Cite this article
Ahmad U, Jamil I, Jamal H *et al.* (2025)
Using machine learning paradigm to predict CBR value of treated expansive soil.
Geotechnical Research **12(4)**: 186–201,
https://doi.org/10.1680/jgere.24.00066

Research Article
Paper 2400066
Received 07/11/2024; Accepted 24/07/2025

Geotechnical Research

Using machine learning paradigm to predict CBR value of treated expansive soil

Umair Ahmad
Department of Civil Engineering, University of Engineering and Technology, Peshawar, Pakistan (corresponding author: 16pwciv4636@uetpeshawar.edu.pk)

Irfan Jamil
Department of Civil Engineering Technology, National Skills University, Islamabad, Pakistan

Hamza Jamal
College of Transportation Engineering, Tongji University, Shanghai, China

Muhammad Bilal Khan
Department of Civil Engineering, University of Engineering and Technology, Peshawar, Pakistan

Oussama Accouche
College of Engineering and Technology, American University of the Middle East, Egaila, Kuwait

Marc Azab
College of Engineering and Technology, American University of the Middle East, Egaila, Kuwait

The California bearing ratio (CBR) value is a fundamental property used to characterise the strength of the subgrade in road pavement design. CBR calculation in the laboratory is complex, time-consuming, costly, and requires careful execution. A machine learning approach is used to predict CBR of soil treated with hydrated-lime activated rice husk ash (HARHA). The prediction uses an algorithmic approach on A-7-6 expansive soil treated with HARHA, added from 0.1% to 12% in 0.1% increments. Using this approach, 121 distinct data sets were produced in the laboratory which are used to accomplish the stated goals. The data set includes six input parameters: HARHA, liquid-limit, plastic-limit, optimum moisture content, clayey activity, maximum dry density and one output: CBR value. Various models were used, including artificial neural networks (ANN), support vector machine (SVM), Gaussian process regression (GPR), and random forest (RF). The models' performance was evaluated using statistical measures including coefficient of determination, mean absolute error, root mean square error, relative absolute error, and root relative squared error. The evaluation indicates that the RF model had superior predictive performance followed by ANN, SVM, and GPR model. Moreover, sensitivity analysis shows that maximum dry density is the most influential factor on CBR value.

Keywords: artificial intelligence/artificial neural network/CBR value/California bearing ratio/expansive soils/hydrated-lime activated rice husk ash/Gaussian process regression/machine learning/pavements and roads/support vector machine

Notation

AdaBoost	adaptive boosting
MA_{Error}	mean absolute error
MS_{Error}	mean square error
R^2	coefficient of determination
R1	hydrated-lime activated rice husk ash
R2	liquid limit
R3	plastic limit
R4	optimum moisture content
R5	clay activity
R6	maximum dry density
R7	CBR value
RMS_{Error}	root mean square error
XGBoost	extreme gradient boosting
σ_p	p-standard deviation
σ_q	q-standard deviation

Introduction

The California bearing ratio (CBR) value is a crucial measure used to estimate and anticipate the stiffness of rocks, especially those found in the subgrade. It serves as a significant and decisive design index. It was present during the repeated stress on the foundation layer (Trong *et al*., 2021). Accurately predicting geomaterials' CBR value is necessary to construct durable pavements (Rehman *et al*., 2017). CBR value may be defined numerically as the ratio of the penetration of a standard plunger in compacted soil to the penetration of the same plunger in standard crushed rocks. Typically, we keep it at a constant pace of 1.27 mm each minute with the same penetration level. The object's depth is standardised to either 2.5 or 5 cm (Yildirim and Gunaydin, 2011).

The traditional approach to assessing soil strength and bearing capacity is undergoing a progressive transformation towards this testing (González Farias *et al*., 2018), standardised by ASTM as D-1883 or BS 1377 (Abdulnabi and Abdulrazzaq, 2020). Laboratory tests are undertaken on both thick and natural ground soil samples under wet and unsoaked circumstances. As per the AASHTO 2003 guidelines, the plunger diameter used in CBR value testing is 50 mm. The plunger is applied at a rate of 1.25 mm/min (Mousavi *et al*., 2014). The plunger's design enables it to penetrate a compact soil sample with an ideal moisture content. The schematic view of the CBR apparatus is shown in Figure 1. This testing can be used for in situ CBR value calculations at various levels and locations, including natural ground surfaces (levelled), modified and prepared subgrade levels, and construction sites within the testing pit. The crushed rocks in this testing standard are used as a reference and made consistent for the remaining rock sample CBR value

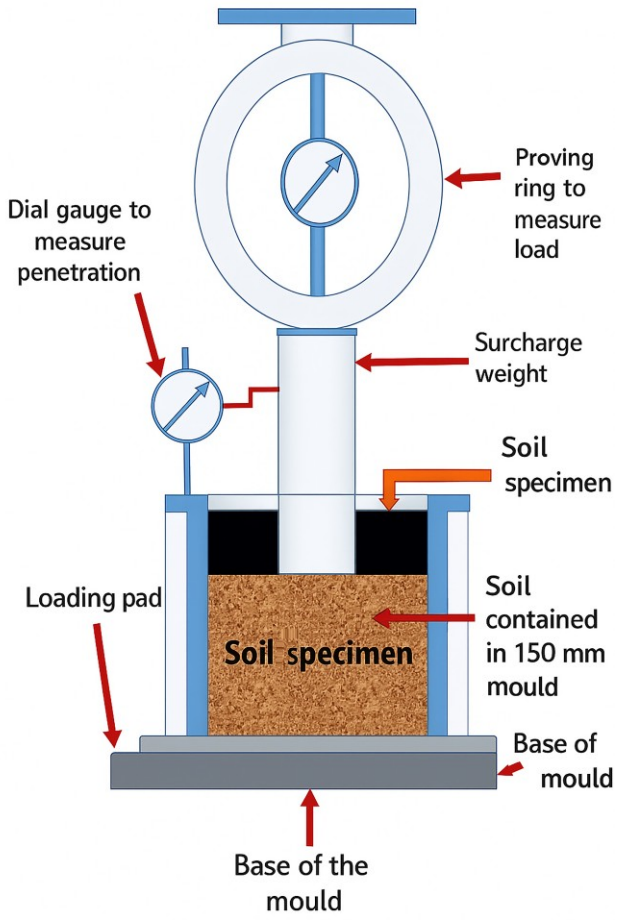

Figure 1. Loading mechanism for CBR value test (www.globalgilson.com)

calculations. Several characteristics and variables may influence the CBR value, including grain size, soil structure, Atterberg limits, and moisture content.

Furthermore, the specimen's dry density might also impact the CBR values (Mishra *et al*., 2010). Multiple studies have shown the influence of soil types and characteristics on CBR values (Taskiran, 2010). Performing these tests in the laboratory is far more intricate and time-consuming. To get more precision, it is necessary to do several iterations of tests and operations (Kişi and Uncuoğlu, 2005).

The CBR value testing is commonly used empirical design technique for flexible pavement design. In essence, a load testing is conducted to assess the durability and operational characteristics of the pavement (Talukdar, 2014). This testing is more often used in developing nations' highway pavement design. The engineering qualities are quantified with respect to strength, stiffness, durability, moisture content, and sensitivity to dimensional changes over time (Onyelowe *et al*., 2020). Obtaining accurate CBR values for design purposes might be problematic because of inadequate soil research and cost constraints. However, conducting the CBR testing in the laboratory is laborious and challenging. Hence, the anticipated outcomes of the created model might be valuable and act as a standard for assessing the accuracy of the CBR data. Therefore, we aim to construct a prediction model that considers the impact of soil index features on CBR value.

Many civil engineering researchers have used soft computing systems, such as artificial neural networks (ANN), support vector machine (SVM), simple linear regressions, and multiple linear regressions, to forecast the CBR value of stabilised soils. Researchers have used the ANN technique to predict the soil CBR value (Taskiran, 2010). ANN are beneficial for analysing and predicting CBR value-stabilised soil. ANNs replicate the neural structure of the human brain (Bhatt *et al*., 2014). The regression coefficient (R-squared) and mean square error (MSE) are the most often used metrics to evaluate the performance of the ANN model. The root mean square error (RMSE) offers a general assessment of the accuracy of the approximation without explicitly referring to individual data points (Bhatt *et al*., 2014). In addition, we have used SVM to forecast and assess the geotechnical properties of stabilised soils. Similarly, we have used SVM to predict the characteristics of stabilised soil, such as compaction, swelling pressure, and foundation settlement, in cohesion-less soil (Islam and Roy, 2020; Rasmin, 2009).

Geotechnical engineers often use the dynamic cone penetrometer (DCP) testing in the field to assess the in situ strength of soils (Paige-Green and Du Plessis, 2009). The process involves continuously striking a metal cone into the ground and measuring the penetration depth after each impact. The DCP testing provides valuable data on the soil's resistance to penetration, allowing for evaluating factors such as compaction and load-bearing capacity. This rapid and efficient testing method is very beneficial for construction and roadworks projects because it can instantly give valuable information on site conditions (Ikechukwu and Mostafa., 2020). The DCP is used in shallow pavement applications according to the Standard Testing method outlined in ASTM D6951-03 (2013) (Thives and Trichês, 2022).

