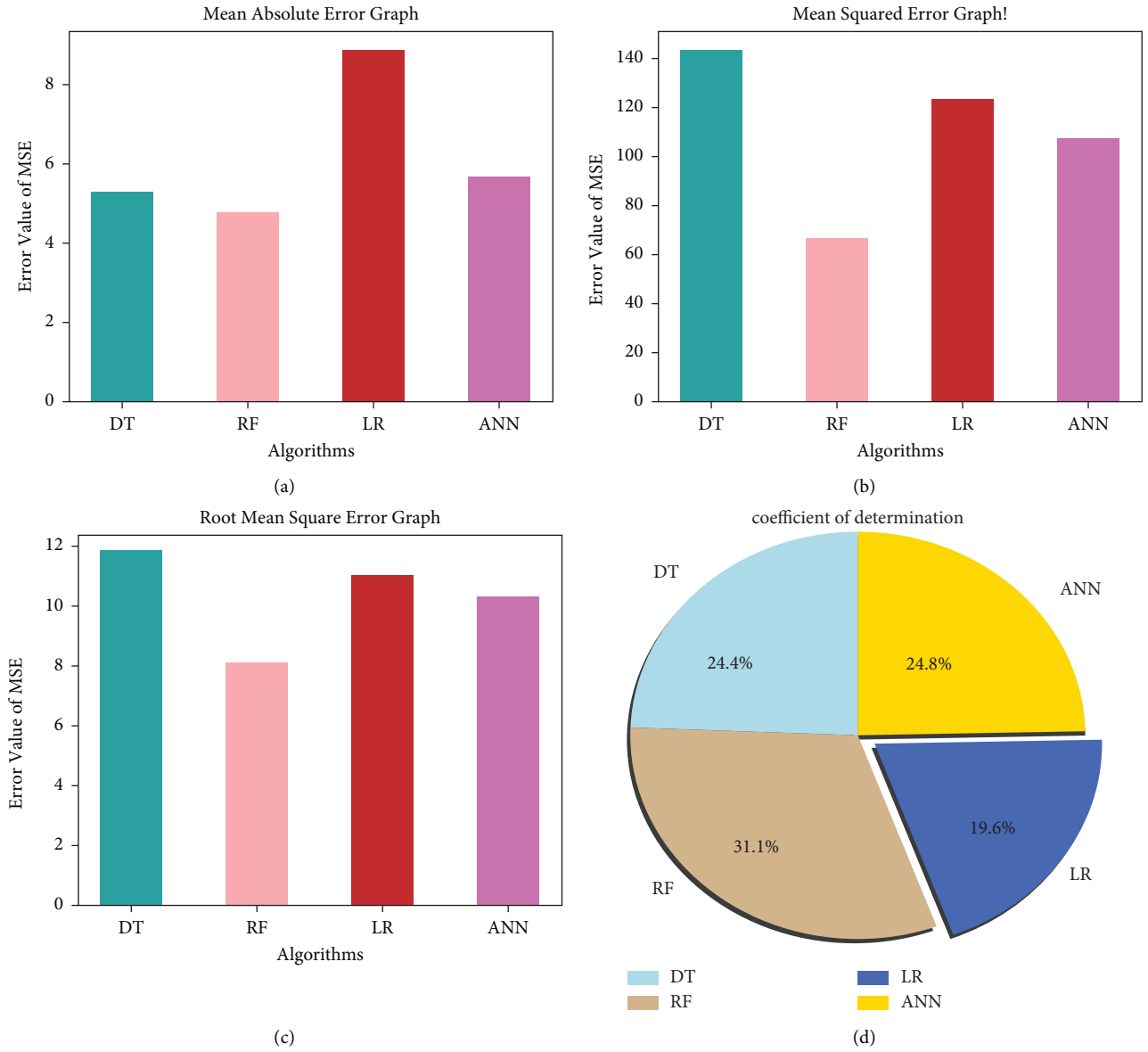

Figure 6: Comparison of the algorithms: (a) mean absolute error, (b) mean squared error, (c) root mean square error, and (d) coefficient of determination.

