Obtaining the California Bearing Ratio Prediction via Hybrid Composition of Random Forest

Bensheng Wu[1], Yan Zheng[2]
Fujian Construction and Engineering Group Co., Ltd.; Fuzhou Fujian, 350000 China[1]
Fujian West Coast Architectural Design Institute Co., Ltd; Fuzhou Fujian, 350000 China[2]

***Abstract*—Artificial intelligence algorithms have become much more sophisticated, so the most complex and challenging problems can be solved with them. California Bearing Ratio (CBR) is a time-consuming testing parameter, and univariate and multivariate regression methods are used to address this challenge. Therefore, the CBR value is an essential parameter in indexing the resistance provided by a structure's subterranean formation or foundation soil. CBR is a crucial factor in pavement design. However, its determination in laboratory conditions can be a time-consuming process. This makes it necessary to look for an alternative method to estimate CBR in the soil subgrade, especially the developed layers of the soil. This study has developed one of the machine learning (ML) models, including Random Forest (RF), to predict the CBR. Additionally, some meta-heuristic algorithms have been used for improving the accuracy and optimizing the output of the prediction, consisting of Gold Rush optimizer (GRO), Stochastic Paint optimizer (SPO), and Electrostatic Discharge algorithm (EDA). The results of the hybrid models were compared via some criteria to choose the desired model. SPO had the most desirable performance when coupled with RF compared to other optimizers, exhibiting high R2 and low RMSE.**

Keywords*—California bearing ratio; gold rush optimizer; stochastic paint optimizer; electrostatic discharge algorithm; random forest

I. Introduction

A. Background

The strength of the soil to be used as a subgrade in the pavement is assessed using the California bearing ratio (CBR) value. The CBR test is a crucial field/laboratory test in geotechnical engineering. This is done to evaluate the resistance provided by the subterranean soil layer or the structure's foundation, particularly for earth embankments, road embankments, abutments, and retaining walls. The CBR value can express the strength of the ground. If the CBR value is low, the pavement thickness will be increased, resulting in higher construction cost, while reducing the pavement thickness will decrease the cost [1–3]. CBR tests can be carried out either in the field or the laboratory. In field CBR tests, the assessment is conducted on the ground surface or within an excavated test pit. Conversely, laboratory CBR tests are typically performed on compressed samples placed in a CBR machine [4].

Performing CBR tests in the laboratory requires less labor, but it is time- and energy-intensive. Hence, a method that can accurately predict CBR values in expansive soils with minimal time and effort is often welcomed. The importance of accurately predicting CBR values in soils, particularly in stabilizer-treated expansive soils, cannot be overstated [5,6]. Precise CBR predictions for modified or treated expansive soils ensure the safety and flexibility of pavement design. When utilized as subgrades, numerous approaches have been proposed to forecast the CBR of expansive soils, whether treated or untreated [7–9]. These approaches have been extensively documented in academic sources. Nevertheless, certain CBR prediction models in the literature demonstrate weak correlation coefficients, suggesting that conventional statistical methods make it difficult to generate accurate CBR estimates [10].

B. Related Works

Due to the robustness of CBR models and the ease with which complex computations can be performed, the recommendation is to employ Machine Learning (ML) techniques for constructing CBR. Several published articles used ML approaches such as random forest (RF), multivariate adaptive regression splines (MARS), and gradient boosting machines (GBM) to predict CBR [11–13]. ML techniques have proved to be effective predictive tools in various engineering disciplines and, hence, were utilized to develop models for predicting CBR in improved soils. The development of CBR estimation using classical statistical methods presents a considerable challenge [14–16]. Employing ML methodologies [17–20] to develop CBR models is advisable due to their inherent robustness and capacity to manage intricate computations proficiently.

Stephens [21] examined the performance of current models for specific native soils using data that had been stored. He looked at the links between CBR and different classification characteristics in both basic and multivariate versions and found that these models were typically insufficient. Additionally, the influence of the clay percentage on CBR was reported. In the interim, shrinkage, and grading moduli were proposed as a means of estimating the lowest CBR values for both shrinking and non-shrinking soils. Another technique for determining CBR was offered by the British Highway Agency [22], which used the plasticity index for British soils compacted at natural moisture content and supplied correlations in a tabular manner. Khasawneh's study [23] focuses on optimizing Resilient Modulus Testing for subgrades and predicting California Bearing Ratio (CBR) using advanced soft computing systems. It introduces a method to forecast the resilient modulus, especially for fine-grained soils, and explores the use of artificial intelligence (AI) techniques for CBR estimation. The research discusses relevant studies on estimating CBR from index properties and compaction characteristics of coarse soils, while also highlighting broader AI applications in fields like quantum

computing and structural engineering. Khasawneh integrates soil mechanics with AI to enhance the accuracy and efficiency of soil property predictions in civil engineering. In contrast, Seman's research [24] emphasizes the significant potential of machine learning methods in reducing prediction errors for plastic soils, acknowledging limitations for non-plastic soils. It identifies soil engineering property variability as a key factor affecting prediction accuracy and suggests addressing the shortage of pedotransfer relationships capable of predicting CBR from other soil measurements. Seman explores case-based reasoning methods and underscores the effectiveness of artificial neural networks in handling complex mappings, offering valuable insights for geotechnical engineering.

C. Research Objectives

The objective of the current investigation is to demonstrate the effectiveness of machine learning (ML) methodologies in developing predictive models for the California Bearing Ratio (CBR). This research specifically utilized artificial neural network (ANN) techniques and the Random Forest (RF) model to estimate CBR values. Additionally, various optimization algorithms, including the Gold Rush Optimizer (GRO), Stochastic Paint Optimizer (SPO), and Electrostatic Discharge Algorithm (EDA), were applied to enhance the accuracy and optimize the predictive output of the RF model. The selection of these optimizers was based on their demonstrated effectiveness in previous studies and their compatibility with the RF model, aiming to further improve the predictive performance of the model by leveraging their respective strengths in optimizing complex engineering problems. The performance of these developed models was assessed using specific evaluation criteria to determine the most suitable combination.

D. Research Significance and Contribution

The study significantly enhances the accuracy of predicting the CBR using advanced machine learning techniques, which is crucial for reliable infrastructure design. By offering an efficient alternative to the time-consuming laboratory determination of CBR, it saves both time and resources in geotechnical engineering projects. The integration of RF with optimization algorithms (GRO, SPO, and EDA) highlights the power of artificial intelligence in addressing complex engineering problems, pushing the boundaries of AI applications. Accurate CBR predictions are essential for pavement design and soil subgrade assessment, making the study's findings highly relevant and beneficial to real-world engineering. Additionally, the study establishes new benchmarks for CBR prediction models and introduces a methodology that can be extended to other predictive modeling tasks in engineering and beyond.

E. Research Organization

The introductory part of this study is divided into five main sections: background, literature review, research objectives, research significances and contributions, and research organization. Following this, Section II provides detailed explanations about the description of performance evaluators, the dataset used and concise descriptions of various machine learning techniques, including models and optimization algorithms. Section three covers the comparative results using metrics and different techniques. In Section IV which titled discussion, three subsections discussed about the limitations of the study, potential future works in the field of study, and the comparison between the results of this study and existing studies. In Section V, the study's conclusions are summarized.

II. Materials and Methodology

A. Random Forest (RF)

Random Forest (RF) is a supervised ML method that belongs to the family of non-parametric regression or classification techniques. It combines multiple decision trees to produce the desired output.

For modeling data, assume that the training $Q = \{(X_i.Y_i) \ldots\ldots (X_n.Y_n)\}$ an n number of samples and d-dimensional features here $X_i \in R^n$, and $Y_i \in R$.

Here is a brief description of the RF:

Produce bootstrap samples$(E_i. \ldots . E_K)$ from Q. The training dataset Q is resampled through bootstrapping, which involves randomly selecting samples with replacement. The size of each bootstrapped sample is equivalent to that of the original training dataset. Recently, many researchers often used small sample sizes for bootstrap samples due to their ease of computation.

Grow a decision tree, $T_m(i = 1. \ldots . M)$ from every bootstrap sample Em using the subsequent alteration:

Step 1: The optimal split for every node is found by picking the best option from a subset that is chosen at random of m_{try} predictors, where m_{try} predictors are chosen from the total d predictors.

Step 2: Similar to the pruning technique utilized in Classification and Regression Trees $(CART)$, the decision tree in this study is grown without any pruning, ensuring its seamless growth. The decision tree is so large that it is impossible to split the nodes further.