Learning, a subset of computational mechanics, is becoming rapidly utilised in geotechnical engineering to address intricate and nonlinear difficulties, owing to its capacity to successfully model diverse systems (Merghadi *et al*., 2018). The National Cooperative Highway Research Programme has developed two prediction models for soils in the USA. However, both models have limits when accurately forecasting CBR value, especially for soils with a coarse-grained texture. Although the research demonstrates some precision in analysing fine-grained soils, it indicates the need for more precise models in CBR value calculations (Rehman *et al*., 2017; Haupt and Netterberg, 2021; Talukdar, 2014; Erzin and Turkoz, 2016).

In the last ten years, artificial intelligence (AI) methods have been used to predict soil CBR value. These techniques include algorithms like gene expression programming (GEP), SVM, ANN, multilayer perceptron (MLP) neural networks, generalised regression neural networks, and the group method of data handling, which draw inspiration from biological processes. These models have shown potential to enhance the understanding and prediction of certain factors or characteristics, although they may have limits in terms of generalisability (Yildirim and Gunaydin, 2011; Erzin and Turkoz, 2016; Roy and Singh, 2020)

Studies on soil prediction accuracy have shown the development of many models using diverse methodologies. Sabat (2015) obtained an R-squared coefficient of 0.96 with the use of SVM models, while Taha *et al*. (2019) attained an R-squared coefficient of 0.88 by using an ANN. Tenpe and Patel (2020) obtained R-squared values ranging from 0.83 to 0.90 using GEP and SVM methodologies (Al-Busultan *et al*., 2020).

The paradox of predictive capacity in models lies in the fact that models trained on fewer data sets often exhibit better accuracy, maybe owing to the phenomenon of network memory and overfitting. This might result in insufficient model generalisation when applied to more extensive data sets. Overfitting and local minima might lead to incorrect conclusions and limited capacity to apply the findings to other situations. A thorough assessment of predicted accuracy, use of regularisation techniques, and validation across several data sets are essential (Ferentinou and Sakellariou, 2007). The research used multiple regression analysis and ANN system to establish a connection between soil data from three districts in Tamil Nadu and Central Barley Relative CBR values. The model also forecasted the CBR value for combining stone dust and low-quality soil, calculated CBR values for different soil types in Jabalpur city, and established a connection between fine-grained soils (Yildirim and Gunaydin, 2011; Mousavi *et al*., 2014).

Utilising comprehensive training data sets in research yields a more precise depiction of geotechnical concerns, impacting models' dependability. The Unified Soil Classification System categorises different soil types that fulfil the criteria of useful prediction models. Hybrid models that combine optimisation techniques with soft computing models are proposed to locate global minima (Roy and Singh, 2020; Ferentinou and Sakellariou, 2007; Bardhan *et al*., 2021). The researchers used a combination of ANSI and ELM methodologies to evaluate the value of CBR in soils. They obtained exceptional outcomes using several machine-learning approaches, resulting in an R-squared value of 0.9 (Bardhan *et al*., 2021; Onyelowe *et al*., 2021). Raza *et al*. (2021) (used data-driven ML methods to forecast the behaviour of subgrade soil reinforced with geosynthetic materials. They used many models, such as ANN, GPR, random forest (RF), and alternating model trees. Consequently, they achieved exceptional prediction accuracy, making a substantial contribution to the advancement of forecasting methods for the features of geosynthetic reinforced subgrade soil (Onyelowe *et al*., 2021; Ali *et al*., 2022).

The objective of this study is to develop many machine learning (ML) models, such as RF, ANN, SVM, and GPR, together with computational techniques, to predict the CBR value of soil treated with rice husk ash (RHA) and hydrated lime. This research included the following input parameters: maximum dry density, clay activity, optimal moisture content, plastic limit, liquid limit, and hydrated lime-activated rice husk ash (HARHA). In addition, we will use many indicators to evaluate the accuracy of the models. Several performance metrics, such as the coefficient of determination, mean absolute error, RMSE, root absolute error, and relative root squared error will be used to compare these models with previously published models in the literature. The study article starts with an introductory section, followed by a comprehensive examination of the data set, correlation analysis, and a concise overview of pertinent literature pertaining to soft computing methodologies. The third portion of this study paper gives an extensive data analysis and discussion. At the same time, the last section offers a comprehensive review of the whole inquiry, including its conclusions and future directions.

Data collection and preparation

Data set

A research group of Nigeria University performed 121 experiments on treated expansive soil's CBR value. These tests were categorised into two data sets: training and testing (Rasmin, 2009). The whole data set was divided into two parts: the training data set, which made up 70% of the total data set, and the testing set, which accounted for the remaining 30%. The whole data set comprises 121 observations. The training data set shall consist of 85 observations, accounting for 70% of the total, whereas the testing data set consists of 36 observations, representing 30% of the total (Luo *et al*., 2016; Mair *et al*., 2000; Tigga and Garg, 2020). As indicated earlier, the selection method ensures that the training and testing sets consistently provide the same results. Table 1 consistently presents the maximum, lowest, mean, and standard deviation. The following figures illustrate the frequency and normal histogram distributions for the qualities suggested in the above database, which are used in the CBR values modelling for reference. Equal significance was assigned to each of the 121 observations for calculation. The data set was established by researchers who used both treated and untreated soil samples (Onyelowe *et al*., 2021). We synthesised a hybrid geometrical binder called HARHA by incorporating 5% hydrated lime, also known as calcium hydroxide, into a binder composed of RHA. This binder was created by blending hydrated lime with rice husk, an agricultural byproduct, and allowing the combination to undergo a whole day of reaction. Subsequently, we introduced hydrated lime as the alkali activator. Rice husk combustion produces a byproduct called RHA (Onyelowe *et al*., 2019). Subsequently, we used the obtained

Table 1. Comparison of key statistical measures in training and testing data

Data set	Parameter	R1	R2	R3	R4	R5	R6	R7
Training set	Minimum	0	27	12.8	16	0.6	1.25	8
	Maximum	12	66	21	19	2	1.99	44.6
	Average	5.96	48.11	17.20	17.99	1.35	1.68	23.95
	Sta. dev.	3.58	11.74	2.46	0.79	0.40	0.25	11.97
Testing set	Minimum	0.3	27.6	13	16.3	0.62	1.27	8.2
	Maximum	11.9	65.6	20.9	19	1.96	1.99	44.6
	Average	6.10	47.74	17.10	18.12	1.34	1.69	24.12
	Sta. dev.	3.37	11.20	2.33	0.71	0.39	0.23	11.36
Total set	Minimum	0	27	12.8	16	0.6	1.25	8
	Maximum	12	66	21	19	2	1.99	44.6
	Average	6.00	48.00	17.17	18.02	1.35	1.69	23.83
	Sta. dev.	3.51	11.54	2.41	0.77	0.40	0.24	11.82

HARHA to treat the soil in varying quantities, spanning from 0.1% to 12%.

ASTM D1883 standard was used to ascertain the soil's CBR value. There are three occurrences, each consisting of 10, 30, and 65 strikes per layer. The relationship between load and penetration was graphically represented for each individual data point. If concavity is present, then adjust each set of mould data by extending the beginning linear segment of the curve until it intersects the horizontal axis. The penetration was measured at distances of 2.54 and 5.08 mm from the revised origin. Measure the forces required for the plunger to penetrate the earth by 0.1 in., equivalent to 5.08 mm. Next, the CBR value is determined by dividing the corrected loads or stresses obtained from the graph by the standard loads or stresses, which are 13.44 and 20.06 kN (1000 and 1500 psi) for 0.1 and 0.2 in., respectively (Kuttah, 2019; Shirur and Hiremath, 2014; Bresfelean, 2007).

An earlier study, as documented in Onyelowe *et al*. (2021), shows that the CBR value (R7) depends on:

1. rice husk ash undergoes hydration and activation (R1)
2. plastic limit (R3)
3. liquid limit (R2)
4. optimal moisture content (R4)
5. clayey activity (R5)
6. maximum dry density (R6)

Hence, the current inquiry included these input factors when formulating the proposed models.

Data preparation

The whole data set is divided into a training set and a testing set to facilitate the development of the ML algorithm model. The division was performed by allocating 70% of the entire data set to the training data set and assigning 30% to the testing data set. The data is only applicable to expansive soil which may or may not contain binder. The data points are more than enough because it is used by many researchers for their studies (Luo *et al*., 2016; Mair *et al*., 2000; Tigga and Garg, 2020; Raschka, 2018). To achieve improved outcomes, we partitioned the data into training, testing, and total data sets, ensuring that the parameters' minimum, maximum, average, and standard deviation remained consistent across all sets. The current data set prevents the model from overfitting (refer to Table 1). Figure 2 displays the histogram distributions and cumulative percentages for all characteristics derived from the database used in the simulation of CBR value. Input and output parameters were used to draw these histograms. The distributions of both parameters exhibited little to no skewness. As shown in Table 2, the calculated skewness and kurtosis values confirm their near-normal distribution. For computation, we gave the same weight to each of the 121 observations (Amjad *et al*., 2022).

