

Article

Natural Gradient Boosting for Probabilistic Prediction of Soaked CBR Values Using an Explainable Artificial Intelligence Approach

Esteban Díaz ^{1,*}  and Giovanni Spagnoli ² ¹ Department of Civil Engineering, University of Alicante, 03690 Alicante, Spain² DMT GmbH & Co., KG, Am TÜV 1, 45307 Essen, Germany

* Correspondence: esteban.diaz@ua.es

Abstract: The California bearing ratio (CBR) value of subgrade is the most used parameter for dimensioning flexible and rigid pavements. The test for determining the CBR value is typically conducted under soaked conditions and is costly, labour-intensive, and time-consuming. Machine learning (ML) techniques have been recently implemented in engineering practice to predict the CBR value from the soil index properties with satisfactory results. However, they provide only deterministic predictions, which do not account for the aleatoric uncertainty linked to input variables and the epistemic uncertainty inherent in the model itself. This work addresses this limitation by introducing an ML model based on the natural gradient boosting (NGBoost) algorithm, becoming the first study to estimate the soaked CBR value from this probabilistic perspective. A database of 2130 soaked CBR tests was compiled for this study. The NGBoost model showcased robust predictive performance, establishing itself as a reliable and effective algorithm for predicting the soaked CBR value. Furthermore, it produced probabilistic CBR predictions as probability density functions, facilitating the establishment of reliable confidence intervals, representing a notable improvement compared to conventional deterministic models. Finally, the Shapley additive explanations method was implemented to investigate the interpretability of the proposed model.

Keywords: machine learning; CBR; soil index properties; subgrade; compaction characteristics; probabilistic model; explainable artificial intelligence



Citation: Díaz, E.; Spagnoli, G. Natural Gradient Boosting for Probabilistic Prediction of Soaked CBR Values Using an Explainable Artificial Intelligence Approach. *Buildings* **2024**, *14*, 352. <https://doi.org/10.3390/buildings14020352>

Academic Editors: Andrea Baliello and Di Wang

Received: 24 December 2023

Revised: 23 January 2024

Accepted: 24 January 2024

Published: 26 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Most civil engineering projects requiring earthwork, such as highways, earth dams, airport runways, and pavements, typically need the establishment of a suitable subbase layer, i.e., subgrade. Subgrade soil plays a vital role in the long-term performance of pavements because it gives them essential foundational support. The subgrade must meet specific engineering criteria, which include factors like bearing capacity, settlement, and swell properties. The California bearing ratio (CBR) is one of the most frequently used tests to assess the stiffness and bearing capacity of subgrade [1]. The CBR is essentially an indirect measure that compares the strength of a specific soil to the strength of a standard crushed rock, typically expressed as a percentage value. The CBR test is considered a penetration test, initially introduced by the California State Highway Department, USA, to assess the subgrade in the process of flexible pavement design. The CBR test involves using soil previously compacted in a cylindrical mould to perform loading tests on the soil's surface with the aid of a plunger [2,3]. The CBR value of any soil sample can be established in soaked and unsoaked conditions. Nevertheless, the CBR value obtained for soaked soil samples is typically lower than the value obtained for unsoaked samples, and CBR values for soaked soil samples are often considered a conservative estimate for design purposes. On the other hand, the soaked CBR test is time-consuming and labour-intensive because it involves soaking the soil sample underwater for 96 h to simulate the

worst-case scenario when the pavement is submerged. Additionally, the soaked CBR test demands a substantial quantity of materials, nearly 6 kg, and involves greater effort in preparing the test specimen. Furthermore, it is important to note that the subgrade soil's properties can exhibit substantial variations even over short distances, primarily due to the heterogeneous nature of soil. This variability can make it challenging to establish consistent design parameters. When building a new road or an earth dam, it becomes necessary to collect numerous soil samples and then estimate the CBR value for each of these samples using the CBR test. This process typically takes at least 4 days per sample (5 to 6 days, including the required modified Proctor test). This can lead to delays in obtaining test results, which may not be practical for making quick decisions related to pavement design, potentially resulting in a significant increase in construction costs. Moreover, the testing method involves expenses related to material transportation (from the construction site to the laboratory) and testing fees. Additionally, test results can occasionally be unreliable due to sample disturbance and issues with the quality of the laboratory conditions. Therefore, it can be challenging to ensure that the soil sample used for the CBR test accurately reflects the in situ conditions. Considering the points previously discussed, many works have tried to establish correlations between the CBR value and readily determinable index properties of soils. These tests are easy to conduct and are typically performed on soil samples as soon as they are brought to the laboratory. The study conducted by Jumikis [4] is one of the early attempts to establish a correlation between CBR values and Atterberg limits. Black [5] formulated a correlation to estimate the CBR value of cohesive soils, utilizing the plasticity index (PI) and the liquidity index (LI) as key parameters. Ring [6] focused on estimating the CBR value using Atterberg limits along with key compaction parameters, i.e., optimum moisture content (OMC) and maximum dry density (MDD). De Graft-Johnson, Bhatia [7] attempted to estimate the CBR value based on Atterberg limits and the soil's grain size distribution. Katte, Mfoyet [8] used the linear regression technique (LR) to establish a relationship between the CBR value and MDD. Additionally, they developed a group of multiple linear regression (MLR) models for estimating the CBR value, taking into account factors such as grain size distribution, Atterberg limits, OMC, and MDD. Patel and Desai [9] used MLR to establish correlations between CBR values and the liquid limit (LL), plastic limit (PL), PI, OMC, and MDD of cohesive soils. Hassan, Alshameri [10] proposed a set of MLR models for estimating the CBR value based on the grain size distribution, Atterberg limits, OMC, and MDD. Thus, numerous researchers have made several attempts to establish a relationship between the soaked CBR value and parameters associated with grain size distribution (such as gravel, sand, and fine content as well as the particle sizes corresponding to 10%, 30%, 50%, and 60% finer materials on the cumulative particle size distribution curve), Atterberg's limits (LL, PL, and PI), and compaction parameters (OMC and MDD). However, a comprehensive review of some conventional studies conducted by Taskiran [11] revealed that satisfactory correlations based on LR or MLR could not be attained in many cases because most of the proposed empirical equations lacked a high degree of accuracy and did not provide a generalized solution. Wang and Yin [12] analysed several traditional expressions and concluded that these expressions could not be considered highly reliable. This was primarily because most of these predictive equations were developed through regression analysis using a limited dataset, typically consisting of a small number of samples, ranging from 20 to 158, and were specific to certain soil types. As a result, it was noted that the degree of accuracy of these relationships was generally unsatisfactory in many cases, leading to poor predictions of CBR values. Therefore, more sophisticated and satisfactory approaches should be employed to estimate CBR values using soil properties. It should be noted that machine learning (ML), with its ability to perform nonlinear modelling, offers a viable tool for simulating complex processes [13,14]. Since the development of ML techniques, numerous studies have adopted them, considering the same input parameters as traditional correlations, and have achieved superior performance metrics compared to traditional methods (e.g., [15,16]). The study of Taskiran [11] marked the first application of ML techniques to predict the