Step3: Take note that the quantity of trees in the woodland is denoted by M while m_{try} represents the number of input variables or predictors that are randomly chosen. The user defines both M and m_{try} while adjusting the parameters of a random forest algorithm. Repeat steps 1-2 until sufficient Tm has been grown.

Step 4: Use the following formula to forecast the answer for an entirely fresh dataset.

$$y_m^*(x) = \frac{1}{M}\sum_{m=1}^{M} y_m(x) \quad (1)$$

In the context of the random forest model, the prediction of the random forest, denoted as $y_m^*(x)$, is obtained by summing the predictions of each tree (mth tree), denoted as $y_m(x)$, for the input vector x [25]. Fig. 1 displays the RF model flowchart.

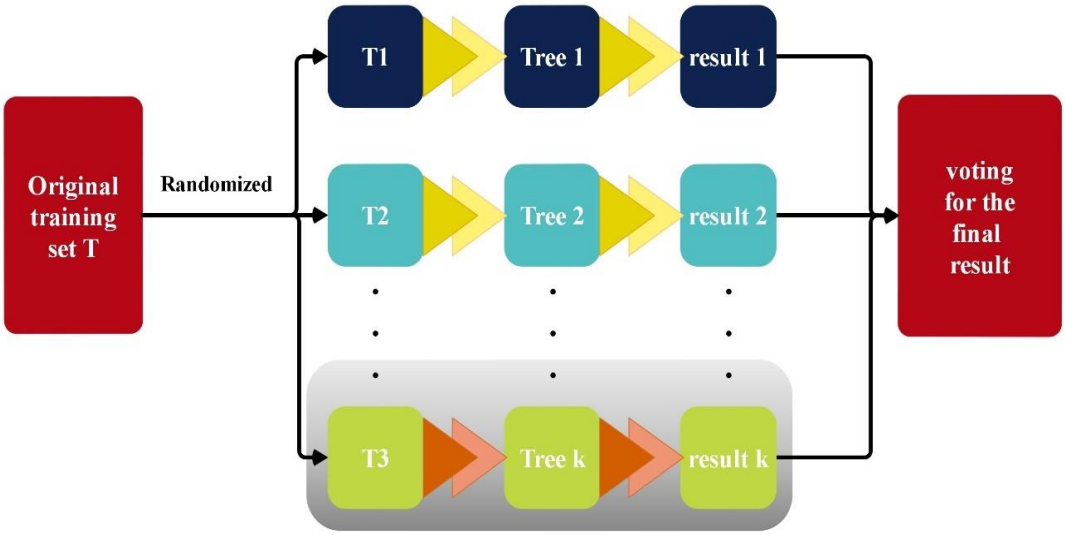


Fig. 1. RF model flowchart.

B. Electrostatic Discharge Algorithm (EDA)

As it is well known, metaheuristic algorithms are mainly inspired by natural behaviors like the feeding action of animals [26]. On the other hand, some people follow the well-known rules of the physical world to do the optimization. As a result, it is always possible to develop a new algorithm that can address some issues better than others. This is the primary motivation for this work. This paper proposes and compares a new metaheuristic algorithm with today's well-known optimization algorithms. Electrostatic Discharge (ESD) events inspire this algorithm and hence are called the Electrostatic Discharge Algorithm ($ESDA$) [27]. The process of optimizing the utilization of $ESDA$ begins with generating a population of individuals. By a fitness value that represents each individual's immunity level, the efficacy of this population is determined. During each implementation or repetition of ESDA, three individuals are randomly selected to undergo discharge. Next, a random value is generated, and depending on its numerical value, one of two scenarios can occur:

Step 1: If the random value is more than 0.5, the discharge is carried out by two personnel at places x1 and x2.

$$x2_{new} = x_2 + 2 \times \beta_1 \times (x_1 - x_2) \quad (2)$$

Step 2: If the random value is smaller than 0.5:

$$x3_{new} = x_3 + 2 \times \beta_1 \times (x_1 - x_3) + 2\beta_3 \times (x_2 - x_3) \quad (3)$$

In the above equations, $x3_{new}$ represents the new position of the individual i and β_i (i = 1, 2, 3) Signify Random Values.

The algorithm then performs extensive checks to ensure that everyone is within bounds. Finally, another check identified those discharged more than three times. This is because the algorithm should consider those individuals as eliminated and replace them with newly generated individuals. This process is repeated for each iteration using fresh individuals, ultimately discovering an optimal solution (i.e., the best solution) [28].

C. Gold Rush Optimizer (GRO)

The optimization problem of damage detection was tackled using the GRO algorithm, a population-based evolutionary algorithm [29]. The GRO algorithm is a population-based evolutionary algorithm with a faster convergence rate than other optimization algorithms. Its primary purpose is to locate areas with gold deposits. Initially, a group of operators is positioned randomly in the search space. Each operator is equipped with a metal detector and is tasked with locating gold deposits. The operators move in groups during each phase and listen to the tone produced by their device. If the tone increases, they stop and investigate the area.

Additionally, they listen to noises made by other devices and observe whether other devices are making loud sounds. During each phase, the group moves to the location with the loudest sound. Finally, the precise location of the gold deposit is determined. The probability of moving towards or away from the loudest sound is described by the parameters α, β, and γ. These parameters are selected within the range of $[0 - 1]$ [29].

Step 1: Initialization

$$location_i^{(0)} = Ib_i + (ub_i - Ib_i) \times rand.i \quad (4)$$
$$= 1.2 \ldots . N$$

Each operator happens to have a position in the search space, as shown in the formula. ub_i and lb_i are upper and lower bounds of the range (search space). $rand$ in the interval [0-1] is a random number. The number of operators is represented by N.

Step 2: Monitoring-Choosing the best locations

A successful operator that discovers the optimal position is referred to as an SOP. Generating an SOP is necessary during this stage. After every iteration, the top 10% of operators should be chosen and documented as part of the SOP.

Step 3: Fitness-distance

The formula is employed to compute the operator that is most likely to extract gold by examining the loudness (rate) of each sound:

$$rate(i) = \frac{D_i}{\rho} \times \frac{sound(highest\ volume) - sound(i)}{(sound(highest\ volume) - sound(lowest\ volume) + \varepsilon)} \quad (5)$$

A small positive number called epsilon (ε) prevents singularities. The coefficients, represented by ρ and Di in Eq. (6), are employed to avoid errors caused by environmental factors. The i and j indicate the two operators' current locations.

$$\rho = 2 - \frac{iter}{max_{iter}} \quad D_i = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} + \cdots \quad (6)$$

Step 4: Think-Decisions-mo

During this stage, each operator chooses an entirely distinct combination of sounds.

$$new\ location(i) = location(i) + md \times [(rate(j) - rate(i)) * (location(j) - location(i)) * ran \quad (7)$$

The coefficients md means move direction determine

$$md = \begin{Bmatrix} +1 \Rightarrow towards\ a\ loudest\ sound? & \alpha > rand \\ -1 \Rightarrow towards\ a\ loudest\ soun? & \alpha < rand \end{Bmatrix} \quad (8)$$

Step 5: Correct locate

If the position obtained from Eq. (7) does not satisfy the problem's constraints, Eq. (8) produces fresh positions. β and γ coefficients are chosen as 0 < β < γ <

$$new\ location(i) = \begin{Bmatrix} choose\ a\ new\ location & rand < \beta \\ select\ a\ new\ location\ randomly & \beta < rand < \gamma \\ do\ not\ move & \gamma < rand \end{Bmatrix} \quad (9)$$

Step 6: Finally, steps 4-6 are repeated in a loop until one of the following exit situations is met:

1) Maximum number of attempts

2) The optimal location did not exhibit any noticeable alteration.

3) The difference between the value of the SOP function and the achieved optimal solution is within the expected threshold. Parameters within the range of [0-1] are chosen.

4) Suppose the disparity between the objective values of the most excellent and poorest positions is lower than the designated accuracy. The GRO's flowchart is displayed in Fig. 2.