Table 1 provides statistical summaries for a data set across several parameters (R1–R7), categorised into train, test, and total sets. Combining the training and testing sets is a merged data set that provides a comprehensive representation of all the features, denoted by the letters R1 through R7, with exceptional detail. Examining the minimum values for each attribute reveals a broad spectrum, ranging from a minimum threshold of 0.3 to a high of 27.6. This version illustrates how much each attribute may reach its minimum value. The maximum values are the uppermost limits that each feature may reach, ranging from 11.9 to 65.6 at the highest point of the scale. The feature averages provide a more comprehensive understanding of the main trends in the combined data set. The range of the averages, spanning from 6.10 to 47.74, provides insight into the typical values that the data tends to concentrate around in Table 1. The variation in averages across different attributes might provide valuable insights into the overall trend or pattern in the collection.

In addition, by analysing the standard deviations across characteristics, we may ascertain the extent of variability or dispersion within each feature. The variability of data points around the average value for each feature is extensive, as shown by the standard deviations, which vary from 0.23 to 11.36 (Table 1). A more significant standard deviation in the data points indicates more

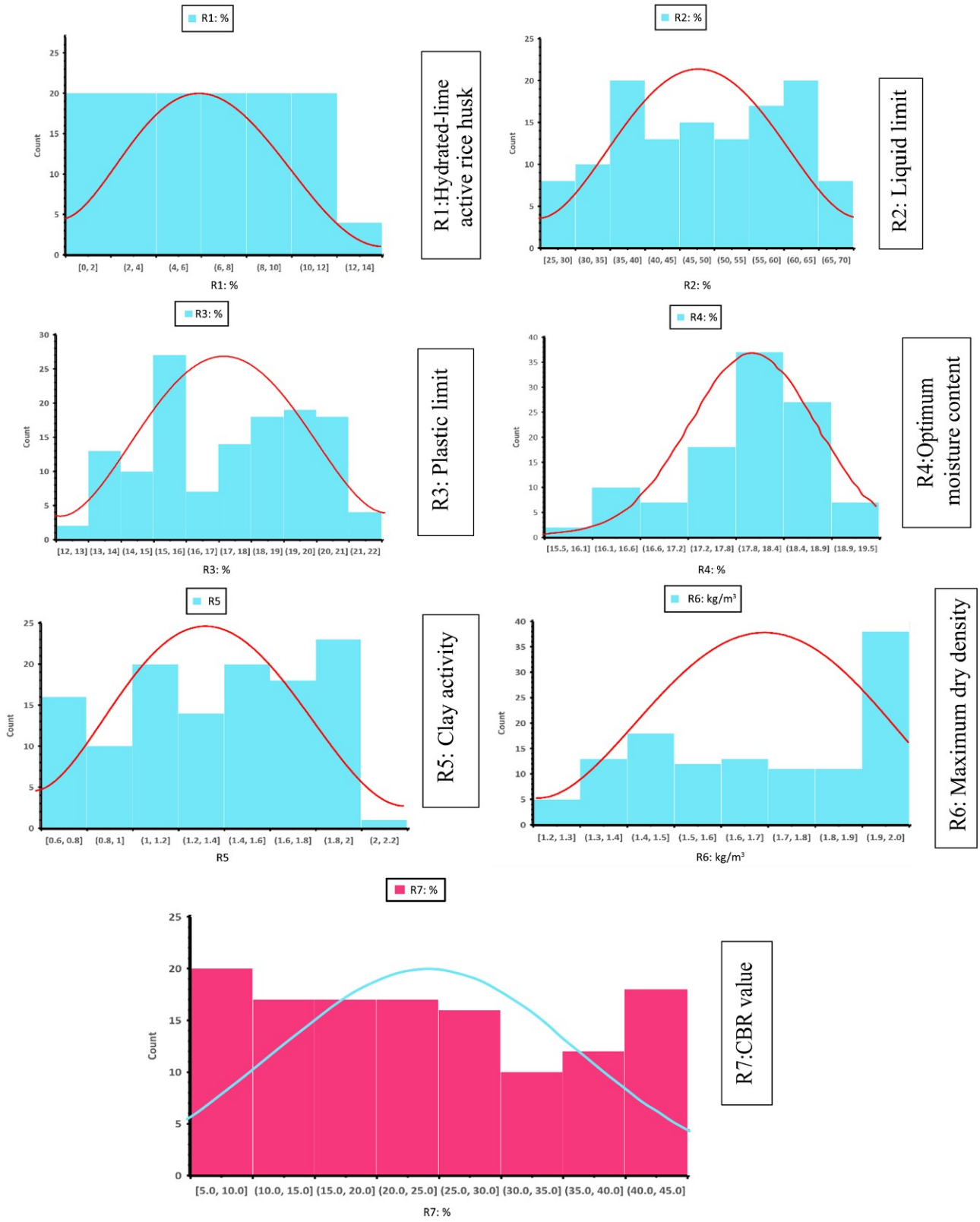

Figure 2. Comparison of input (aqua) and output (pink) distributions shown as histograms

Table 2. Pearson correlation matrix for the input and output parameters in the model

Parameters	R1	R2	R3	R4	R5	R6
R1	1.0000	—	—	—	—	—
R2	−0.9972	1.0000	—	—	—	—
R3	−0.9893	0.9915	1.0000	—	—	—
R4	0.2014	−0.1435	−0.1749	1.0000	—	—
R5	−0.9939	0.9975	0.9846	−0.1204	1.0000	—
R6	0.9858	−0.9818	−0.9770	0.2394	−0.9742	1.0000

volatility, whereas a lower standard deviation implies greater stability. The table presents statistical data for each of the seven parameters (R1–R7) in three distinct sets: the training set, the testing set, and the combined total set. The data includes information on the lowest, highest, average, and standard deviation values. These statistics facilitate comprehension of the distribution and variability of the data across different parameters.

Correlation analysis

Employ Pearson's correlation coefficient to ascertain the existence of linear associations between the input and outcome parameters of the investigation. The magnitude and orientation of these connections were evaluated using the correlation coefficient (ρ). The absolute value of the correlation coefficient determines the degree of correlation. A value of $|\rho| > 0.8$ indicates a strong association, $|\rho| = 0.3$–0.8 suggests an average connection, and $|\rho| = 0.30$ indicates a weak relationship. This work used Pearson's correlation coefficient to determine the linear relationship between input and output parameters (Kumar and Chong, 2018).

1. $$\rho(s,t) = \frac{cov(s,t)}{\sigma_s \sigma_t}$$

The symbols "cov" reflect the concept of covariance, "s" represents standard deviation, and "t" represents standard deviation as well. The Pearson correlation coefficient (ρ) was used to ascertain the associations between each pair of variables (Kumar and Chong, 2018). Table 2 summarises the correlation between all parameters based on the absolute value of ρ. Table 2 clearly demonstrates a substantial correlation between R1 and R6 (highlighted in red) and R7, with correlation coefficients of $|\rho| = 0.9779$ and $|\rho| = 0.9524$, respectively. R2, R3, and R4 (in green) exhibit a mild association with CBR value, with correlation coefficients of −0.9802, −0.9635, and −0.9820, respectively. R4 (in yellow) also shows a moderate link with CBR value, with a correlation coefficient of 0.0927.

Machine learning algorithms

Random forest algorithms

RF is a widely used ML technique created by Breiman in 2001, i.e. often used for regression analysis and classification tasks (Breiman, 2001). This learning technique is a complete approach that enhances the performance of a single decision tree by using majority voting or means outcomes to increase forecast accuracy (Amjad *et al*., 2022). A RF model is an ensemble of decision trees used to predict the ultimate result, denoted as Rn. Tree-growing methods use partitioning to divide training sets into smaller subgroups and then randomly choose predicted pieces. The Gini diversity index, RMSE, and MSE are used to determine the exact moment tree development stops. The optimal RF model selects trees that provide accurate predictions, effectively mitigating overfitting by randomly picking the final set of decision trees and predictor parameters (Amjad *et al*., 2022). The RF structure, seen in Figure 3, integrates predictions from individual trees in classification tasks, whereas regression calculates the average predictions for each activity.

For majority voting classification

2. $$\hat{Z}_{RF} = \frac{1}{NT}\sum_{i=1}^{NT}(Z_i = Z_i)$$

For averaging regression

3. $$\hat{Z}_{RF} = \frac{1}{NT}\sum_{i=1}^{NT}\hat{Z}_i$$

Artificial neural network

ANNs are statistical methods that construct logical models of linked neurons in computer networks, imitating the functions of the brain and spinal cord. They can resolve modelling problems such as estimation, classification, and pattern recognition. ANNs may be categorised into supervised and unsupervised. In supervised learning, the network weights are adjusted to make accurate predictions based on given goal values. On the other hand, unsupervised learning involves providing inputs to the network without accompanying target values which is shown in Figure 4 (Tenpe and Patel, 2020). The MLP is a popular feed-forward neural network that requires careful selection of appropriate values for its hidden layers during training. The output layer acts as the input layer for the following stages in the forward computation process.