CBR value in soils. A summary of these can be consulted in Bardhan, Gokceoglu [17], Khasawneh, Al-Akhrass [18], Othman and Abdelwahab [19], or Verma, Kumar [20]. It is worth emphasizing the study conducted by Bardhan, Gokceoglu [17], in which they employed four soft computing techniques and analysed 312 experimental results for soaked CBR values. They concluded that multivariate adaptive regression splines with piecewise linear models achieved the highest level of accuracy with a coefficient of determination (R^2) value of 0.969. On the other hand, this study analysed most of works where ML techniques were applied to predict the CBR value [11,21–35]. The study concluded that the works conducted with small datasets (consisting of fewer than 160 data points) exhibited higher predictive accuracies, with R^2 values between 0.81 and 1.00, in contrast to the studies conducted with larger datasets ranging from 358 to 389 soil test results [21,22]. When small datasets are utilized for ML models, they often yield exceptionally high performance; however, these models tend to suffer from overfitting and lack generalization ability. Conversely, the results also indicated that the prediction accuracy of ML models decreased as the number of samples increased, with accuracies ranging from 0.78 to 0.80 in terms of R^2 , which is less than desirable. It is worth noting that studies based on larger datasets are generally considered to be more reliable than those conducted with a limited number of samples. Posteriorly, Hao and Pabst [36] compiled 39 samples for predicting CBR values and the resilient modulus of crushed waste rock. They used MLR, k-nearest neighbours (KNN), decision tree (DT), random forest (RF), multilayer perceptron (MLP), and neuroevolution of augmenting topologies (NEAT). The RF model provided better results with an R^2 value of 0.926. Khasawneh, Al-Akhrass [18] utilized 110 data points and investigated three ML techniques, namely, artificial neural networks (ANNs), M5P model tree, and KNN, alongside conventional methods such as MLR and nonlinear regression (NLR). The study concluded that the most effective model for predicting the CBR value was the ANN model, which achieved an R^2 value of 0.905, followed by KNN, MLR, M5P, and NLR in descending order of performance. Khatti and Grover [37] compiled 182 training, 15 testing, 36 validation, and 12 cross-validation data points for fine-grained soil. They employed gene expression programming (GEP) and relevance vector machine (RVM) models for predicting the soaked CBR value. The genetic algorithm-optimized Laplacian kernel-based SRVM model outperformed the others, achieving a correlation coefficient (r) of 0.9874. Othman and Abdelwahab [19] developed 240 ANNs with various hyperparameters. They achieved an R^2 value of 0.945, surpassing the performance of traditional MLR models. The size of the dataset was 77 samples. Verma, Kumar [20] collected 1011 soil samples and used kernel ridge regression, KNN, and Gaussian process regression (GPR) to predict the soaked CBR value of soils. GPR exhibited the highest performance, achieving an r value of 0.83 on the test set. Kamrul Alam and Shiuly [38] considered three distinct models developed using three different methodologies: a fuzzy inference system (FIS), an ANN, and an adaptive neuro-fuzzy inference system. To conduct the study, a dataset of 2000 soil samples was collected. The hybrid model of ANN and FIS (ANFIS) outperformed both ANN and FIS in terms of predictive accuracy, achieving an r value of 0.92. It should be noted that in some cases, when dealing with highly complex datasets, a better/more optimized network or model may not necessarily improve the model accuracy. Reducing the complexity of a dataset can also potentially improve the accuracy of the ML model [39,40]. Thus, in order to obtain good performance metrics, it is necessary to employ ML models to predict the CBR value based on a large number of samples, which allows the algorithm to generalize when it is tested with new samples.

However, ML methods have limitations for field applications, presenting two main challenges. Firstly, the uncertainty in predicted results makes reliability analysis difficult. Secondly, the models often lack interpretability. To enable reliability analysis, quantifying the uncertainty for each data point is necessary. There are two primary sources of uncertainty in data sets: (1) aleatoric uncertainty stems from the inherent variability in the data generation process, and (2) epistemic uncertainty is caused by limitations in the predictive capabilities of the model. Probabilistic forecasts hold greater value than deterministic ones

because they offer additional information about prediction uncertainty. These data can be utilized by engineers to reduce risk and make more advantageous decisions. Thus, the deterministic perspective offers insight into the model's accuracy in making precise predictions, while the probabilistic perspective sheds light on the model's ability to understand and quantify the uncertainty inherent in those predictions. Together, they provide a nuanced and in-depth evaluation of the model's overall effectiveness and reliability in a regression setting. On the other hand, due to the growing complexity of these models, interpreting their results presents a significant challenge. A recently developed technique in the field of explainable artificial intelligence (XAI), known as Shapley additive explanation (SHAP) within the ML community, can be employed to interpret these models. Using the SHAP explainable ML model enables the determination of the importance and directional impact of each input variable on the predicted results, thereby providing insights into the relationships between input variables and the output.

Thus, this study attempted to predict the soaked CBR values through a probabilistic approach, using variables previously correlated with it. These variables encompassed gravel content (G), sand content (S), fine content (F) expressed in terms of silt and clay content, LL, PI, OMC, and MDD.

The main contributions of this study include the following: (1) the proposal and application of a reliable model for predicting the soaked CBR value considering a large database to ensure proper generalisation, (2) the ability of the proposed model to provide not only accurate point prediction but also estimation of the predictive uncertainty for reliable decision-making, and (3) an explanation approach based on explainable artificial intelligence to understand the importance of the input factors and estimate the quantitative results of the impact of every feature on the soaked CBR value.

2. Methodology

In the present investigation, the variables G, S, F, LL, PI, OMC, and MDD were used as input parameters to predict the soaked CBR value. For this aim, the study followed the framework illustrated in Figure 1. The adopted framework consists of five main parts:

- (1) Data preparation. In this case, an outlier analysis and a normalization of the values were performed.
- (2) ML model selection. Four probabilistic ML models were compared to select the best one.
- (3) ML model optimization. The artificial bee colony algorithm was selected to obtain the definitive fine-tuned model.
- (4) ML model evaluation. First, it is evaluated from a deterministic perspective and then from a probabilistic perspective to establish the confidence intervals and uncertainty in the predictions.
- (5) Implementing an explanation approach based on XAI.