X_i and y_i are the i^{th} experimental and measured outputs, which are considered to be a considerable degree of correlation between the actual and estimated values when the R^2 number approaches 1 [23]. Correlation coefficients are scaled so that they range from −1 to +1, where 0 denotes the absence of a linear or monotonic association. As the coefficient approaches an absolute value of 1, the relationship becomes stronger and eventually resembles a straight line. [23]. Second, RMSE is preferred because RMSE = 0 represents the least errors and greater residual errors are handled more delicately. On the other hand, there are situations where RMSE is not the best option for obtaining a greater level of accuracy; in these situations, MAE is used because it works with both smooth and continuous data. A more reliable model performance and proper calibration are also indicated by higher R-values and lower RMSE and MAE values.

3. Results

The followings are the experimental setup that follows the study. From 252 total records and 8 attributes, the data are divided into training and testing sets at 80% and 20%, respectively. The author uses random forest, decision tree, linear regression, and artificial neural network algorithms to predict the CBR. Mean squared error, root mean square error, mean absolute error, and relative error are used to assess the effectiveness of the algorithms. The output of the method is shown in Table 4.

As it is shown in the above table, random forest is the minimum value of MAE, RMSE, and MSE. This means that it has the smallest error to predict the CBR, and there is a highest value of R^2 which indicates that there is a good relationship between parameters. The following Figure 6 depicts the error value of the algorithms.

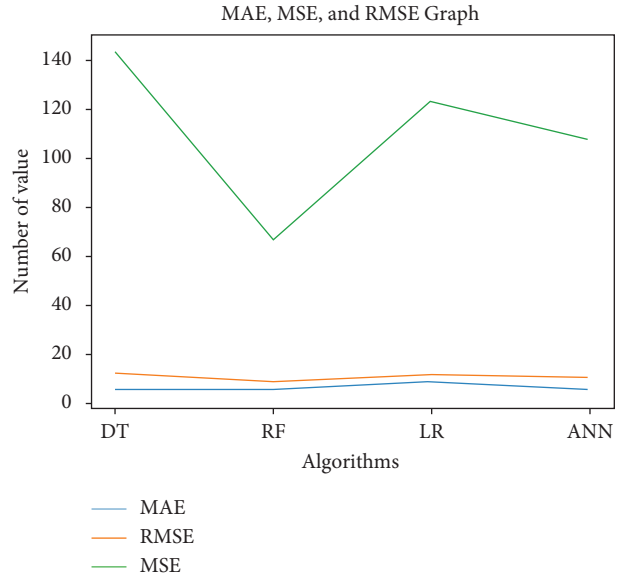

Figure 7: Comparisons of MAE, MSE, and RMSE of the algorithms.