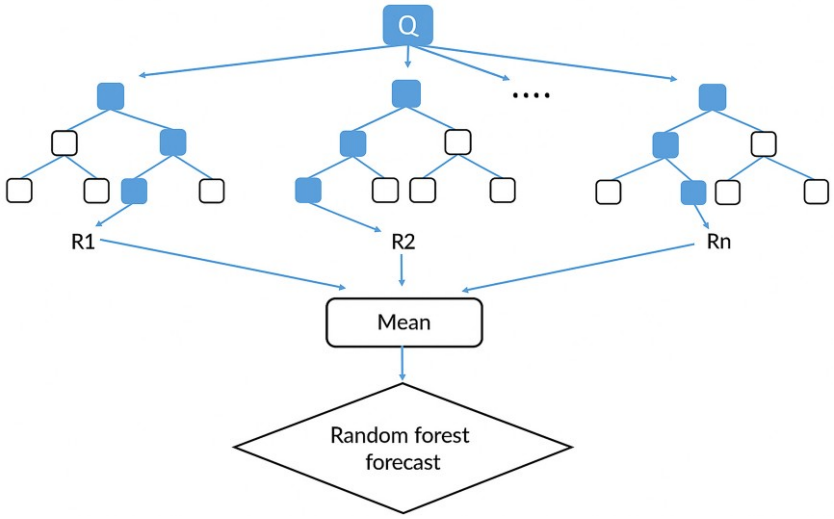

Figure 3. The structure of the random forest

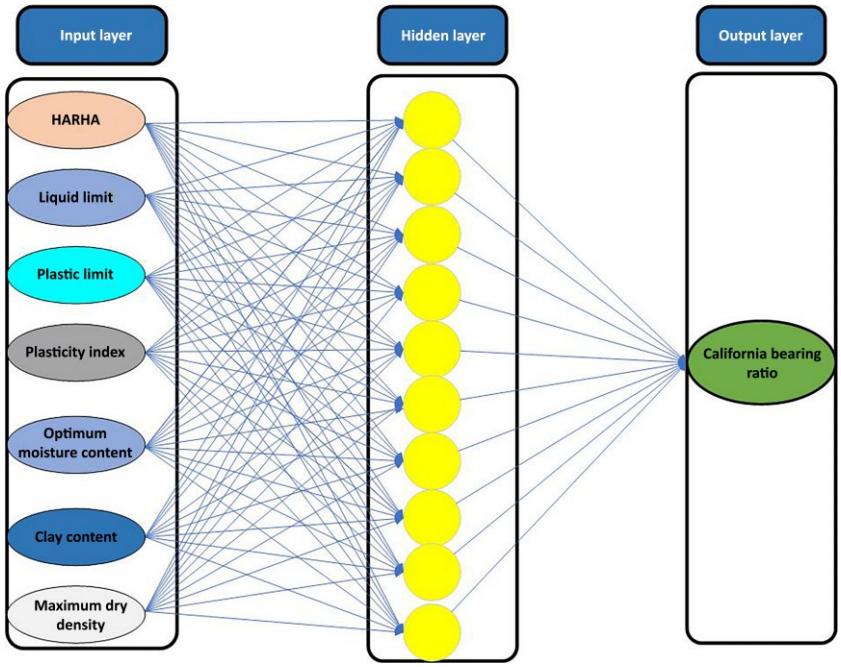

Figure 4. Artificial neural network mechanisms

$$O_j = f\left(\sum_{i=1}^{N} w_{ij} \times I_i + b_j\right) \quad (4)$$

where O_j is the output for neuron j which is determined by many factors: the total number of neurons in the previous layer N, W_{ij} is the weight between the corresponding neuron of the previous layer I_i and neuron j, the output neuron of the previous layer, and the bias term for neuron j. For forward computation,

$$\theta = \frac{\partial l}{\partial p} \quad (5)$$

The back propagation learning method creates a relationship between inputs and outputs by assigning random weights to

input data and then modifying them. In neural computing, employing multiple transfer functions, such as tangent sigmoid functions for hidden layers and linear functions for output layers, is used to improve the input-output behaviour. For classification tasks, the output layer often employs a softmax activation function (Jogin *et al*., 2018).

6. $$p_i = \frac{e^{RO_i}}{\sum_{i=1}^{N} e^{RO_i}}$$

where p_i is the probability score for each class while the network raw output for the i and j neurons are denoted by RO_i and RO_j. ANNs rely on mathematical concepts such as neurons, activation functions, and optimisation algorithms to model intricate data linkages (Basheer and Hajmeer, 2000).

Support vector machine

The fundamental concept behind SVM is to use a function that transforms the training data from the input space to a feature space of greater dimensions. This enables the construction of a separation hyperplane with the largest possible margin in the feature space, as seen in Figure 5. SVM uses a training data set and an offset scalar to determine the direction of a hyperplane b. These correspond to the amount of training data and class labels for positive instances and negative examples, respectively (Kecman, 2005). SVM uses a training data set to determine a hyperplane direction and an offset scalar. These values are related to the amount of training data and the class labels for positive and negative instances, respectively (Kecman, 2005).

Let us consider a set of training data, where each data point represents the target value for a certain input sample in the input space (Peshkin *et al*., 2004). The objective of the regression problem is to ascertain a mathematical function that can accurately forecast forthcoming values.

The default format of the SVR estimate function is as follows:

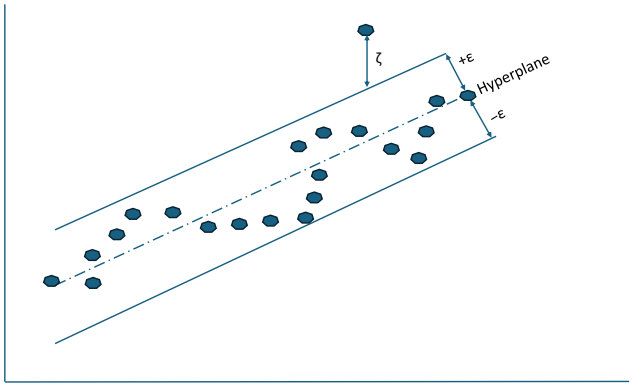

Figure 5. Mechanics of support vector machine

7. $$f(y) = \omega \times \varphi(y) + c$$

This scenario demonstrates a transition from a low-dimensional environment to a high-dimensional one, i.e. not linear. Our objective is to ascertain the values of w and c to minimise the danger of regression.

8. $$R_{reg}(f) = D\sum_{i=1}^{N} e^{RO_i} \times \varphi(y) + c$$
$$R_{reg}(f) = D\sum_{i=1}^{N} f(y_i - z_i) + \frac{1}{2}\|\omega\|^2$$

Given that D is a constant in this scenario, the vector w may be expressed in terms of data points as follows:

9. $$\omega = \sum_{i=1}^{N}(a_i - a_i^*)\psi(y_i)$$

10. $$f(y) = \sum_{i=1}^{N}(a_i - a_i^*)\psi(\mathrm{y}) + \mathrm{c}$$

11. $$= \sum_{i=1}^{N}(a_i - a_i^*)k(y_{i,}y) + c$$

The dot product may be substituted in Equation (11) with *ith* the kernel function, also known as the function (Islam and Roy, 2020). Kernel functions enable the computation of the dot product in a feature space with a large number of dimensions, using data provided from a space with a lower number of dimensions without requiring knowledge of the transformation. The Mercer condition is a requirement that all kernel functions must satisfy, and it relates to the inner product of a certain feature space. The radial basis function is often used as the regression kernel.

12. $$k(y_{i,}y) = exp\left[-ve|y - y_{i,}|^2\right]$$

Gaussian process regression

Gaussian process regression (GPR) is a probabilistic, non-parametric supervised learning approach that may generalise complex, non-linear, and hidden function mappings in data sets. The GPR model is a method that follows Rasmussen and Williams' recommendation that neighbouring observations should provide information (Kumar *et al*., 2013) by explicitly including a prior throughout the whole function space. The mean and covariance of a Gaussian distribution are represented by vectors and matrices,

respectively. In contrast, a Gaussian process may be described as an over function. The GPR model may be used to find a prediction distribution like the testing input. A GPR is a set of random variables that follows a joint multivariate Gaussian distribution for any finite number of variables.

To get more information on GPR and other covariance functions, go to Kuss (Kumar *et al*., 2013). The specifics of kernel functions are as follows: The GPR design process employs the kernel function. A comprehensive range of kernels has been investigated in the literature (Alam *et al*., 2020; Sabat, 2015). This study uses the Pearson universal kernel (PUK) functions listed below:

PUK:

$$k(p,q) = \left\{ \frac{1}{\left\{ \left[\sqrt{\|p-q\| \sqrt{2^{1/\omega} - \frac{1}{\sigma}}} \right]^2 \right\}^{\omega}} \right\} \quad (13)$$

Construction of prediction models

The primary objective of the research was to investigate the CBR value as the dependent variable (y), while the independent variables R1, R2, R3, R4, R5, R6, and R7 constituted the input set (x). Choosing the optimal amount of training and testing data sets is crucial at every stage of the modelling process. The models were developed using 70% of the research data and then evaluated using the remaining data. To clarify, we used 85 sets for training and 36 sets for testing the models (Luo *et al*., 2016; Mair *et al*., 2000; Tigga and Garg, 2020; Raschka, 2018). All models were optimised in an iterative process to improve the accuracy of CBR value predictions. Figure 6 depicts the creation process of the prediction models.