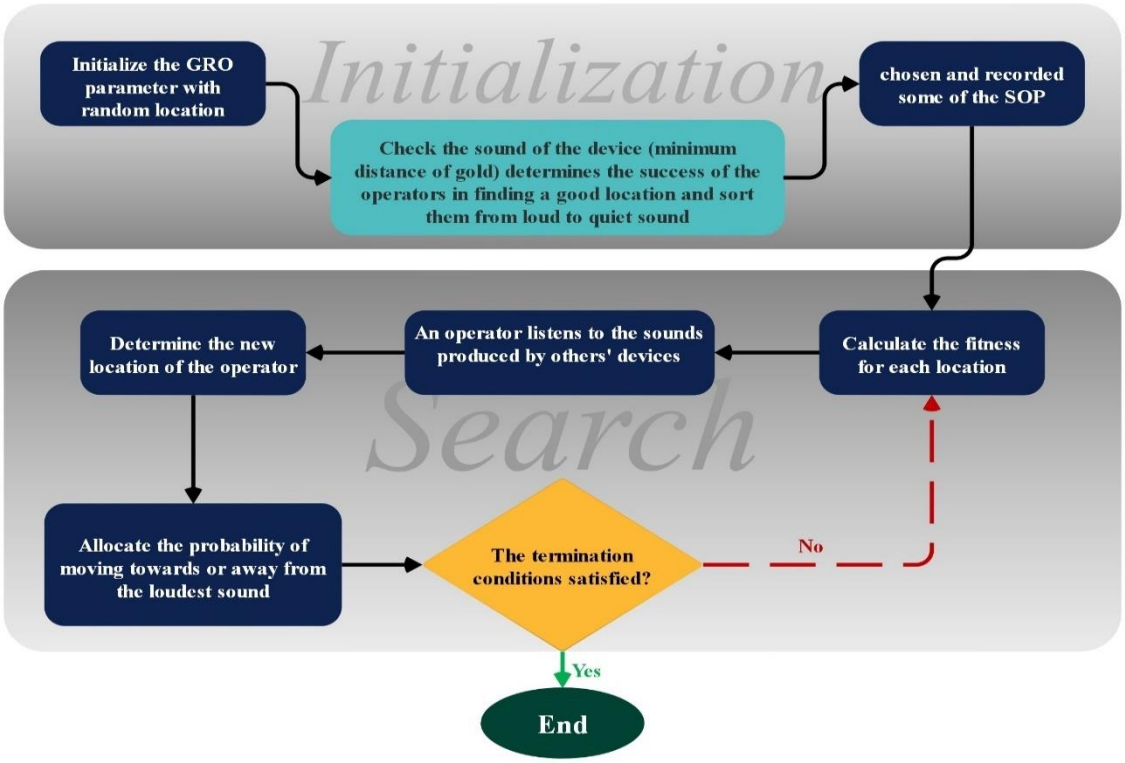


Fig. 2. GRO's flowchart.

D. Stochastic Paint Optimizer (SPO)

This section proposes a novel meta-heuristic algorithm, namely the Stochastic Paint Optimizer (SPO), based on the principles governing the use of colors in paint. The canvas serves as the defined search space wherein solutions, represented by a set of design variables involving certain colors, are considered paint strokes to produce the final output. The aesthetic value of various paints is appraised and categorized in ascending order based on their respective beauty index, representing the objective function values. Adding any fresh hue to a canvas contributes significantly to the overall interpretation of the artwork. As such, each hue is assigned a corresponding grade or value based on the hierarchical classification of colors in the color wheel, with primary colors being deemed most superior, followed by secondary colors as good, and tertiary

colors as inferior. Due to the equal categories, including parameters in the algorithm is deemed unnecessary. This algorithm can produce the most optimal pigments or solutions using the provided combination techniques for color mixing.

Step 1: Initialization

For an nc-dimensional search object, the selection of initial colors for all paints is made randomly.

$$C_{i.0} = C_{min} + rand \times (C_{max} - C_{min}).\, i \quad (10)$$
$$= 1.2.3 \ldots . nc$$

where, $C_{i.0}$ is the initial color of i the paint. C_{min} and C_{max} are the $lower$ and $upper$ limits of the design variable i, $rand$ is a random number with its range $[0, 1]$, and nc is the number of variables or colors. It is noteworthy that combining all colors produces a paint that serves as a solution or design for optimization problems. Subsequently, the objective function is assessed for each painting. Thus, the aesthetic quality of each painting is elucidated.

Step 2: Evaluation, Sorting, and Clustering

The paints are methodically arranged in ascending order concerning their corresponding objective function, thus serving as a direct outcome of the problem. Ultimately, these entities are grouped into $three$ equal categories, including primary (the most favorable), secondary (favorable), and tertiary (the least favorable).

Step 3: Utilizing Combination Techniques

This phase synthesizes novel paint formulations using four distinct combination methodologies.

Step 4: Evaluating and Updating.

The new beauty index of the paints is assessed, and if it is superior to the previous index, the old paint is substituted with the new one.

Step 5: Checking Termination

Upon the completion of a series of iterations, the cycle of optimization is considered finished. If the established criteria are not satisfied, a new Phase 2 process will be arranged. However, if the criteria are met, the process will be terminated, and the optimal solution will be reported [30]. Fig. 3 displays the SPO flowchart.