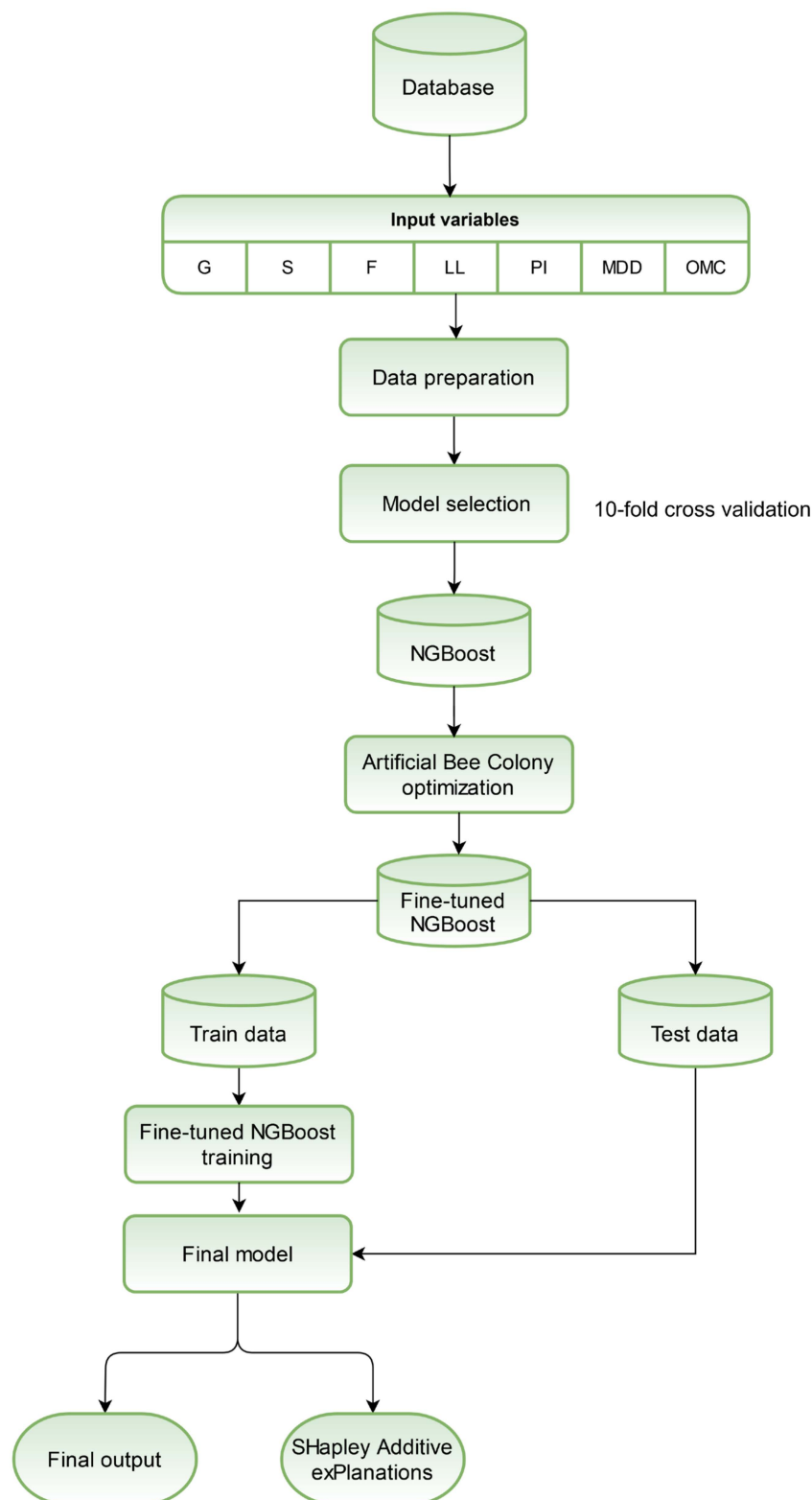


Figure 1. Framework of the methodology performed.

3. Database

A total of 2130 soil samples were collected from several construction projects in Spain. These soil samples were divided into training and testing sample sets in an 80/20 ratio. The laboratory tests were conducted in accordance with Spanish standard specifications, which are mostly equivalent to European standards. These tests were grain size distribution [41],

Atterberg limits [42], modified Proctor [43], and the soaked CBR [44]. Through these laboratory tests, numerous geotechnical parameters were collected such as the gravel content (G), sand content (S), silt and clay content termed fine content (F), liquid limit (LL), plasticity index (PI), maximum dry density (MDD), optimum moisture content (OMC), and soaked CBR value. The laboratory-obtained soil database comprised a wide spectrum of soils, ranging from soils classified as gravel according to the USCS, accounting for 33.5% of the total, sands at 29.4% of the total, and fine soils making up 37.1% of the total. Regarding plasticity characteristics, there were non-plastic soils or low-plasticity soils, i.e., $LL < 50$ (97.5%), and high-plasticity soils, i.e., $LL \geq 50$ (2.5%). The main statistics of the considered properties are summarized in Table 1.

Table 1. Descriptive statistics of the dataset.

	G (%)	S (%)	F (%)	LL (%)	PI (%)	MDD (g/cm ³)	OMC (%)	CBR
Mean	21.08	38.18	40.75	28.18	9.89	1.97	10.28	29.01
Std. deviation	27.01	18.11	26.34	10.29	7.97	0.23	3.66	27.23
Minimum	0	1.80	0.70	0	0	1.46	4.10	0.90
25th percentile	0	26.20	15.40	20.60	4.00	1.79	6.50	7.10
Median	1.90	33.80	39.40	25.90	7.30	1.96	11.40	17.50
75th percentile	53.80	50.50	61.88	33.30	13.60	2.12	12.90	55.00
Maximum	73.90	90.20	98.20	82.30	53.80	2.52	20.00	144.00
Skewness	0.69	0.54	0.34	1.65	1.89	0.13	−0.17	0.91
Kurtosis	−1.38	−0.20	−1.04	5.04	5.82	−0.71	−0.93	−0.59