Hyperparameter optimisation

Optimisation of hyperparameters is essential in ML projects to get optimum results. The parameter values substantially impact the efficiency of architectures like RF, decision tree, SVM, XGBoost, and AdaBoost. The objective of the optimisation technique is to determine the optimal parameters for these models (Smithson *et al*., 2016; Yu and Zhu, 2020). Hyperparameter tuning is seen to be the trickiest part of creating ML models. The bulk of these ML algorithms offer the hyperparameters' default settings. However, many different ML programs may not necessarily perform effectively with default parameters. To discover the best combination that will give you the best outcomes, you must optimise or adjust them. The model parameters were carefully chosen and adjusted during the experimental trials to achieve the improved metrics, as shown in Table 3.

Table 3. Values of algorithm hyperparameters

ML algorithm	Hyperparameter	Symbol	Tuned value
RF	Bag sized percent	P	100
	Batch size	I	100
	Number of excess slots	Num-slot	1
	—	M	1
	—	V	0.001
	Speed	S	1
ANN (multilayer perceptron)	Learning rate	L	0.3
	Momentum	M	0.2
	Training time	N	500
	Hidden layers	H	a
SVM (Poly kernel)	Exponent	E	1
	—	C	1
GPR (PUK kernel)	Omega	0	0.3
	Sigma	S	0.3
	Noise	N	0.3

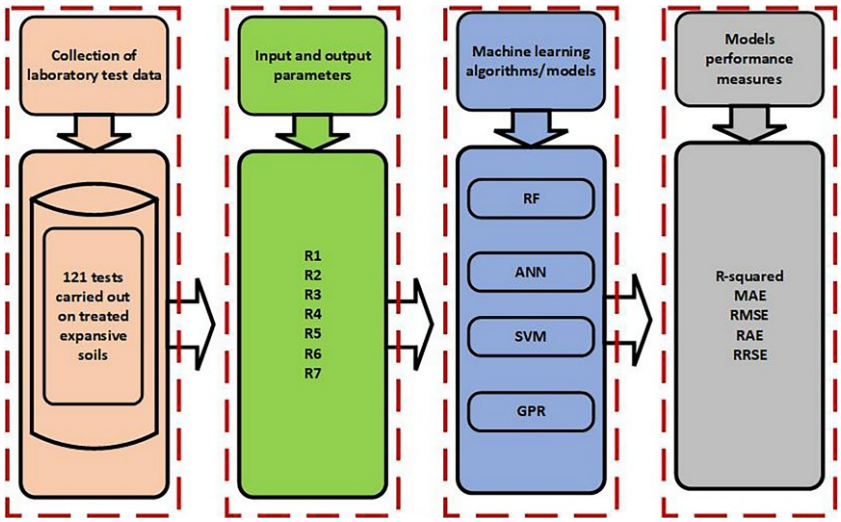

Figure 6. The flowchart used to anticipate CBR values using a data-driven method

Algorithms for ML have parameters that need to be adjusted for best results. To guarantee accurate predictions, the optimisation method focuses on determining the most efficient SVM and ANN parameters. This work defines the meanings of these hyper parameters and adjusts a number of important parameters in the SVM and ANN models. After the models' tuning parameters were established, they were adjusted during trials until they reached the ideal values. Overfitting was reduced by using cross-validated tuning of the hyperparameters. The ANN momentum, along with the SVM kernel rules, was tuned. A held-out test validation was also used. The model generalisation was carefully ensured through these steps.

Evaluation parameters of the proposed model

The performance measures used to evaluate the effectiveness of the recommended models are R-squared, mean absolute error (MA_{Error}), root mean squared error (RMS_{Error}), relative absolute error (RA_{Error}), and relative root squared error (RRSError). The equations are labelled in Equation (14) to Equation (18) provide the mathematical representations for several statistical measures (Amjad *et al*., 2022).

14. $$R-squared = 1 - \left\{\frac{SSregression}{SStotal}\right\} = 1 - \frac{\sum_{i=1}^{N}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{N}(p_i - \bar{p}_i)}$$

15. $$MA_{Error} = \frac{\sum_{i=1}^{N}|p_i - \hat{p}|}{N} = \frac{\sum_{i=1}^{N}|E_i|}{N}$$

16. $$RA_{Error} = \frac{\sum_{i=1}^{N}|p_i - \hat{p}i|}{\sum_{i=1}^{N}|\hat{p}_i - \bar{p}|}$$

17. $$RMS_{Error} = \sqrt{\frac{\sum_{i=1}^{N}|p_i - \hat{p}|}{N}}$$

18. $$RRS_{Error} = \sqrt{\frac{\sum_{i=1}^{N}(p_i - \hat{p}i)^2}{\sum_{i=1}^{N}(p_i - \bar{p}i)^2}}$$

R-squared is a statistical metric that quantifies the proportion of variability in a dependent variable that one or more independent variables can explain in a regression model. The variable may assume two distinct values: a perfect positive correlation (+1) and a perfect negative correlation (−1). A model with an R-squared of 0.50 may explain about 50% of the variability in the data. The MA_{Error} is a metric used to measure the statistical differences between matched data that relate to the same event. The RA_{Error} is a metric used to assess the effectiveness of a prediction model. RMS_{Error} is a frequently used statistic for evaluating the degree to which a model accurately represents the data. A zero number signifies an ideal match of the data. The RRS_{Error} is easily comprehensible since it is based on the simple concept of averaging the realised values in the fundamental model. A model is considered superior if its RRS_{Error} is lower (Taskiran, 2010; Kişi and Uncuoğlu, 2005).

Result and discussions

Models comparison

This section evaluates the efficacy of the model. Table 4 summarises the relevant data, whereas Figures 7 and 8 display the predictive accuracy of the guiding and testing data sets in regression format, respectively. This study compared the proposed and previously documented models in the literature (Luo *et al*., 2016; Amjad *et al*., 2022). Table 4 indicates that the comparison used many performance indicators based on the criteria for achieving the great testing performance. The RF model had the lowest error rate, so the feasibility and practicality of the CBR value prediction were demonstrated using the soft computing approach. The proposed models, including ANN, GPR, SVM, and RF, illustrate the results acquired for training and testing.

Comparing the RF model to ANN, SVM, and GPR, the RF model had the highest accuracy in its predictions. The RF model had an R-squared value of 0.9999, a MA_{Error} of 0.1271, and an RMS_{Error} of 0.1617. In comparison, the ANN model had an R-squared value of 0.9989, an MA_{Error} of 0.4052, and an RMS_{error} of 0.5588. The SVM model had an R-squared value of 0.9972, a MA_{Error} of 0.6924, and an RMS_{Error} of 0.8982. Lastly, the GPR model had an R-squared value of 0.9941, an MA_{Error} of 1.031, and an RMS_{Error} of 1.3390. This is also corroborated by the RA_{Error} and RRS_{Error}

Table 4. Analysing machine learning models' performance for CBR value prediction on training and testing sets

Model	Symbol	Data set	R-squared	MA_{Error}	RMS_{Error}	RA_{Error}	RRS_{Error}
Random forest	RF	Training set	0.9999	0.1271	0.1617	1.2316%	1.3592%
	—	Testing set	0.9997	0.1992	0.2782	2.0942%	2.4832%
Artificial neural network	ANN	Training set	0.9989	0.4052	0.5588	3.9260%	4.6975%
	—	Testing set	0.9986	0.5147	0.6589	5.4104%	5.8807%
Support vector machine	SVM	Training set	0.9972	0.6924	0.8982	6.7079%	7.5512%
	—	Testing set	0.9968	0.7060	0.9380	7.4214%	8.3714%
Gaussian process regression	GPR	Training set	0.9941	1.031	1.3390	9.9880%	11.2571%
	—	Testing set	0.9931	1.0894	1.5060	11.4518%	13.4408%

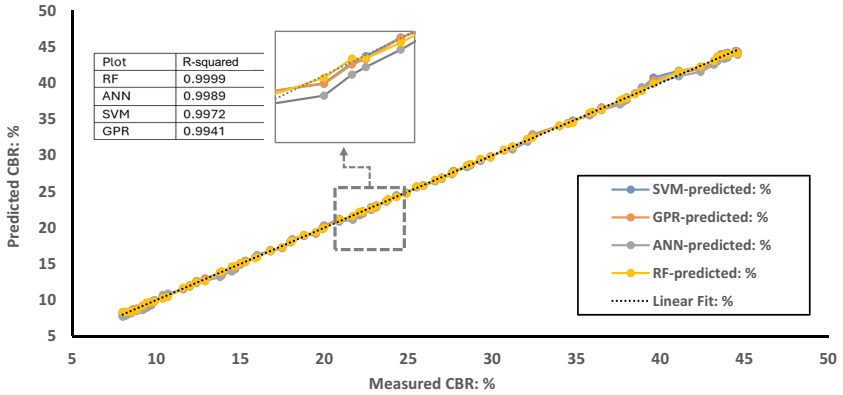

Figure 7. Comparison between measured and predicted CBR values for model training