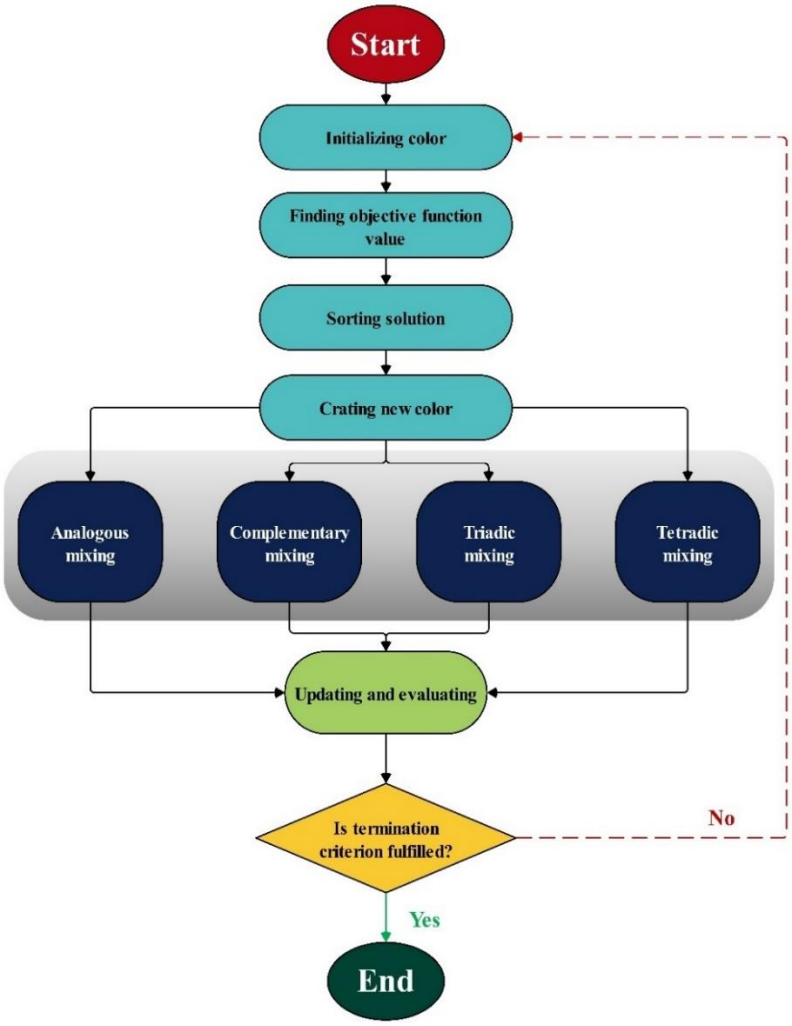


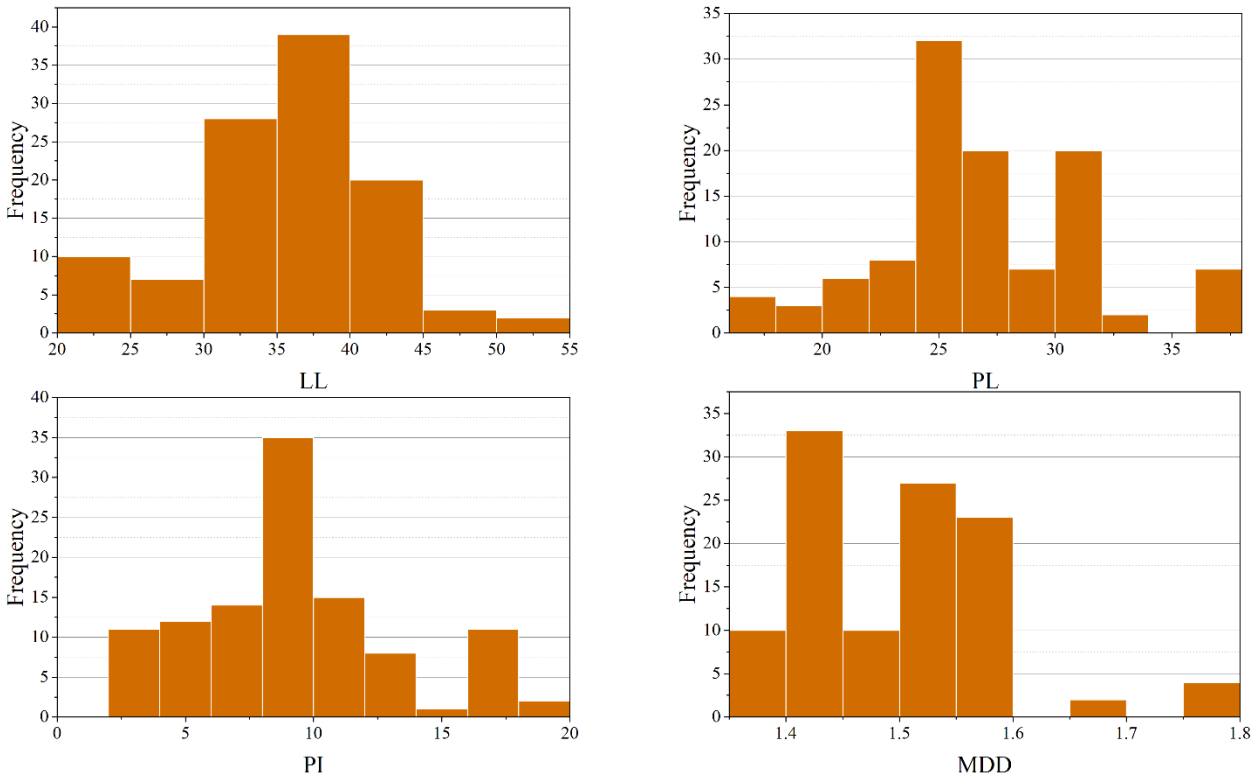
Fig. 3. SPO flowchart.

E. Data Gathering

The goal of this work is to accurately estimate the California Bearing Ratio (CBR), a crucial variable in civil engineering projects, using a novel ML (ML) technique. In order to do this, the dataset is carefully split into three stages: a significant 70% is set aside for training, and the remaining 30% is put aside for testing. The visual representation of the input and output variables is shown in Fig. 4, and Table I provides a thorough summary of the statistical properties for the major contributing factors, such as the crucial CBR, Silt and Dust Amount as a Percentage ($SDA\%$), Quartz and Dirt Percentage ($QD\%$), Plastic Limit (PL), Plasticity Index (PI), Maximum Dry Density (MDD), Optimum Moisture Content (OMC), and Silt and Dust Amount as a Percentage ($SDA\%$). This study utilizes the Random Forest (RF) model to enhance the design and construction of CBR in the broader infrastructure landscape, overcoming challenges in empirical data acquisition. The proposed framework for civil engineering predicts the strength of concrete by analyzing a vast CBR dataset. This comprehensive approach offers valuable insights, enabling informed decisions and ensuring the robustness of structural designs in civil engineering projects [31–33].

TABLE I. THE STATISTIC PROPERTIES OF THE INPUT VARIABLE OF CBR

Variables		Indicators			
	Category	*Min*	*Max*	*Avg*	*St.Dev.*
LL	Input	21.200	52.100	35.846	6.154
PL	Input	17.900	37.200	26.683	4.281
PI	Input	2.100	19.500	9.162	4.115
MDD	Input	1.365	1.777	1.493	0.088
OMC	Input	18.900	29.500	24.143	2.427
SDA (%)	Input	0.000	20.000	10.661	7.155
QD (%)	Input	0.000	20.000	10.642	8.196
OPC (%)	Input	2.000	8.000	4.945	2.380
CBR (%)	Output	19.690	66.750	39.959	10.867



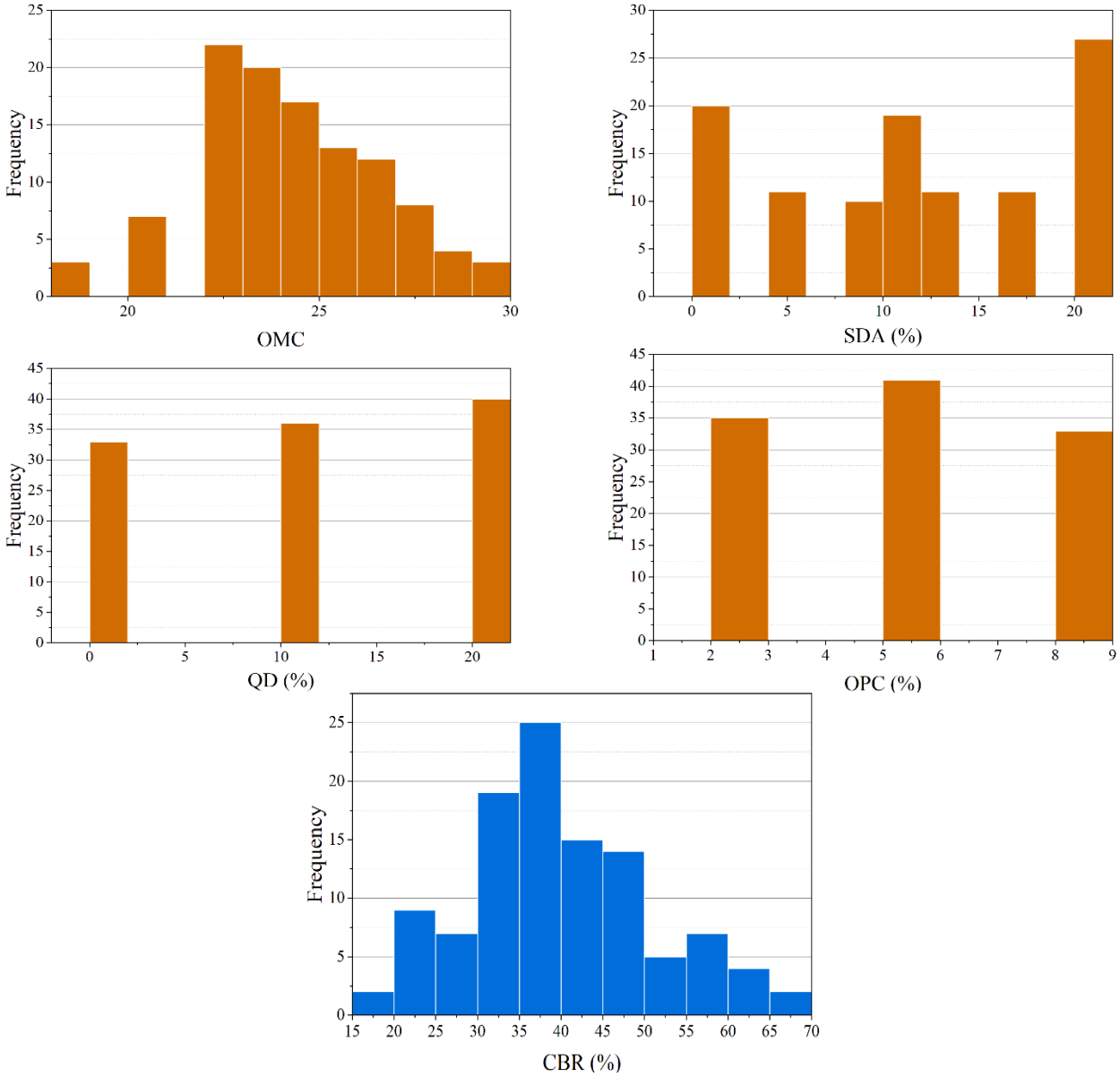


Fig. 4. The histograms plot for input and output.