From the analysis of Table 1, it can be concluded that in the dataset, G had a mean of 21.08%, with a high standard deviation of 27.01%, suggesting substantial variability. The skewness and kurtosis indicated a moderate skew and a flatter than normal distribution. S had a mean of 38.18% and a standard deviation of 18.11% reflecting moderate variability. The data followed a slightly skewed, nearly normal distribution. F showed a mean of 40.75% and a standard deviation of 26.34%, indicative of a broad spread of data points. The skewness and kurtosis pointed to a slightly right-skewed and somewhat flatter distribution. LL had a mean of 28.18% and a standard deviation of 10.29%, showing lesser but significant variation. The data indicated a significantly skewed and leptokurtic distribution. The PI presented a mean of 9.89% and a standard deviation of 7.97%, suggesting moderate variability. Its high skewness and kurtosis indicated a highly skewed and leptokurtic distribution. MDD had a mean of 1.97 g/cm³ and a low standard deviation of 0.23 g/cm³, showing very little variation in values. MDD followed a nearly symmetric, slightly platykurtic distribution. OMC had a mean of 10.28% and a standard deviation of 3.66%, indicating moderate variability. Its skewness of −0.17 and kurtosis of −0.93 suggested a nearly symmetric, slightly flat distribution. Lastly, CBR values had a median of 17.5, a mode of 2.0, a mean of 29.01, and a high standard deviation of 27.23, showing a wide spread in these values. The distribution of CBR values exhibited a moderate right skew, indicating that there is a clustering of data points on the lower end of the scale with a few higher-value data points. Additionally, it displayed a flatter peak and lighter tails compared to a normal distribution, suggesting a lower occurrence of extreme values.

In Figure 2, scatter plots of the analysed parameters are depicted, providing a descriptive overview of the data distribution. Furthermore, each scatterplot included the best-fit line and the correlation coefficient (r) value for added context. These graphs are useful for detecting linear correlations between variables, as if it were a correlation matrix.

The analysis of these plots revealed several interesting relationships between variables. A strong negative correlation existed between G and F (−0.77), suggesting that as one increases, the other tends to decrease significantly, an obvious fact. Similarly, the F and CBR values showed a strong negative correlation (−0.76). This pattern was also seen with G and OMC (−0.83), indicating inverse relationships. Conversely, G positively correlated with MDD (0.73) and the CBR value (0.83), implying that these pairs of variables tend to

increase or decrease together. LL and PI exhibited a very high positive correlation (0.90), suggesting they are likely to change in the same direction, which is quite normal as these variables are highly related. Other correlations, such as S with G (-0.37) and S with MDD (-0.31), were less pronounced but still notable.

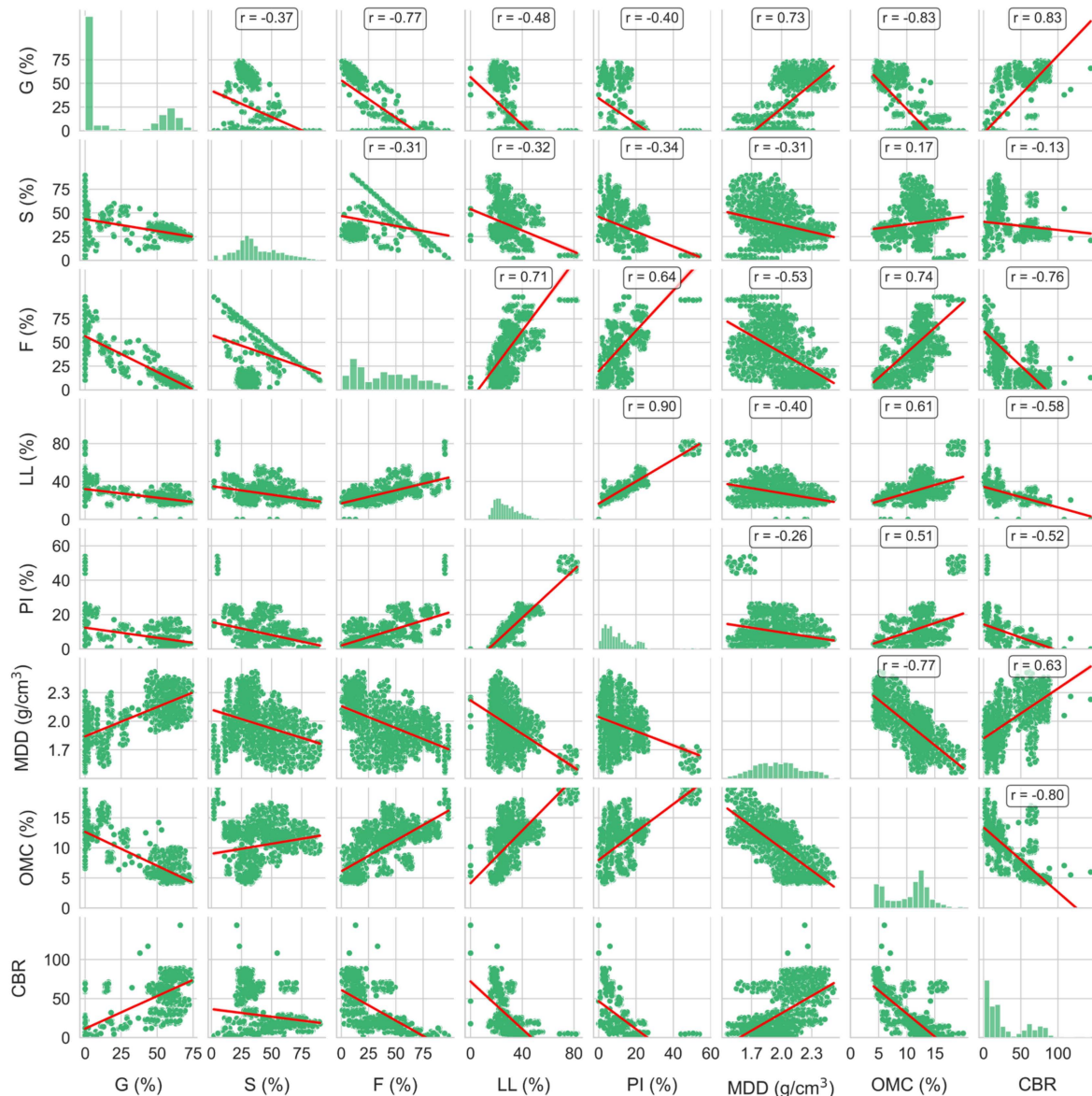


Figure 2. Correlation analysis including scatter plots and distribution histograms of the considered variables.