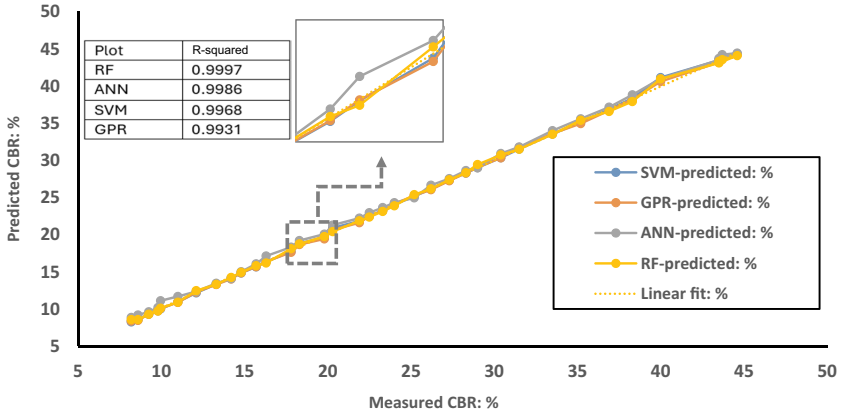

Figure 8. Comparison between measured and predicted CBR values for model validation

outcomes in the table provided below.

During the testing phase, the RF model demonstrated superior performance to the ANN, SVM, and GPR models regarding various metrics. The RF model achieved an R-squared value of 0.9986, MA_{Error} of 0.05147, RMS_{Error} of 0.6589, RA_{Error} of 5.4104%, and RRS_{Error} of 5.8807%. In comparison, the SVM model achieved an R-squared value of 0.9968, MA_{Error} of 0.7060, RMS_{Error} of 0.9380, RA_{Error} of 7.4214%, and RRS_{Error} of 8.3714%. Lastly, the GPR model achieved an R-squared value of 0.9931, MA_{Error} of 1.0894, RMS_{Error} of 1.5060, RA_{Error} of 11.4518%, and RRS_{Error} of 13.4408%.

Models in the area of ML need assessment to determine their operational effectiveness. Different model types use different assessment methodologies. After developing a machine-learning model for CBR value forecasting, the next important step is to evaluate the model's predictive capabilities (Amjad *et al*., 2022; Brenning, 2005). The study confirmed the accuracy of the CBR value predictions generated by the proposed models by comparing the anticipated and actual values of CBR value. Figures 9 and 10 provide a comparison between the projected and observed values of CBR value for both the training and testing sets. The training set's anticipated and actual values demonstrate a significant level of coherence, as seen in Figure 9. The projected value is often accurate, even if a few data points in the predicted value of the testing set exhibit significant discrepancies when compared to the actual value. In general, the prediction is satisfactory and appropriate (Brenning, 2005).

Figure 9 presents a scatter plot comparing the actual and projected values for the training and testing sets. This visualisation helps us understand the degree of alignment between the predicted values and the real values. The CBR values for the training and testing sets range from 8.2% to 44.5%, as seen in Figure 10. There are few disparities in the testing set and a strong correspondence between the expected and observed values in both sets. Notably, there are a few instances in the testing set where the predicted value

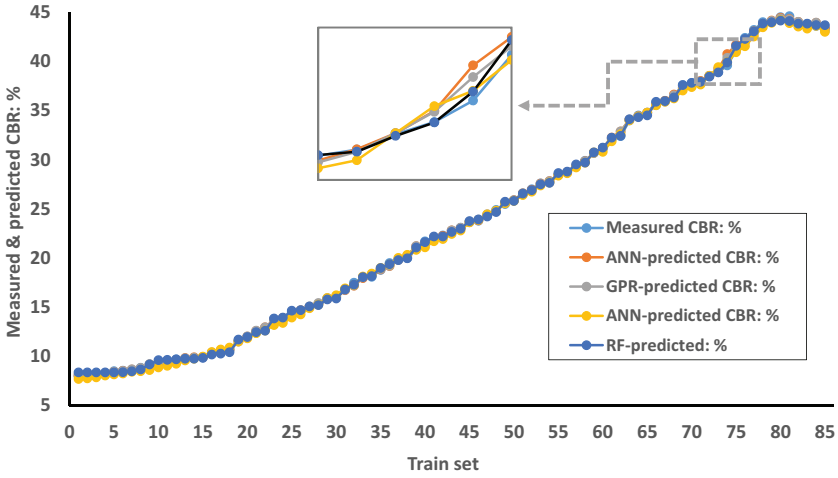

Figure 9. Visualisation of model accuracy for CBR value prediction on training set

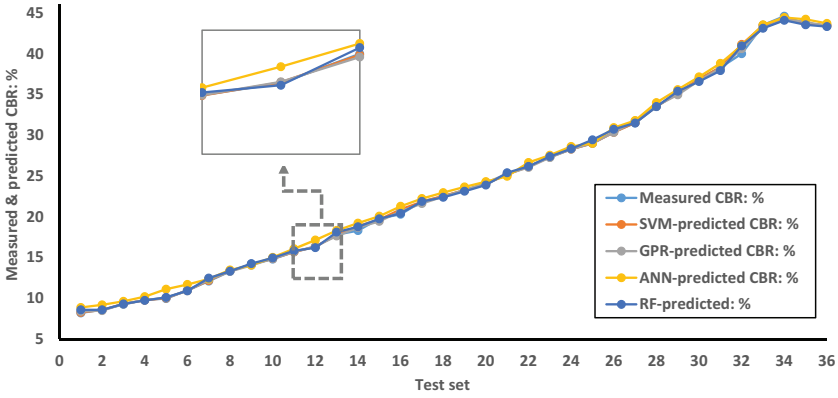

Figure 10. Visualisation of model accuracy for CBR value prediction on training set and testing set

reached 40.614%, whereas the actual CBR value was almost 40%, suggesting more significant discrepancies (refer to Figure 10). Nevertheless, the little deviations in individual data points do not impact the overall predictive capability of the offered models. The error analysis graphs of the recommended models indicate that the RF model has the lowest error rate compared to the ANN, SVM, and GPR models. Thus, the study reaffirms that the RF model accurately predicts the CBR value without any occurrences of overfitting.

Residual values comparison

The training and testing data sets indicate that none of the models can attain flawless accuracy, as seen by relative errors ranging from −0.9 to 1.5 (see Figures 11(a) and 11(b)). The ANN, SVM, and GPR models provide the highest level of precision in predictions of CBR value, with the RF-predicted model following closely after. The RF model has low error rates in both data sets and an exceptional ability to forecast CBR values within the given input data intervals (Botchkarev, 2018; Bannach-Brown *et al*., 2019; Amershi *et al*., 2015; Chen *et al*., 2004).

Sensitivity analysis

Sensitivity analysis is the technique used to ascertain the impact of individual input parameters on an output parameter. Essentially, it computes the output yields that arise from modifying the input. The approach developed by Yang and Zhang (1997) was used to evaluate the impact of input parameters on the sensitivity outcomes of the XGBoost model. This approach is outlined as follows and has been used in several investigations (Amjad *et al*., 2022; Ahmad *et al*., 2019; Ahmad *et al*., 2021).

19. $$q_{ij} = \frac{\sum_{k=1}^{n}(R_{im} x R_{om})}{\sqrt{\sum_{k=1}^{n} R_{ik}{}^{2} \sum_{k=1}^{n} R_{ok}{}^{2}}}$$