F. Performance Evaluation Methods

Concrete estimative evaluations typically incorporate five commonly utilized performance indicators. Moreover, their utilization was employed to evaluate the ML approach presented in this manuscript. The correlation coefficient (R^2) provides a quantitative metric of the extent to which the explanatory variables can successfully account for the variable's observed response. This statement assesses the model's aptness for the intended purpose by evaluating the degree to which it aligns with the data or phenomena under consideration. The estimation capacity of the model under consideration can be adequately assessed by observing an elevated R^2 coefficient value. The Root Mean Squared Error ($RMSE$) is a statistical measure utilized to assess the accuracy of a forecast. The $RMSE$ is a statistical measure employed to assess the variance of a response variable, which can be effectively characterized using models. The Mean squared Error (MSE) is a statistical measure that calculates the mean magnitude of errors in the predictions made by a given model. The subsequent section elaborates on the mean absolute percentage error (MAPE) measures as well as the variance account factor (VAF).

$$R^2 = \left(\frac{\sum_{i=1}^{n}(p_i - \bar{p})(r_i - \bar{r})}{\sqrt{[\sum_{i=1}^{n}(p_i - p)^2][\sum_{i=1}^{n}(r_i - \bar{r})^2]}} \right)^2 \quad (11)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - p_i)^2} \quad (12)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(r_i - p_i)^2 \quad (13)$$

$$MAPE = \frac{100}{n}\sum_{i}^{n}\frac{|r_i|}{|p_i|} \quad (14)$$

$$T_{state} = \sqrt{\frac{(n-1)MSE}{RMSE^2 - MSE}} \quad (15)$$

where, $\bar{r}$ and $\bar{p}$ show the averages of the observed and predicted, respectively, where p_i and r_i determine the predicted and observed values. n is the sample number as well.

III. RESULTS

This section delves into model discussion based on specific criteria. To ensure a robust evaluation, the dataset underwent random partitioning, creating distinct training and test sets. The model construction relied on 70% of the training data, while the remaining 30% assessed the built model's reliability. Realistic interconnections among elements were established. Table II presents key findings as follows:

The R^2 metric ranges from 0.9959 ($RFSPS_{test}$) to 0.9773 ($RFED_{train}$). Similarly, RMSE values vary from 7.623 ($RFED_{train}$) to 2.1546 ($RFSP_{test}$). Notably, the most favorable MSE, at 9.260, corresponds to $RFGR_{test}$, while the least desirable, reaching 58.110, aligns with $RFED_{train}$. In terms of MAPE, RFSPtest excels at 4.0378 while RFEDtrain lags at 9.7295. Testate values highlight $RFGR_{test}$'s suitability at 0.5758, contrasting starkly with $RFED_{train}$'s 2.1246, indicating inferior performance. The outcomes collectively suggest thorough training for all models, with minor exceptions where test data performance deviates.

Regarding model parameters, RFED experiences a general decrease, though initially high values render it unsuitable as a predictive model. Moreover, RFGR shows marginal increases, barring R^2, with other parameters decreasing. This demonstrates its suitability and high predictive accuracy. In essence, the assessment underscores the models' varied performances. While RFED's parameter reduction might suggest improvement, its unsuitably high initial values limit its predictive efficacy.

Conversely, the slight parameter fluctuations within RFGR, coupled with predominantly decreasing trends, affirm its suitability and accuracy in forecasting. Overall, the evaluation emphasizes the nuanced performance differences among models. RFED's parameter reductions hint at enhancement, yet its unsuitably high initial values limit its predictive prowess. Conversely, RFGR's minor parameter fluctuations alongside predominantly decreasing trends validate its suitability and precision in forecasting.

TABLE II. THE ACHIEVING RESULTS OF PRESENTED MODELS

Models	**RFGR**		**RFSP**		**RFED**	
Section	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
RMSE	4.754	3.043	3.189	2.155	7.62	5.357
R2	0.9876	0.9930	0.9935	0.9959	0.9773	0.9814
MSE	22.60	9.260	10.17	4.643	58.11	28.70
MAPE	5.882	5.272	4.551	4.038	9.730	8.723
Tstate	1.633	0.576	1.071	1.378	2.125	0.115

In Fig. 5, this study visually illustrates a Scatter plot depicting predicted and measured CBR values across distinct testing and training phases. The shape's determination relies on two evaluative metrics: R^2 and RMSE. R^2 assesses the likelihood within a given sequence, while RMSE gauges data dispersion or density. X = Y coordinates form the central line, and the linear regression underwent two phases of experimentation and evaluation. The angle divergence between these lines measures model effectiveness. The RFSP model displays lower RMSE and higher R^2 in training than in testing.

Consequently, there is minimal variance between the linear fit angle and the line, indicating less scatter in training compared to testing. In contrast, the RFED model shares similarities with the RFSP model but exhibits considerably high RMSE and R^2 values, rendering it unsuitable for forecasting. Conversely, the RFGR model showcases favorable RMSE and R^2 values. Notably, this model displays a higher degree of data point dispersion compared to the other two models. To sum up, Fig. 5 visually demonstrates the Scatter plot portraying predicted and measured CBR values in distinct testing and training phases. R^2 and RMSE serve as evaluative metrics, depicting model effectiveness and data dispersion. While RFSP shows promising results in training compared to testing, RFED's high RMSE and R^2 values make it unsuitable for forecasting. Conversely, RFGR exhibits favorable RMSE and R^2 but displays a higher data point dispersion compared to the other models.

Fig. 6 compares the anticipated and observed CBR values during two distinct stages of experimentation, namely training and testing. In an optimal scenario, the anticipated and observed conduct exhibit comparable conformity. Both RFGR and RFSP models have relatively similar performance, and the max difference CBR for both models is equal to 50, but the RFED model has a relatively significant difference with the other two models, while the max for this model is equal to 15. The well-known RFSO and RFGR models can be generally concluded to be less accurate than the combined model of the two phases.

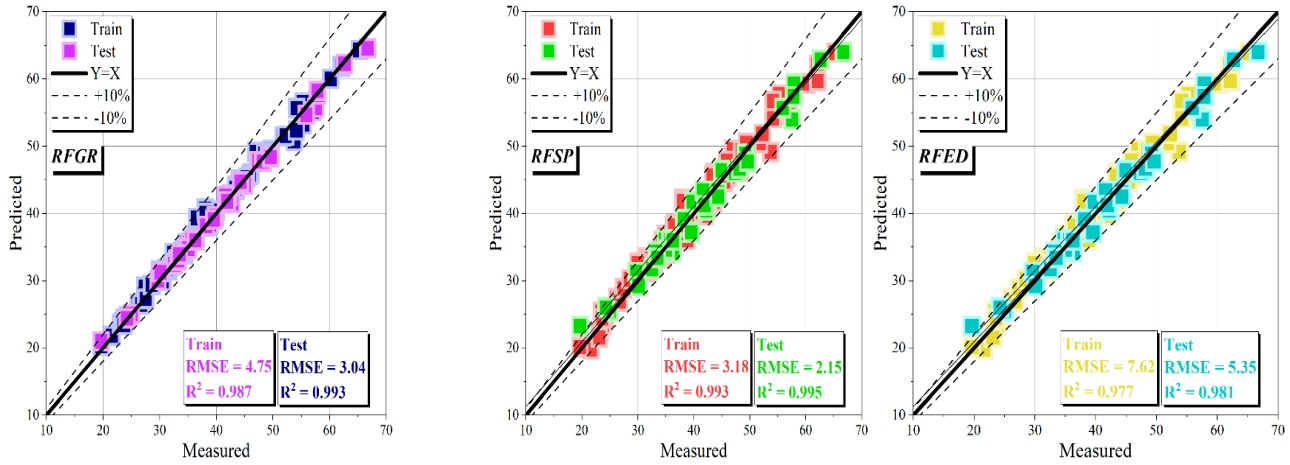
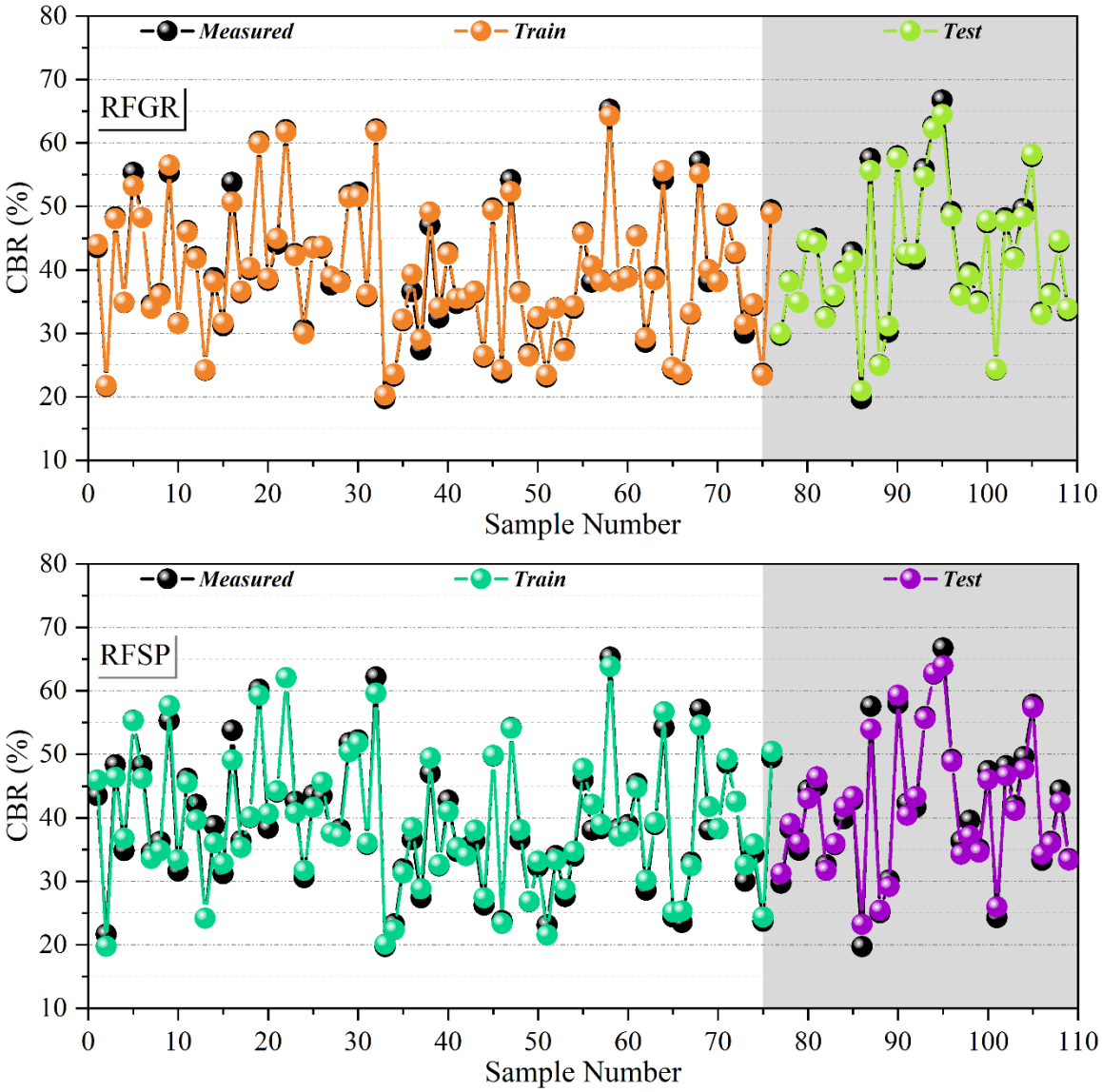