4. Machine Learning Process

4.1. Data Preparation

Before starting the ML process, it is necessary to conduct an outlier study on the dataset because outliers can significantly distort statistical analyses and predictive modelling, leading to inaccurate results. Identifying these anomalies ensures more reliable and generalizable insights from the data. The isolation forest (IF) algorithm [45] was employed for this purpose. This is an effective algorithm for anomaly detection and is particularly suited for high-dimensional datasets. It operates on the principle that anomalies are data points that are few and different, making them easier to 'isolate' than normal points. The algorithm selects a feature at random and then chooses a split value between the maximum and minimum values of the selected feature. This process creates 'isolation trees' or

random decision trees. For each data point, the path length from the root to the node is calculated. Anomalies are expected to have shorter path lengths on average, as they are more susceptible to isolation. This path length forms the basis for the anomaly score, which is used to determine outliers: the shorter the path length, the more anomalous the data point. Thus, in an IF, anomalies are indeed 'isolated' more efficiently compared to normal points. After applying the IF algorithm to our dataset, 64 data points were removed. On the other hand, a normalization process was conducted to standardize the data values to a common scale while preserving the original differences in value ranges. This was achieved using min-max normalization, which scaled the data between 0 and 1.

4.2. Model Selection

Once the initial data were processed, four probabilistic ML algorithms were initially selected as candidates to choose the optimal prediction model through a cross-validation process with $k = 10$. The selection of $k = 10$ in k -fold cross-validation is a widely accepted practice that strikes a balance between bias and variance in the performance estimation of ML models [46]. Opting for a higher k value can decrease variance but may lead to increased bias. Conversely, a lower k tends to increase variance while reducing bias. Therefore, taking into account the dataset's size, $k = 10$ constitutes a practical compromise, ensuring that each fold includes a sufficiently large sample size and offering 10 distinct training-testing splits for comprehensive model evaluation. In this case, the following models were used: natural gradient boosting (NGBoost), quantile regression forest, Bayesian ridge regression, and Gaussian process regression.

NGBoost [47] represents a probabilistic model based on gradient boosted decision trees. However, in contrast to quantile regression or conditional mean estimation, NGBoost is designed to learn the complete probability distribution as an inherent part of its methodology. The fundamental concept behind NGBoost involves the utilization of natural gradients instead of traditional gradients, enabling the algorithm to effectively model a probability distribution across the outcome space, taking into account the predictor variables. Quantile regression forest [48,49] is a non-parametric method for estimating conditional quantiles, offering flexibility for modelling complex relationships and handling non-Gaussian distributions. It utilizes an ensemble of decision trees to predict different quantiles of the response variable, providing a comprehensive view of uncertainty in high-dimensional datasets. Bayesian ridge regression is a probabilistic technique employed to formulate and estimate statistical models [50]. Its probabilistic perspective offers significant advantages, particularly in scenarios where the data distribution is non-ideal or when the dataset is limited. Bayesian ridge regression leverages probability distributions to make predictions. Instead of yielding a single point estimate, this approach provides a full probability distribution over the predicted values. Gaussian process regression (GPR) [51] is a non-parametric approach that stands out in making probabilistic predictions by modelling data points as variables in a Gaussian process. By utilizing a covariance kernel, it effectively captures intricate non-linear relationships and dependencies between data points. Through Bayesian inference, GPR updates beliefs about the underlying function, providing predictions as Gaussian probability distributions, characterized by the mean and variance.

In Figure 3 are shown raincloud plots representing the distribution of R^2 and mean absolute error (MAE) values obtained in the cross-validation process. A raincloud plot combines elements of a half-violin plot, a boxplot, and a strip plot. This comprehensive visualization enables full statistical interpretation of the data. Specifically, it showcases the distribution's shape (half-violin plot), the five-number summary statistics including the minimum, first quartile (25th percentile), median, third quartile (75th percentile), and maximum (boxplot component), and the sample size and exact positions of individual data points (strip plot). Additionally, their means are highlighted in a darker colour. This format provides a nuanced view of the data, highlighting both its aggregate properties and individual variations. In this figure, to enhance visualization, Gaussian process regression is not included due to its low metrics (R^2 below 0.5).

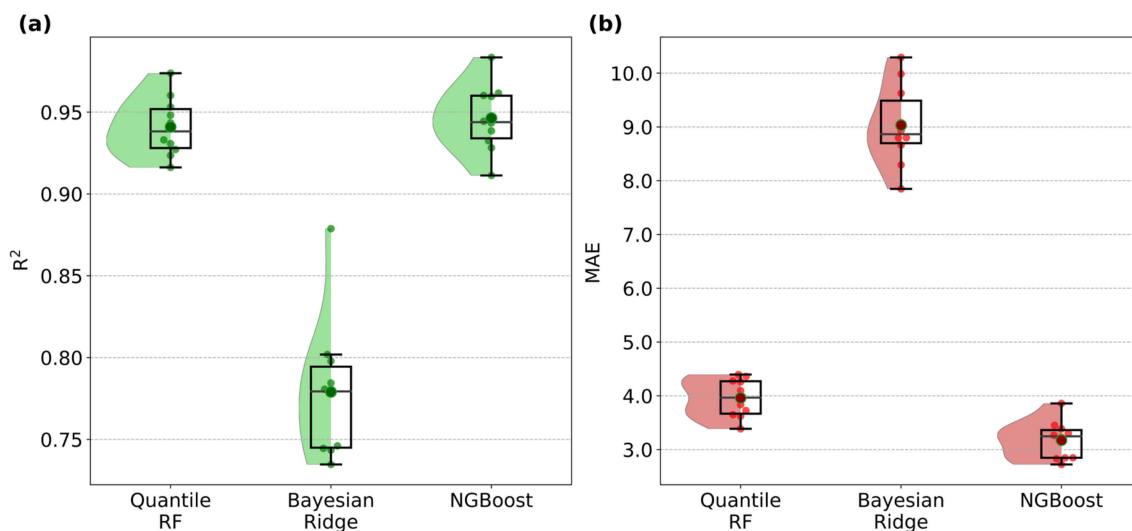


Figure 3. Raincloud plots of the initial probabilistic models considered. (a) R^2 , (b) MAE.

Based on the analysis of Figure 3, it is evident that NGBoost exhibits the best overall performance among the algorithms, followed by quantile regression forest and Bayesian ridge. Conversely, Gaussian process demonstrates the poorest performance. Consequently, NGBoost was selected as the definitive model.