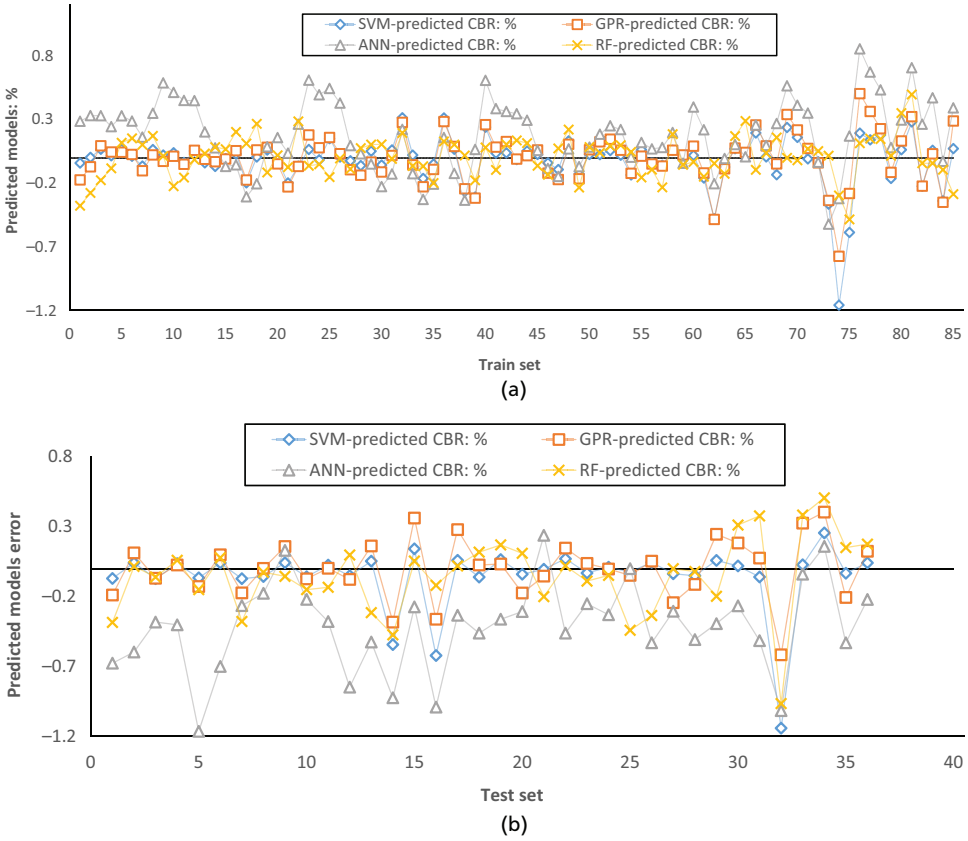

Figure 11. (a) Training set error levels for the suggested models and (b) testing set error levels for the suggested models

In this context, R_{im} and R_{om} represent the input and output variables, respectively, while N represents the total number of values, which is 85 in this particular situation. The maximum q_{ij} values, ranging from zero to one, reflect the CBR value for each input parameter (Amjad *et al*., 2022). Figure 12 displays the q_{ij} scores corresponding to each input variable. Figure 12 illustrates that R1, R6, and R4 have the highest degree of effect (q_{ij} = 19%) in predicting CBR value. However, R5 has the smallest impact on the prediction of CBR value in Figure 12.

Compare with other studies in the subject

Table 5 displays the results of studies that examined the use of ML in assessing the load-bearing capability of piles. Prior studies have shown that ML algorithms have a high level of accuracy in predicting CBR value outcomes, as indicated by R-squared values ranging from 0.81 to 0.96. However, in this research, the range is from 0.9941 to 0.9999. The comparison of these outcomes is not feasible due to the use of disparate data sets. It is essential to conduct a comprehensive analysis using several data sources to provide a broad framework for foundation engineering (Amjad *et al*., 2022; Kurt and Kayfeci, 2009).

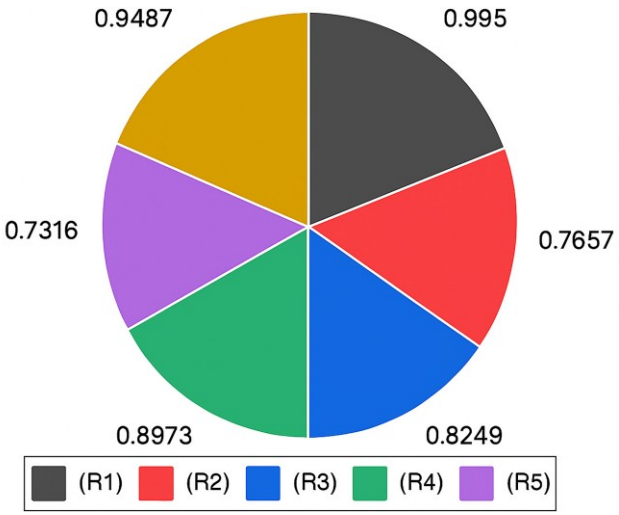

Figure 12. Sensitivity analysis showing the impact of each input variable on model output

Conclusions and ideas for further research

This research used ML techniques to predict the CBR value of treated soil. The performance of the generated models was assessed by using statistical measures. The measures including the coefficient of determination (R^2), mean absolute error (MA_{Error}), root mean square error (RMS_{Error}), relative absolute error (RA_{Error}), and relative root square error (RRS_{Error}). Below is a concise overview of the research's findings:

Table 5. Results matrix of the previous studies

Previous studies	Developed model(s)	Evaluation parameter(s)	Number of data points used
Ahmad *et al*. (2021)	MARSE	R-squared is 96%; RMSE is 3.59%	362
Ahmad *et al*. (2021)	ELM	R-squared between 81% and 91%	312
Rafizul and Roy (2020)	LMBP, BP, CG	R-squared is 90%	129
Yang and Zhang (1997)	Elastic net regularisation regression (ENRR), Lazy K-star, M-5 model tree, Gaussian	R-squared between 90% and 92%	97

- The suggested models yielded R-squared along with MA_{Error}, RA_{Error}, RMS_{Error}, and RRS_{Error} values to forecast CBR values. The RF model (R-squared = 0.9999) produced higher prediction results with MA_{Error} = 0.1271, RMS_{Error} = 0.1617, RA_{Error} = 1.2316%, and RRS_{Error} = 1.3592%) followed by the ANN model (R-squared = 0.9989, MA_{Error} = 0.4052, RMS_{Error} = 0.5588, RA_{Error} = 3.9260%, and RRS_{Error} = 4.6975%), the SVM model (R-squared = 0.9972, MA_{Error} = 0.6924, RMS_{Error} = 0.8982, RA_{Error} = 6.7079%, and RRS_{Error} = 7.5512%), and the GPR model (R-squared = 0.9941, MA_{Error} = 1.031, RMS_{Error} = 1.3390, RA_{Error} = 9.9380%, and RRS_{Error} = 11.2571%).
- The sensitivity analysis results indicate that HARHA (R1) is the parameter with the most significance when predicting the CBR value followed by maximum dry density (R6).
- This work shows the efficiency of AI models in evaluating the CBR value of soil treated with activated RHA, water, and lime, thereby emphasising its use in assessing soil mechanical characteristics. This study also investigates the use of sophisticated ML techniques, including deep learning, to predict the CBR value in soil treated with HARHA.

Recommendations for future work

- Increase experimental data: Future study should aim to collect more experimental data in order to improve the universality and resilience of the suggested technique. The accuracy and dependability of AI models may be further enhanced with a larger and more varied data set.
- Explore advanced ML algorithms: This study mainly concentrated on traditional AI methods. The investigation of sophisticated ML methods, such as deep learning, for CBR prediction, is a potential area for future study. These algorithms could provide forecasts and insights that are even more accurate.

Funding statement

The author(s) received no specific funding for this study.

Ethics statement

Not applicable for studies not involving humans or animals.

Conflicts of interest

The authors declare no conflicts of interest to report regarding the present study

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Irfan Jamil[1], Umair Ahmad[2]; data collection: Umair Ahmad[1], Hamza Jamal[2], Muhammad Bilal Khan[1]; analysis and interpretation of results: Umair Ahmad[1], Irfan Jamil[1]; draft manuscript preparation: Hamza Jamal[2], Oussama Accouche[3], Marc Azab[3]. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author [Umair Ahmad] upon reasonable request

REFERENCES

Abdulnabi TY and Abdulrazzaq ZG (2020) An estimated correlation between California bearing ratio (CBR) with some soil parameters of gypseous silty sandy soils. *Tikrit Journal of Engineering Sciences* **27(1)**: 58–64.

Ahmad M, Ahmad F, Wróblewski P *et al.* (2021) Prediction of ultimate bearing capacity of shallow foundations on cohesionless soils: a Gaussian process regression approach. *Applied Sciences* **11(21)**: 10317.

Ahmad M, Kamiński P, Olczak P *et al.* (2021) Development of prediction models for shear strength of rockfill material using machine learning techniques. *Applied Sciences* **11(13)**: 6167.

Ahmad M, Tang XW, Qiu JN and Ahmad F (2019) Evaluating seismic soil liquefaction potential using Bayesian belief network and C4.5 decision tree approaches. *Applied Sciences* **9(20)**: 4226.

Alam SK, Mondal A and Shiuly A (2020) Prediction of CBR value of fine grained soils of Bengal Basin by genetic expression programming, artificial neural network and Krigging method. *Journal of the Geological Society of India* **95(2)**: 190–196.

Al-Busultan S, Aswed GK, Almuhanna RR and Rasheed SE (2020) January Application of artificial neural networks in predicting subbase CBR values using soil indices data. In *IOP Conference Series: Materials Science and Engineering* **671(1)**: 012106.

Ali T, Haider W, Ali N and Aslam M (2022) A machine learning architecture replacing heavy instrumented laboratory tests: in application to the pullout capacity of geosynthetic reinforced soils. *Sensors (Basel, Switzerland)* **22(22)**: 8699.

Amershi S, Chickering M, Drucker SM, *et al.* (2015). Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 337–346).

Amjad M, Ahmad I, Ahmad M *et al.* (2022) Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation. *Applied Sciences* **12(4)**: 2126.

Breiman L (2001) Random forests. *Machine Learning* **45(1)**: 5–32.

Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences* **5(6)**: 853–862.