Fig. 5. Scatter plot for correlation between the predicted and measured CBR.



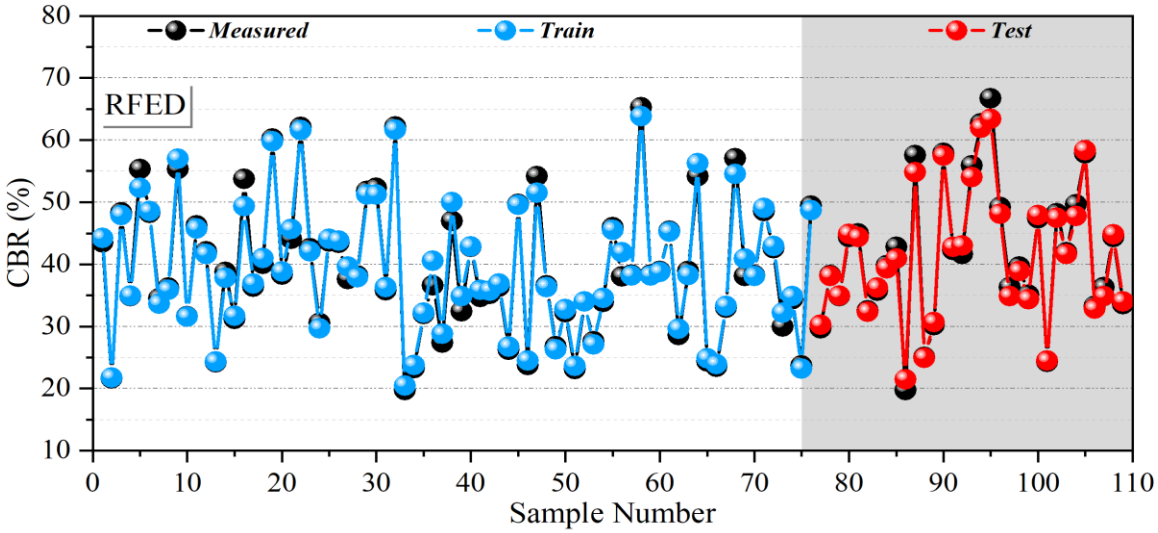
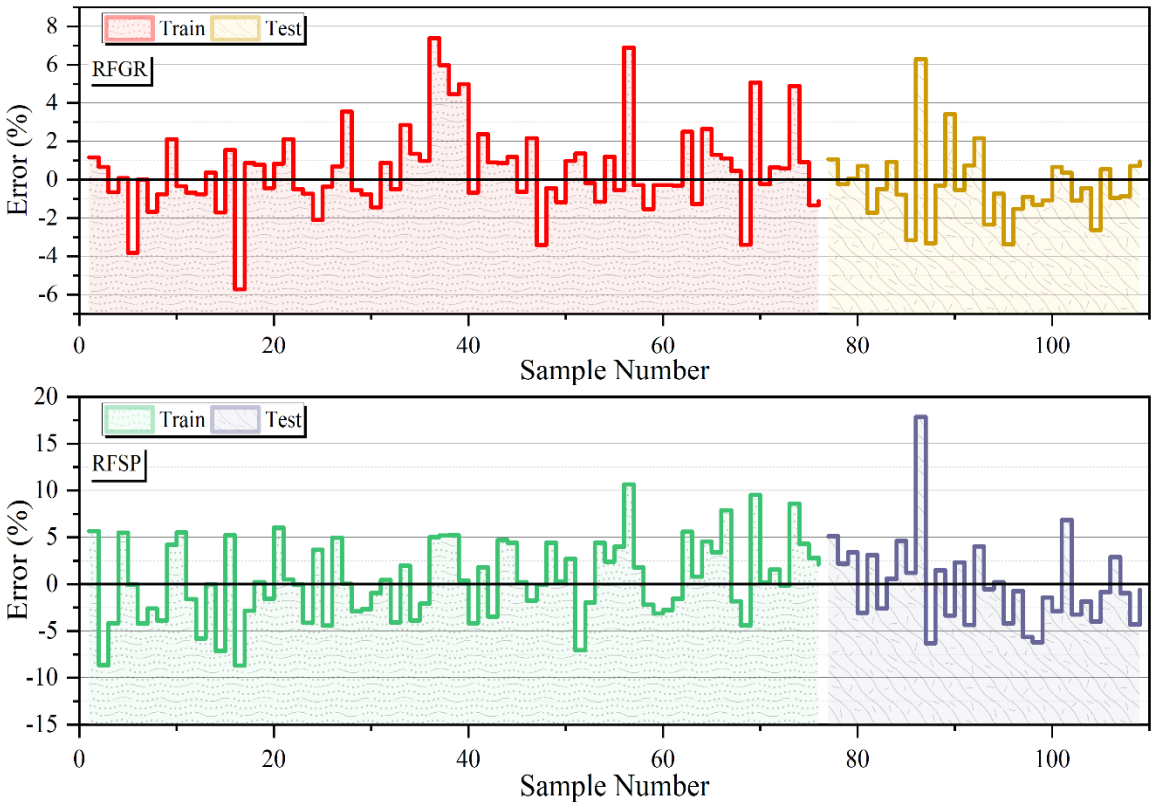


Fig. 6. Comparison between predicted and measured CBR.

Fig. 7 shows the percentage of errors observed during both the training and test stages. As indicated, most samples, specifically 70%, were associated with the training segment, while the remaining 30% were attributed to the testing component. In the RFGR model, the high error percentage was 26%, and the lowest was 11%. These numbers have been reduced to -2% and 12% for the test data. For the RFSP model, the test data has decreased significantly, from -9 to 29, which has decreased to -9 and 12 for the test data. However, the RFED model did not have any special change; the highest error data was 36%, and the lowest was -19%, which decreased to 26% and -14% after training the model. Finally, the percentage of the numbers obtained for the error is closer to zero, the better the model is trained and more appropriate, such as the models RFGR and RFSP.