4.3. Artificial Bee Colony Optimization

After conducting the model selection study and determining that the algorithm to be used in this study for predicting the CBR value is NGBoost, an optimization of its hyperparameters was carried out. To achieve this, the artificial bee colony (ABC) algorithm was employed. The ABC algorithm [52] is a metaheuristic optimization algorithm inspired by the foraging behaviour of honey bee colonies. It works by iteratively updating a population of artificial bees, which represent potential solutions to the optimization problem. The ABC algorithm has three types of bees: employed bees, onlooker bees, and scouts. Employed bees explore the search space and evaluate the quality of the solutions they find. Onlooker bees then choose solutions to explore based on the quality of the solutions found by employed bees. Scouts explore new areas of the search space. The ABC algorithm is a powerful and versatile optimization technique that can be used to solve a wide range of complex optimization problems (e.g., [53]). This process was performed three times. The first time was used to choose the base model of NGBoost where the following were considered the next models: RF, extremely randomized trees (ET), and DT. The best base model was ET, and then this model was tuned by the ABC algorithm. The parameters to be adjusted in the ET base model were the number of trees in the forest ($n_estimators$), the maximum depth of them (max_depth), the minimum samples required for node splitting ($min_samples_split$), the minimum number of samples required in a leaf node ($min_samples_leaf$), and the maximum features to consider for node splitting ($max_features$). In Table 2, the search values and the selected values for each of them are included. In all optimization processes carried out, the search ranges for the hyperparameters were primarily determined through the collective expertise of the authors and common values in ML. Furthermore, an additional layer of validation in the selection process ensured that the chosen hyperparameters were not at the extremes of their respective ranges.

Once the base model was optimized, the process was repeated with NGBoost. For this purpose, a normal distribution was set as the probability distribution to model the predictions. The tuned ET base model was also used, and the optimization process was carried out using the ABC algorithm. In this case, the hyperparameters to be optimized were the number of trees in the forest ($n_estimators$), the fraction of data per mini-batch during training ($minibatch_frac$), and the step rate at which the model learns from the

training data (learning_rate). The results of this optimization process are summarized in Table 3.

Table 2. Inputs for tuning hyperparameters of the ET base model using ABC.

Hyperparameter	Search Range	Optimal Value
n_estimators	1–200	51
max_depth	1–20	10
min_samples_split	2–15	10
min_samples_leaf	1–15	4
max_features	sqrt, log2, None	None

Table 3. Inputs for tuning hyperparameters for the NGBoost model using the ABC algorithm.

Hyperparameter	Search Range	Optimal Value
n_estimators	1–200	50
minibatch_frac	0–1	0.5
learning_rate	0.005–0.5	0.05

4.4. Fine-Tuned NGBoost

After tuning the hyperparameters, the mean prediction accuracy of the fine-tuned NGBoost was assessed using the test set. In this way, the training and testing sets were divided in an 80/20 proportion. The metrics obtained are included in Table 4. This analysis included R^2 , MAE, root mean squared error (RMSE), and the a20-index [54]. The a20-index provides a clear and practical interpretation for engineering applications. It specifically denotes the percentage of samples where the predicted values are within a $\pm 20\%$ range of deviation from the experimentally measured values.

Table 4. Summary of performance metrics for the fine-tuned NGBoost.

Set	R^2	MAE	RMSE	a-20 Index
Train	0.975	1.97	3.68	0.9025
Test	0.967	2.53	4.26	0.8697

In Figure 4a are presented scatterplots comparing actual versus measured CBR values, showcasing the performance of the NGBoost model post hyperparameter fine-tuning for predicting the CBR mean values. From the analysis of Table 4 and Figure 4a, it can be observed that the NGBoost algorithm has been adequately trained, exhibiting similar performance on both the training and testing sets. On the testing set, the R^2 was 0.967, and the RMSE and MAE values were 4.26 and 2.53, respectively. Nearly 87% of the mean prediction results fell within the $\pm 20\%$ relative error boundary for the testing set. This suggested that NGBoost's regression capacity for scalar prediction was satisfactory, and it offers the added benefit of probabilistic predictions. Figure 4b includes the residuals obtained by comparing the model's mean predictions with the actual measured values for both the training and testing datasets. The results from the analysis of residuals indicated that in the training set, the mean of residuals was 0.16, the median was -0.07 , and the standard deviation of residuals was 3.25. In the testing set, the mean of residuals was 0.09, the median was -0.15 , and the standard deviation of residuals was 3.74. Generally, the analysis of the residuals and error metrics for both the training and testing sets revealed that the metrics were quite similar across both sets. This similarity in metrics suggested that the model performed consistently on both the training and testing data. Regarding the Durbin–Watson statistic, it was 1.999 for the residuals in the training set and 2.121 for the testing set. The Durbin–Watson statistic ranges from 0 to 4, where a value close to 2 suggests there is no significant autocorrelation in the residuals. The obtained values, being approximately 2, indicated that there was likely no positive or negative autocorrelation in the residuals of

both sets. This suggested that the model was effective in capturing correlation in the data and that the residuals were approximately independent of each other.

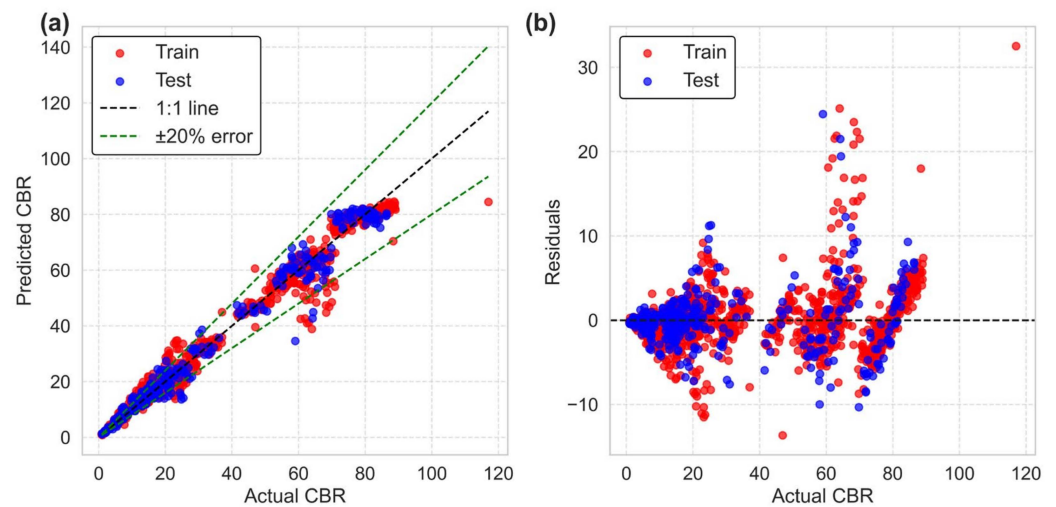


Figure 4. Fit between the mean predictions and the actual values for the training and testing sets in the NGBoost model. (a) Actual vs. predicted CBR values for the training and testing sets, (b) residuals of the predictions.

In Figure 5, the ratios between the test results and predictions are shown as functions of database variables. The NGBoost model demonstrated a consistently reasonable level of prediction accuracy across the entire range of database variables. The slopes of the linear regression lines (in pink) were nearly zero for all database features, confirming the suitability of the chosen input parameters. As mentioned earlier, almost all predictions fell within a $\pm 20\%$ error range, with no observed systematic deviations in the value ranges for any variables.