Bresfelean VP (2007) Analysis and predictions on students' behavior using decision trees in Weka environment. In *2007 29th International Conference on Information Technology Interfaces* (pp. 51–56). IEEE.

Botchkarev A (2018) Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. *arXiv Preprint arXiv:1809.03006*

Basheer IA and Hajmeer M (2000) Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods* **43(1)**: 3–31.

Bannach-Brown A, Przybyła P, Thomas J *et al.* (2019) Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews* **8(1)**:12–23.

Bardhan A, Gokceoglu C, Burman A *et al.* (2021) Efficient computational techniques for predicting the California bearing ratio of soil in soaked conditions. *Engineering Geology* Vol. 291: 106239.

Bardhan A, Samui P, Ghosh K *et al.* (2021) ELM-based adaptive neuro swarm intelligence techniques for predicting the California bearing ratio of soils in soaked conditions. *Applied Soft Computing* Vol. **110**: 107595.

Bhatt S, Jain PK and Pradesh M (2014) Prediction of California bearing ratio of soils using artificial neural network. *Am. Int. J. Res. Sci. Technol. Eng. Math* **8(2)**: 156–161.

Chen DR, Wu Q, Ying Y and Zhou DX (2004) Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research* Vol. 5, pp: 1143–1175.

Erzin Y and Turkoz D (2016) Use of neural networks for the prediction of the CBR value of some Aegean sands. *Neural Computing and Applications* **27(5)**: 1415–1426.

Ferentinou MD and Sakellariou MG (2007) Computational intelligence tools for the prediction of slope performance. *Computers and Geotechnics* **34(5)**: 362–384.

González Farias I, Araujo W and Ruiz G (2018) Prediction of California bearing ratio from index properties of soils using parametric and non-parametric models. *Geotechnical and Geological Engineering* **36(6)**: 3485–3498.

Haupt FJ and Netterberg F (2021) Prediction of California bearing ratio and compaction characteristics of transvaal soils from indicator properties. *Journal of the South African Institution of Civil Engineering* **63(2)**: 47–56.

Ikechukwu AF and Mostafa MM (2020) Performance assessment of pavement structure using dynamics cone penetrometer (DCP). *International Journal of Pavement Research and Technology* **13(5)**: 466–476.

Islam MR and Roy AC (2020) Prediction of California bearing ratio of fine-grained soil stabilized with admixtures using soft computing systems. In *Journal of Civil Engineering*. Science and Technology, **11(1)**, pp. 28–44.

Jogin M, Madhulika MS, Divya GD, *et al.* (2018). Feature extraction using convolution neural networks (CNN) and deep learning, In *2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT)* (pp. 2319–2323).

Kecman V (2005) Basics of machine learning by support vector machines. In *Real World Applications of Computational Intelligence*. Springer, Berlin, Heidelberg, pp. 49–103.

Kuttah D (2019) Strong correlation between the laboratory dynamic CBR and the compaction characteristics of sandy soil. *International Journal of Geo-Engineering* **10(1)**: 7.

Kişi Ö and Uncuoğlu E (2005) Comparison of three back-propagation training algorithms for two case studies, *Indian Journal of Engineering and Materials Sciences* **12(5)**.

Kumar S and Chong I (2018) Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states. *International Journal of Environmental Research and Public Health* **15(12)**: 2907.

Kurt H and Kayfeci M (2009) Prediction of thermal conductivity of ethylene glycol–water solutions by using artificial neural networks. *Applied Energy* **86(10)**: 2244–2248.

Kumar SA, Kumar JP and Rajeev J (2013) Application of machine learning techniques to predict soaked CBR of remolded soils. *International Journal of Engineering Research & Technology (IJERT)* **2(6)**: 3019–3024.

Luo Y, Szolovits P, Dighe AS and Baron JM (2016) Using machine learning to predict laboratory test results. *American Journal of Clinical Pathology* **145(6)**: 778–788.

Mair C, Kadoda G, Lefley M *et al.* (2000) An investigation of machine learning based prediction systems. *Journal of Systems and Software* **53(1)**: 23–29.

Merghadi A, Abderrahmane B and Tien Bui D (2018) Landslide susceptibility assessment at Mila Basin (Algeria): a comparative assessment of prediction capability of advanced machine learning methods. *ISPRS International Journal of Geo-Information* **7(7)**: 268.

Mishra D, Tutumluer E and Butt AA (2010) Quantifying effects of particle shape and type and amount of fines on unbound aggregate performance through controlled gradation. *Transportation Research Record: Journal of the Transportation Research Board* **2167(1)**: 61–71.

Mousavi F, Abdi E and Rahimi H (2014) Effect of polymer stabilizer on swelling potential and CBR of forest road material. *KSCE Journal of Civil Engineering* **18(7)**: 2064–2071.

Onyelowe K, Salahudeen AB, Eberemu A, *et al.* (2019). *Oxides of carbon entrapment for environmental friendly geomaterials ash derivation*. (pp. 58–67): Springer International Publishing, Cham.

Onyelowe KC, Onyia ME, Onukwugha ER *et al.* (2020) Polynomial relationship of compaction properties of silicate-based RHA modified expansive soil for pavement subgrade purposes. *Építőanyag: Journal of Silicate Based & Composite Materials* **72(6)**.

Onyelowe KC, Jalal FE, Onyia ME *et al.* (2021) Application of gene expression programming to evaluate strength characteristics of hydrated lime activated rice husk ash-treated expansive soil. *Applied Computational Intelligence and Soft Computing* **2021(1)**: 1–17.

Paige-Green P and Du Plessis L (2009) Use and interpretation of the dynamic cone penetrometer (DCP) test. *CSIR Built Environment Pretoria* Vol. 2.

Peshkin DG, Hoerner TE and Zimmerman KA (2004) In *Optimal Timing of Pavement Preventive Maintenance Treatment Applications*. Transportation Research Board, Vol. **523**.

Rasmin R (2009) *Comparison between sieve analysis & hydrometer with laser particle analyzer to determine particle size distribution* (Doctoral dissertation).

Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning. *Computer Science Machine Learning*, 10.48550/arXiv.1811.12808.

Rafizul IM and Roy AC (2020) Prediction of California bearing ratio of fine-grained soil stabilized with admixtures using soft computing systems. *Journal of Civil Engineering* **11(1)**.

Raza SAM, Maqsood A, Umer M *et al.* (2021) Application of data-driven machine learning approach for the estimation of California bearing ratio of the subgrade soils. *Arabian Journal for Science and Engineering* **46(10)**: 10113–10131, 10.1007/s13369-021-05809-y.

Roy B and Singh MP (2020) An empirical-based rainfall-runoff modelling using optimization technique. *International Journal of River Basin Management* **18(1)**: 49–67.

Rehman ZU, Khalid U, Farooq K, Mujtaba and H, T (2017) Prediction of CBR value from index properties of different soils. *Technical Journal of University of Engineering & Technology Taxila* **22(2)**.

Sabat AK (2015) Prediction of California bearing ratio of a stabilized expansive soil using artificial neural network and support vector machine. *Electron. J. Geotech. Eng* **20(3)**: 981–991.

Shirur NB and Hiremath SG (2014) Establishing relationship between CBR value and physical properties of soil. *IOSR Journal of Mechanical and Civil Engineering* **11(5)**: 26–30.

Smithson SC, Yang G, Gross WJ and Meyer BH (2016) Neural networks designing neural networks: multi-objective hyper-parameter optimization. In *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (pp. 1–8). IEEE.

Taha S, Gabr A and El-Badawy S (2019) Regression and neural network models for California bearing ratio prediction of typical granular materials in Egypt. *Arabian Journal for Science and Engineering* **44(10)**: 8691–8705, 10.1007/s13369-019-04046-8.

Taskiran T (2010) Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. *Advances in Engineering Software* **41(6)**: 886–892.

Talukdar DK (2014) A study of correlation between California bearing ratio (CBR) value with other properties of soil. *International Journal of Emerging Technology and Advanced Engineering* **4(1)**: 559–562.

Tenpe AR and Patel A (2020) Application of genetic expression programming and artificial neural network for prediction of CBR. *Road Materials and Pavement Design* **21(5)**: 1183–1200.

Thives LP and Trichês G (2022) Methodology of layers' compaction control through dynamic cone penetrometer (DCP). *Arabian Journal of Geosciences* **15(13)**: 1224.

Tigga NP and Garg S (2020) Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science* Vol. 167: 706–716.

Trong DK, Pham BT, Jalal FE *et al.* (2021) On random subspace optimization-based hybrid computing models predicting the California bearing ratio of soils. *Materials (Basel, Switzerland)* **14(21)**: 6516.

Yang Y and Zhang Q (1997) A hierarchical analysis for rock engineering using artificial neural networks. *Rock Mechanics and Rock Engineering* **30(4)**: 207–222.

Yildirim B and Gunaydin O (2011) Estimation of California bearing ratio by using soft computing systems. *Expert Systems with Applications* **38(5)**: 6381–6391.

Yu T and Zhu H (2020) Hyper-parameter optimization: a review of algorithms and applications. *Computer Science Machine Learning*, 10.48550/arXiv.2003.05689.