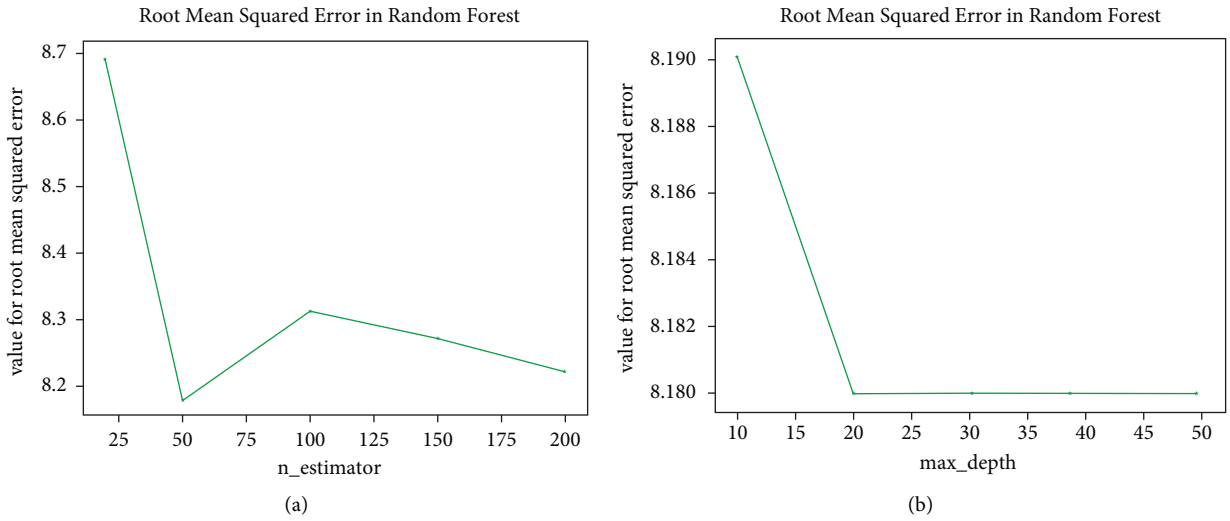

Figure 8: RMSE of random forest algorithm: (a) n_estimator and (b) max_depth value.

From the above figure, RF yields a lower mean absolute error, mean squared error, and root mean square error than the others which is a lower error rate in prediction. Besides, it scores the greater value in relative error (coefficient of determination). The highest value in relative error implies that there is a highest correlation or relationship among variables.

4. Discussion

Now, it is a time to discuss the result. Generally, in this study, the author employed random forest, decision tree, linear regression, and artificial neural network algorithms in 80% training and 20% testing data. From this, RF predicts well, and there is a minimum error rate when it compares to the others in the value of MAE, MSE, and RMSE. In addition to this, RF scores a good value in R^2 or coefficient of determination than the other which indicates that the relationship between attributes becomes stronger. It is not possible to conclude that one algorithm is always fit in different studies and data. The dataset and the parameters have their own influence to achieve the result. For this purpose, the author tried to identify the best algorithm for the collected data and identify the determinant factor which is more important to predict the CBR value. Therefore, random forest is the final selected algorithm for the study to predict CBR. RMSE is a cost function, and it is used to calculate the difference between the actual target values in the testing set and the values predicted by the model. The RMSE comparison of the algorithms is described in the following Figure 7.

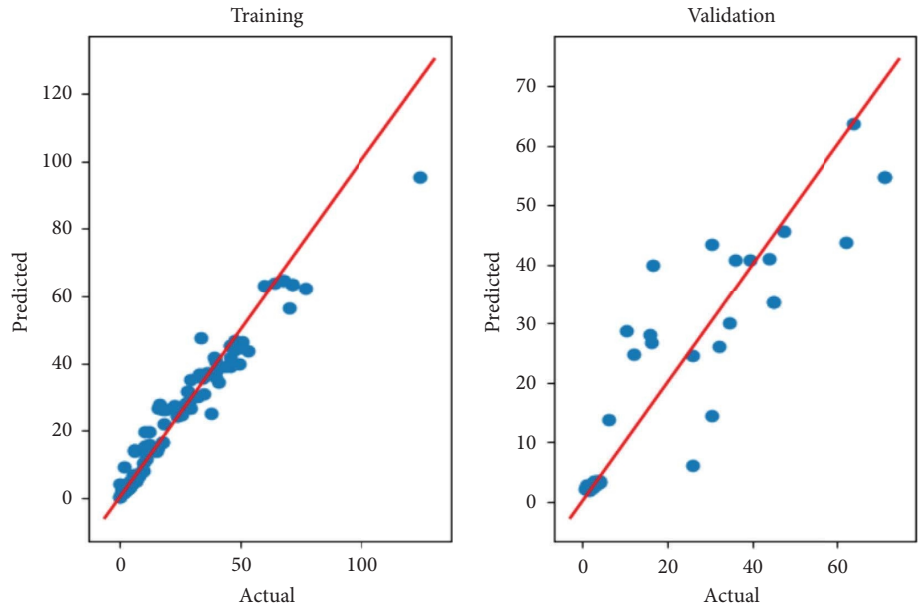

Figure 9: The actual and predicted values of training and testing sets using RF.