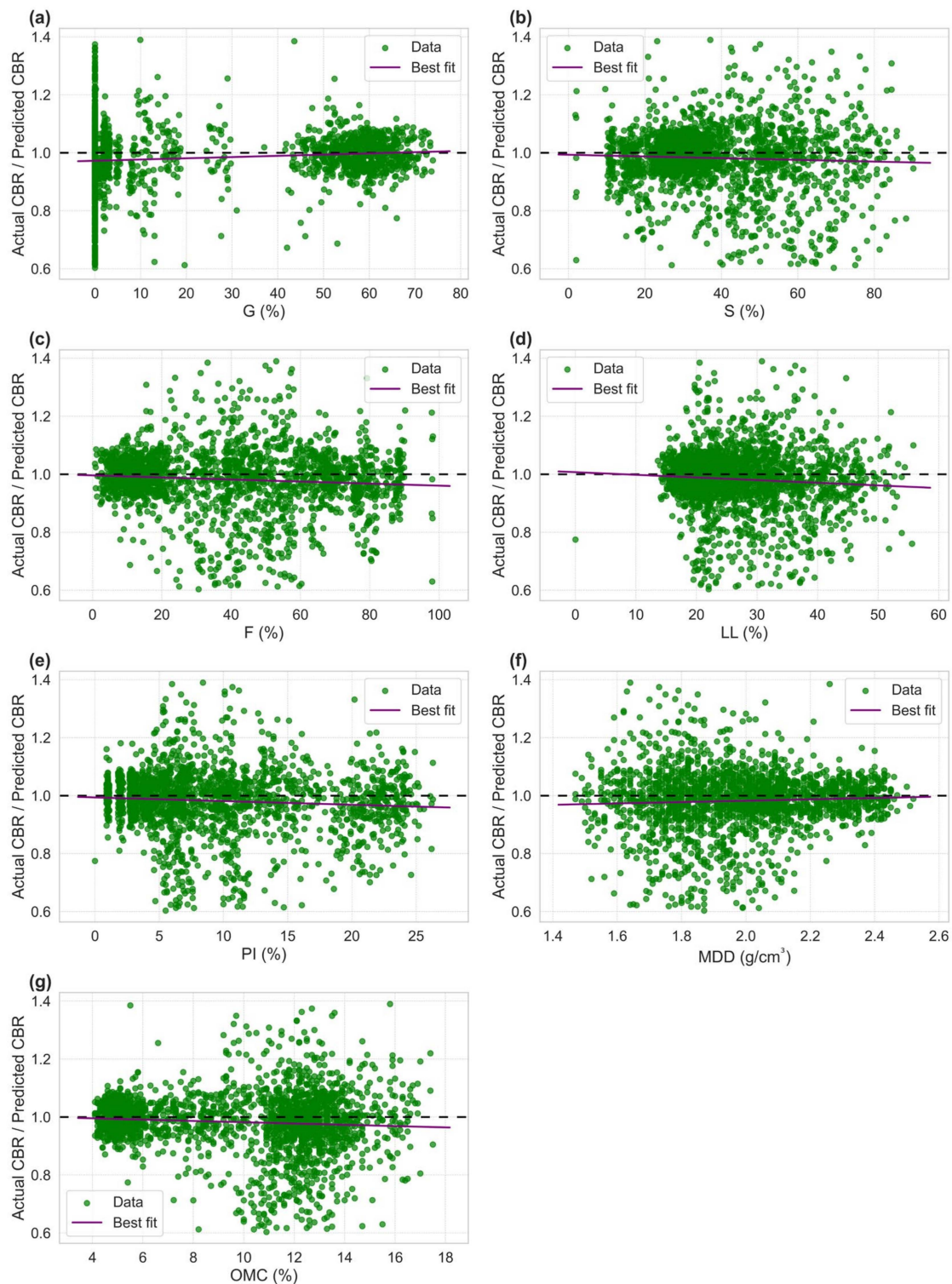


Figure 5. Prediction ratios versus database variables considering the entire dataset. (a) Gravel content, (b) sand content, (c) fine content, (d) liquid limit, (e) plasticity index, (f) maximum dry density, and (g) optimum moisture content.

5. Analysis of Uncertainty in the Predictive Model

In this section, the applicability of the probabilistic approach of NGBoost for predicting soaked CBR values in soils is analysed. It should also be highlighted that, in addition to providing reasonably accurate point estimates, the NGBoost-based probabilistic prediction model can also offer the variance of the point to account for the inherent uncertainty in

the response. The NGBoost’s probabilistic predictions, based on a normal distribution, are displayed in Figure 6. In addition to the mean prediction (μ), two confidence intervals, dependent on the standard deviation (std), were also provided using the NGBoost model. It can be observed that the majority of the test data (77.44%) fell within the interval $\mu \pm \text{std}$, corresponding to a confidence level of 68.2%. Furthermore, 96.15% of the test data fell within the interval $\mu \pm 2 \text{ std}$, corresponding to a confidence level of 95.4%. The evaluation of the NGBoost model for CBR values below 20 demonstrated a high level of reliability in the estimations, as evidenced by the average width of the confidence intervals, which was 4.88 for 68.2% and 9.75 for 95.4%. This suggested that the model provided estimations with a reasonably well-defined degree of uncertainty, which is crucial for users to understand the potential margins of error in their analyses. Furthermore, the model’s reliability was reinforced by the fact that 92.41% of the actual data fell within the $\mu \pm \text{std}$ interval and that 98.66% fell within the $\mu \pm 2 \text{ std}$ interval. This indicated a strong ability of the model to predict accurately within this range of CBR values. This is significant because this range of values is where soil values are predominantly found, especially in soils where the greatest support capacity problems are encountered. On the other hand, the evaluation of the NGBoost model for CBR values above 40 exhibited distinct characteristics when compared to lower CBR values. The average width of the confidence intervals was 8.88 for 68.2% and 17.76 for 95.4%. These wider values suggested that the model provided estimations with a higher degree of uncertainty in this upper range of the CBR value, which is an important consideration for users when interpreting the results and understanding potential margins of error. In terms of reliability, it could be observed that 57.26% of the actual data fell within the $\mu \pm \text{std}$ interval and that 94.02% fell within the $\mu \pm 2 \text{ std}$ interval. Although most of the actual values still fell within these confidence intervals, the proportion was significantly lower compared to the CBR range below 20. This indicated a greater level of uncertainty in predictions for soils with higher bearing capacity. This fact can be explained by the scarcity of samples in the training dataset for CBR values > 40 (approximately half compared to $\text{CBR} < 20$), which means that the model had less information to learn the characteristics of that range. This could lead to more uncertain estimations because the model was not sufficiently exposed to data in that specific range. For the range of CBR values between 20 and 40, the behaviour was intermediate between the ones analysed. Obviously, these findings could not have been obtained by deterministic ML methods.