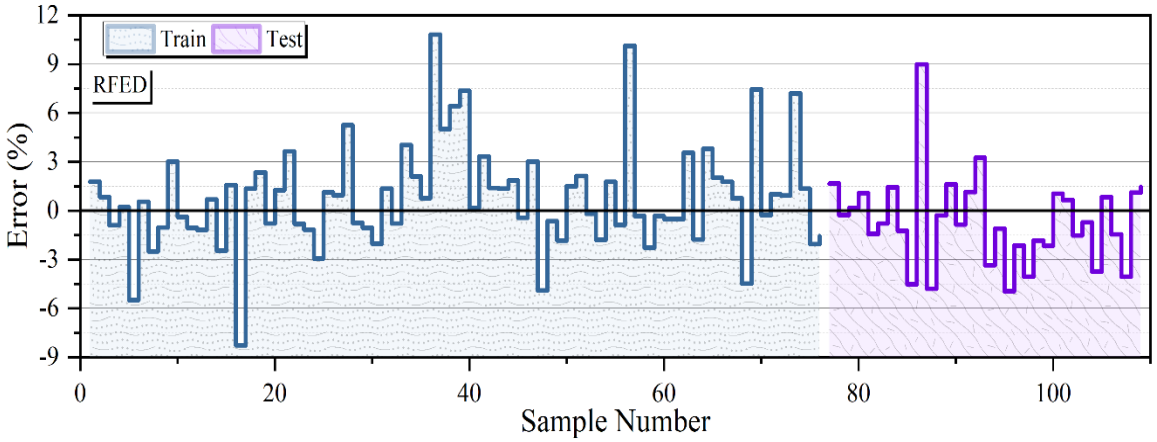


Fig. 7. Percentage of error in the test and training phase.

Fig. 8 displays the half-box plot for the error percentage of the models that are being presented. In the RFGR model, the scatter for error is high and slightly reduces the scatter for test data. In the RFSP model, the dispersion for error is very low compared to the other two models, and it has also decreased due to the test data ranging between -15% and -15%, which indicates the appropriateness and correct training of the model. On the other hand, the RFED model also decreased for the test data, but since it had a very high dispersion from the beginning, it is unsuitable.

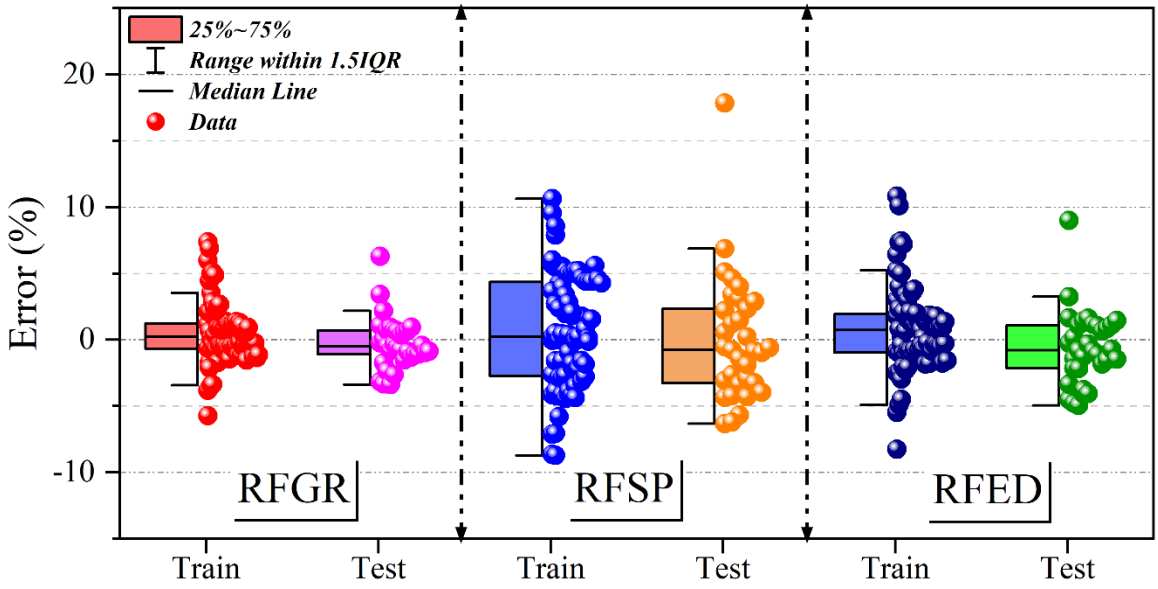


Fig. 8. Half-box plot for the error percentage of the presented models.

IV. Discussion

A. Limitations of the Study

The limitations of this study include potential constraints related to the methodology, data, and scope. Firstly, the reliance on machine learning techniques, while beneficial, may be limited by the quality and quantity of available data for model training and validation. Insufficient or biased data could affect the accuracy and generalizability of the predictive models developed. Additionally, the scope of the study may focus primarily on specific soil types, geographic regions, or pavement conditions, limiting the applicability of the findings to broader contexts. Furthermore, the complexity of integrating multiple meta-heuristic algorithms into hybrid models may introduce challenges in model interpretation, implementation, and computational efficiency. Finally, the study may not account for all potential factors influencing CBR prediction accuracy, such as variations in testing protocols, environmental conditions, or pavement maintenance practices, thus warranting further investigation and refinement in future research endeavors.

B. Potential Future Works

Potential future works in this area could encompass several avenues for advancement. Firstly, there's the opportunity for further refinement and optimization of hybrid models by exploring additional meta-heuristic algorithms or fine-tuning the parameters of existing ones to enhance predictive accuracy. Secondly, the integration of advanced machine learning techniques beyond random forests, such as deep learning or

ensemble methods, could be explored to improve the accuracy and robustness of CBR prediction models. Additionally, researchers could investigate the inclusion of additional input variables, such as environmental factors, traffic loads, or pavement materials, to broaden the predictive capabilities of the models and capture a wider range of influencing factors. Long-term monitoring of pavement performance using the developed models could also be conducted to assess their reliability over time and to update the models based on new data and insights gained from ongoing monitoring activities. Finally, there's the potential for integrating the CBR prediction models into decision support systems for pavement design and management, providing engineers and decision-makers with valuable insights and recommendations for optimizing pavement performance and longevity.

C. Compare the Results of the Present Study and Previous Studies

Numerous studies have been conducted on CBR prediction. Notably, Nawaz et al. [34] utilized a Gene Expression Programming (GEP) model, Khan et al. [35] employed Gaussian Process Regression (GPR), and Bhatt et al. [36] implemented an Artificial Neural Network (ANN) for their predictions. Of the methods summarized in Table III, the ANN model stood out, delivering exceptional results with an R^2 value of 0.99 and an RMSE of 10.01 in the study by Nawaz et al. [34]. In this study, the primary framework utilized was the Random Forest (RF) model, which was augmented through hybridization with three optimization algorithms: the Gold Rush Optimizer (GRO), the Stochastic Paint Optimizer (SPO), and the Electrostatic Discharge Algorithm (EDA). Upon analyzing the results, the integration of the SPO with the RF model exhibited outstanding performance, achieving an R^2 value of 0.9959 and an RMSE of 2.155. This combination outperformed the other two hybrid models evaluated in this research.

TABLE III. COPARISON BETWEEN THE RESULTS OF PREVIOUS ARTICLES AND PRESENT STUDY

Name	Model	Results	
		RMSE	R^2
Bhatt et. al. [36]	ANN	0.202	0.9895
Nawaz et. al. [34]	GEP	10.01	0.99
Khan et. al. [35]	GPR	0.1609	0.8139
Present study	**RF+SPO**	**2.155**	**0.9959**

V. CONCLUSION

Ensuring the accuracy and reliability of California Bearing Ratio (CBR) predictions is crucial for the establishment and deployment of robust and adaptable pavement systems. Unfortunately, the conventional CBR testing protocol employed to ascertain the CBR of subgrades encounters challenges primarily attributed to the prolonged duration required by the testing methodology. Consequently, there arises a need to explore alternative methodologies to approximate the CBR of expansive soil subgrades, with a particular emphasis on the construction of predictive models. To address the limitations associated with conventional testing procedures, particularly their time-intensive nature, the adoption of machine learning (ML) has emerged as a viable solution. By employing ML techniques, the reliance on traditional, labor-intensive experimentation has been significantly reduced. This paradigm shift not only expedites the CBR prediction process but also opens avenues for more accurate and efficient assessments of subgrade characteristics. The integration of ML in CBR prediction offers a progressive step towards enhancing the effectiveness of pavement development and implementation, promoting both safety and flexibility in infrastructure design. The random forest (RF) model, an ML technique for CBR prediction, was also meant to be introduced in this article. Furthermore, the corresponding model was also combined with three meta-heuristic algorithms to form a hybrid model to increase accuracy, which include the electrostatic discharge algorithm (EDA), the stochastic paint optimizer (SPO), and the gold rush optimizer (GRO). Additionally, the performance of developed hybrid models was assessed by several metrics, including R^2, RMSE, MSE, MAPE, and Tstate. Consequently, the SPO model exhibited optimal performance compared to the other two models in conjunction with RF. Results reveal that RFSP consistently excels across various metrics, exhibiting the lowest RMSE and MSE values, highest R^2 values, and statistically significant Tstate values. This underscores RFSP's superior predictive accuracy and robustness compared to RFGR and RFED. RFED also demonstrates commendable performance, particularly in achieving the lowest MAPE values. In contrast, RFGR exhibits relatively lower performance metrics, suggesting lower accuracy and statistical significance in comparison to the other models.