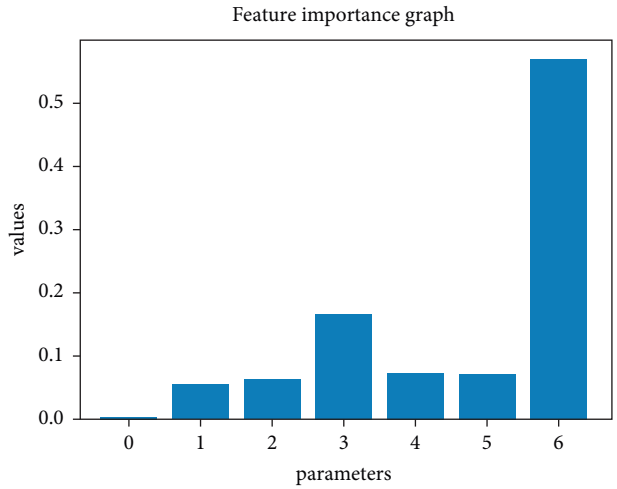

Figure 10: Feature importance graph. 0 = soil classification, 1 = liquid limit, 2 = plastic limit, 3 = plasticity index, 4 = maximum dry density, 5 = optimum moisture content, and 6 = swell.

As it is shown from the above figure, the random forest algorithm scores a better result that the other which is a smaller value in MSE (mean squared error), MAE (mean absolute error), and RMSE (root mean square error). In the random forest regressor algorithm, the author tried to change the parameters such as n_estimators value as 20, 50, 100, 150, and 200 and max_depth value as 10, 20, 30, and 50 to get the better result. The n_estimators and max_depth values variation in the random forest algorithm for RMSE is depicted in the following Figure 8.

As it can see from the above figure, the n_estimators value is a smaller error at 50. Besides this, in max_depth, there is a smaller error in 20, and there is no difference in errors after this point.

The following Figure 9 displays the actual and predicted values in training and testing sets.

As it can be seen from the above table, the actual values are plotted with red color, and the predicted values are plotted with green color. In addition to this, to identify the most influential parameter from the input, the author used the feature importance technique. From this, the three ranked important features that are used to predict the CBR are swell, PI, and MDD depicted in the following Figure 10. In the figure, swell is the first important value to predict the CBR. Next to this, plastic index (PI) and maximum dry density (MDD) are also the important one in their order.

5. Conclusions

In this study, the CBR of soils was predicted using the random forest, decision tree, linear regression, and artificial neural network algorithms, which were trained and built-in

80% training and 20% testing set. The models' input variables include SC, LL, PL, PI, OMC, MDD, swell, and CBR. The author tried to employ these algorithms and conduct the study to predict the CBR values in the specified soil. The key findings of the study are to assess the methods in this dataset and identify the good algorithm that predicts well in this dataset. From this, the RF algorithm achieved a coefficient of determination (R^2) value of 0.84, which is higher than the prediction algorithms, which means that there is a stronger relationship between attributes. In addition to this, it also yields a minimum error in MAE, MSE, and RMSE values than the other algorithm. The dataset and methods used will affect how accurate the predictions are. In addition, as part of future studies, the researchers may use RF, DT, LR, and ANN models which can be further improved by utilizing more input data, and the outcomes can be compared to those of other ML models.

Data Availability

All data are available within the article.

Conflicts of Interest

All authors declare that there are no conflicts of interest.

Authors' Contributions

Semachew Molla Kassa and Betelhem Zewdu Wubineh are contributed equally.

Acknowledgments

The authors would like to acknowledge the anonymous reviewers to give constructive comments and suggestions and also acknowledge Ethiopian Road Administration Laboratories.

References

[1] S. M. Lakshmi, S. Subramanian, M. Lalithambikhai, A. M. Vela, and M. Ashni, "Evaluation of soaked and unsoaked CBR values of soil based on the compaction characteristics," *Malaysian Journal of Civil Engineering*, vol. 28, no. 2, 2016.