REFERENCES

[1] Kin MW. California bearing ratio correlation with soil index properties. Master Degree Project, Faculty of Civil Engineering, University Technology Malaysia 2006.

[2] Salehi M, Bayat M, Saadat M, Nasri M. Prediction of unconfined compressive strength and California bearing capacity of cement-or lime-pozzolan-stabilised soil admixed with crushed stone waste. Geomechanics and Geoengineering 2022:1–12.

[3] Yildirim B, Gunaydin O. Estimation of California bearing ratio by using soft computing systems. Expert Syst Appl 2011;38:6381–91.

[4] Kassa SM, Wubineh BZ. Use of Machine Learning to Predict California Bearing Ratio of Soils. Advances in Civil Engineering 2023;2023.

[5] Sabat AK. Prediction of California bearing ratio of a soil stabilized with lime and quarry dust using artificial neural network. Electronic Journal of Geotechnical Engineering 2013;18:3261–72.

[6] González Farias I, Araujo W, Ruiz G. Prediction of California bearing ratio from index properties of soils using parametric and non-parametric models. Geotechnical and Geological Engineering 2018;36:3485–98.

[7] Yildirim B, Gunaydin O. Estimation of California bearing ratio by using soft computing systems. Expert Syst Appl 2011;38:6381–91.

[8] Huang L, Jiang W, Wang Y, Zhu Y, Afzal M. Prediction of long-term compressive strength of concrete with admixtures using hybrid swarm-based algorithms. Smart Struct Syst 2022;29:433–44.

[9] Kassa SM, Wubineh BZ. Use of Machine Learning to Predict California Bearing Ratio of Soils. Advances in Civil Engineering 2023;2023.

[10] Behnam Sedaghat, Tejani GG, Kumar S. Predict the Maximum Dry Density of soil based on Individual and Hybrid Methods of Machine Learning. Advances in Engineering and Intelligence Systems 2023;002. https://doi.org/10.22034/aeis.2023.414188.1129.

[11] Khatti J, Grover KS. Relationship Between Index Properties and CBR of Soil and Prediction of CBR. Indian Geotechnical Conference, Springer; 2021, p. 171–85.

[12] Khasnabis C, Motsch KH, Achu K, Al Jubah K, Brodtkorb S, Chervin P, et al. About the CBR guidelines. Community-Based Rehabilitation: CBR Guidelines 2010.

[13] Tamassoki S, Daud NNN, Wang S, Roshan MJ. CBR of stabilized and reinforced residual soils using experimental, numerical, and machine-learning approaches. Transportation Geotechnics 2023;42:101080. https://doi.org/https://doi.org/10.1016/j.trgeo.2023.101080.

[14] Taskiran Tja. Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. Advances in Engineering Software 2010;41:886–92.

[15] Nagaraju TV, Bahrami A, Prasad CD, Mantena S, Biswal M, Islam MR. Predicting California Bearing Ratio of Lateritic Soils Using Hybrid Machine Learning Technique. Buildings 2023;13:255.

[16] Vu DQ, Nguyen DD, Bui Q-AT, Trong DK, Prakash I, Pham BT. Estimation of California bearing ratio of soils using random forest based machine learning. Journal of Science and Transport Technology 2021:48–61.

[17] González Farias I, Araujo W, Ruiz G. Prediction of California bearing ratio from index properties of soils using parametric and non-parametric models. Geotechnical and Geological Engineering 2018;36:3485–98.

[18] Akbarzadeh MR, Ghafourian H, Anvari A, Pourhanasa R, Nehdi ML. Estimating Compressive Strength of Concrete Using Neural Electromagnetic Field Optimization. Materials 2023;16:4200.

[19] Tavana Amlashi A, Mohammadi Golafshani E, Ebrahimi SA, Behnood A. Estimation of the compressive strength of green concretes containing rice husk ash: a comparison of different machine learning approaches. European Journal of Environmental and Civil Engineering 2023;27:961–83. https://doi.org/10.1080/19648189.2022.2068657.

[20] Khajeh A, Ebrahimi SA, MolaAbasi H, Jamshidi Chenari R, Payan M. Effect of EPS beads in lightening a typical zeolite and cement-treated sand. Bulletin of Engineering Geology and the Environment 2021;80:8615–32. https://doi.org/10.1007/s10064-021-02458-1.

[21] Stephens DJ. Variation of the California bearing ratio in some synthetic clayey soils. Civil Engineering= Siviele Ingenieurswese 1992;1992:379–80.

[22] Kin MW. California bearing ratio correlation with soil index properties. Master Degree Project, Faculty of Civil Engineering, University Technology Malaysia 2006.

[23] Khasawneh MA, Al-Akhrass HI, Rabab'ah SR, Al-sugaier AO. Prediction of California bearing ratio using soil index properties by regression and machine-learning techniques. International Journal of Pavement Research and Technology 2024;17:306–24.

[24] Seman PM. Machine learning approaches to CBR prediction for unsurfaced airfields. Transportation Systems Workshop, 2008.

[25] Ikeagwuani CC. Estimation of modified expansive soil CBR with multivariate adaptive regression splines, random forest and gradient boosting machine. Innovative Infrastructure Solutions 2021;6:199.

[26] Safayenikoo H, Nejati F, Nehdi ML. Indirect Analysis of Concrete Slump Using Different Metaheuristic-Empowered Neural Processors. Sustainability 2022;14:10373.

[27] Masugi M. Multiresolution analysis of electrostatic discharge current from electromagnetic interference aspects. IEEE Trans Electromagn Compat 2003;45:393–403.

[28] Bouchekara HREH. Electrostatic discharge algorithm: a novel nature-inspired optimisation algorithm and its application to worst-case tolerance analysis of an EMC filter. IET Science, Measurement & Technology 2019;13:491–9.

[29] Sarjamei S, Massoudi MS, Esfandi Sarafraz M. Gold Rush Optimization Algorithm. Iran Univ Sci Technol 2021;11:291–327.

[30] Kaveh A, Talatahari S, Khodadadi N. Stochastic paint optimizer: theory and application in civil engineering. Eng Comput 2020:1–32.

[31] Taskiran Tja. Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. Advances in Engineering Software 2010;41:886–92.

[32] Karimiazar J, Sharifi Teshnizi E, Mirzababaei M, Mahdad M, Arjmandzadeh R. California bearing ratio of a reactive clay treated with nano-additives and cement. Journal of Materials in Civil Engineering 2022;34:4021431.

[33] Bardhan A, Gokceoglu C, Burman A, Samui P, Asteris PG. Efficient computational techniques for predicting the California bearing ratio of soil in soaked conditions. Eng Geol 2021;291:106239.

[34] Nawaz MN, Qamar SU, Alshameri B, Nawaz MM, Hassan W, Awan TA. A robust prediction model for evaluation of plastic limit based on sieve# 200 passing material using gene expression programming. PLoS One 2022;17:e0275524.

[35] Khan MHA, Jafri TH, Ud-Din S, Ullah HS, Nawaz MN. Prediction of soil compaction parameters through the development and experimental validation of Gaussian process regression models. Environ Earth Sci 2024;83:129. https://doi.org/10.1007/s12665-024-11433-4.

[36] Bhatt S, Jain PK, Pradesh M. Prediction of California bearing ratio of soils using artificial neural network. Am Int J Res Sci Technol Eng Math 2014;8:156–61.