[2] B. Gunaydin and O. Gunaydin, "Estimation of California bearing ratio by using soft computing systems," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6381–6391, 2011.

[3] M. Vaquero Barnadas, *Machine Learning Applied to Crime Prediction*, Universitat Politècnica de Catalunya, Barcelona, Spain, 2016.

[4] S. Taha, S. El-Badawy, A. Gabr, A. Azam, and U. Shahdah, "Modeling of California bearing ratio using basic engineering properties," in *Proceedings of the 8th International Engineering Conference*, Sharm Al-Sheikh, Egypt, November 2015.

[5] A. K. Sabat, "Prediction of California bearing ratio of a stabilized expansive soil using artificial neural network and support vector machine," *Electronic Journal of Geotechnical Engineering*, vol. 20, no. 3, pp. 981–991, 2015.

[6] D. Q. Vu, D. D. Nguyen, Q.-A. T. Bui, D. K. Trong, I. Prakash, and B. T. Pham, "Estimation of California bearing ratio of soils using random forest based machine learning," *Journal of Science and Transport Technology*, vol. 1, pp. 48–61, 2021.

[7] D. K. Trong, B. T. Pham, F. E. Jalal et al., "On random subspace optimization-based hybrid computing models predicting the California bearing ratio of soils," *Materials*, vol. 14, no. 21, p. 6516, 2021.

[8] A. Bardhan, C. Gokceoglu, A. Burman, P. Samui, and P. G. Asteris, "Efficient computational techniques for predicting the California bearing ratio of soil in soaked conditions," *Engineering Geology*, vol. 291, Article ID 106239, 2021.

[9] L. S. Ho and V. Q. Tran, "Machine learning approach for predicting and evaluating California bearing ratio of stabilized soil containing industrial waste," *Journal of Cleaner Production*, vol. 370, Article ID 133587, 2022.

[10] A. R. Patel and A. Patel, "Utilization of support vector models and gene expression programming for soil strength modeling," *Arabian Journal for Science and Engineering*, vol. 45, no. 5, pp. 4301–4319, 2020.

[11] T. Taskiran, "Prediction of California bearing ratio (CBR) of fine grained soils by AI methods," *Advances in Engineering Software*, vol. 41, no. 6, pp. 886–892, 2010.

[12] M. Arsyad, I. B. Mochtar, and N. E. Mochtar, "Analysis of settlement of the road with full scale geotextile reinforcement on the very soft soil (case study in tapin regency, south kalimantan)," in *MATEC Web of Conferences*vol. 280, EDP Sciences, Article ID 03012, 2019.

[13] J. Connelly, W. Jensen, and P. Harmon, *Proctor Compaction Testing*, The Constructor Building Ideas, Chennai, India, 2008.

[14] A T89, "Determining the plastic limit and plasticity index of soils," *Standard Specifications for Transportation Materials and Methods of Sampling*, The Constructor Building Ideas, Chennai, India, 2014.

[15] Astm, "3282/AASHTO M 145," *Practice for Classification of Soils and Soil-Aggregate Mixtures for Highway Construction Purposes*, ASTM International, Pennsylvania, USA.

[16] V. R. Schaefer, D. J. White, H. Ceylan, and L. J. Stevens, "Design guide for improved quality of roadway subgrades and subbases," *Iowa Highway Research Board (IHRB Project TR-525)*, vol. 7, pp. 8–72, 2008.

[17] H. F. Southgate, *Comparison of Rigid Pavement Thickness Design Systems*, Chennai, India, 1988.

[18] F. Calders and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[19] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, 2017.

[20] S. W. Kwok and C. Carter, "Multiple decision trees," in *Uncertainty in Artificial Intelligence*, vol. 9, pp. 327–335, Elsevier, 1990.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] X. Yan and X. Su, *Linear Regression Analysis: Theory and Computing*, World Scientific, Singapore, 2009.

[23] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *Anesthesia and analgesia*, vol. 126, pp. 1763–1768, 2018.