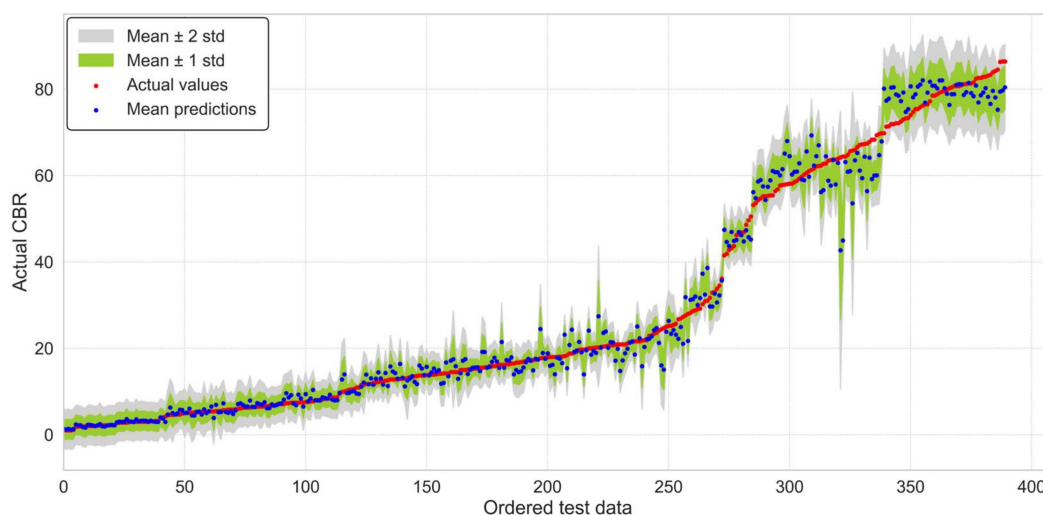


Figure 6. Performance of the NGBoost model for predicting the probabilistic distribution of CBR values. Note that the predicted CBR values are in ascending order. The grey and green areas indicate the $\mu \pm 2 \text{ std}$ (95.4%) and $\mu \pm \text{std}$ (68.2%) prediction intervals, respectively.

Additionally, to showcase the probabilistic prediction capabilities of the NGBoost model, Figure 7 includes three examples of density plots depicting the predicted probability

distributions for three specific data points. These data points were the closest to the CBR values of (a) 5.0, (b) 25, and (c) 50. The true value (red) and mean (green) are marked along with shaded regions for ± 1 std and ± 2 std from the mean. The distributions characterize the uncertainty in the predictions at those data points.

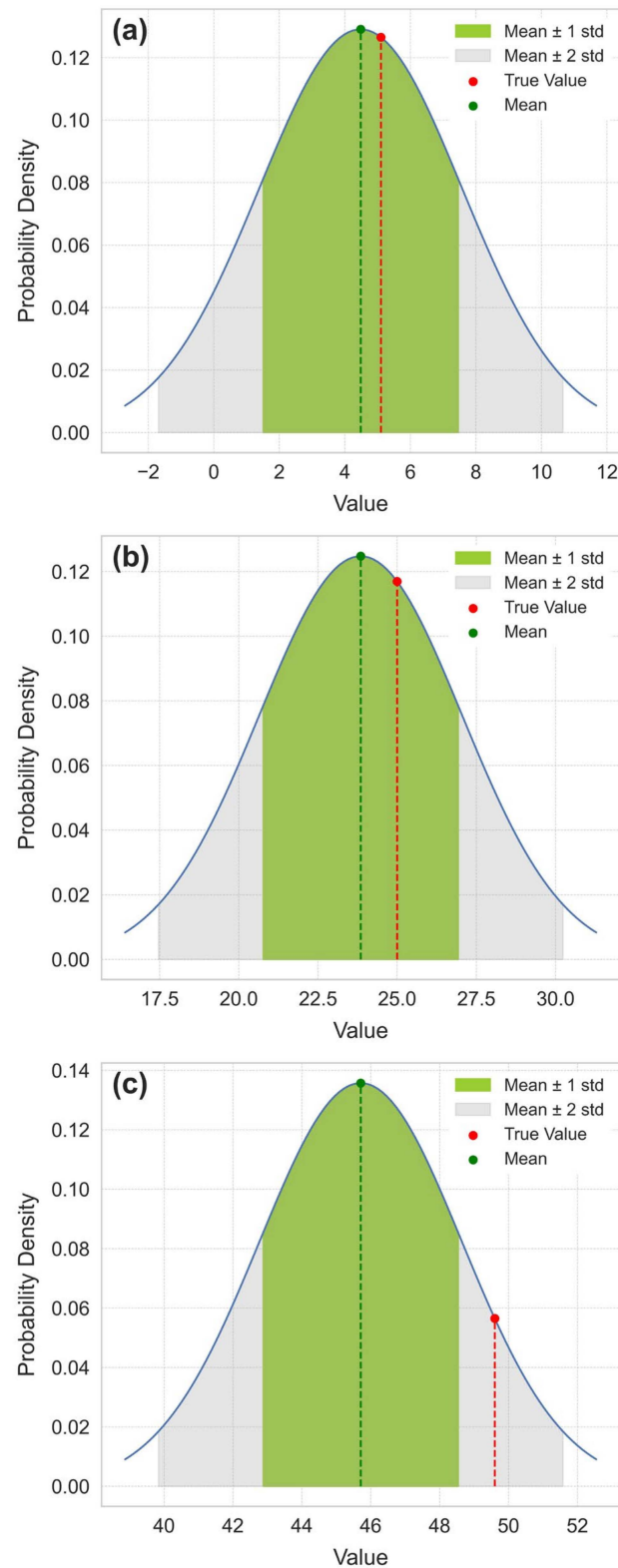


Figure 7. Probability density distributions for three selected data points in the test set. (a) CBR value closest to 5, (b) 25, and (c) 50.

These graphs provide a visual representation of how the NGBoost model generated probabilistic predictions for specific data points, which helps to understand the uncertainty associated with these predictions and how they relate to the real values. Consequently, engineers can have greater confidence in their decisions, especially when considering predetermined target reliability levels. These visualizations are valuable for the interpretation and evaluation of probabilistic models, something that is not achieved with average performance metrics as is usually used in deterministic ML models.

6. Interpretability Analysis of the Predictive Model

SHAP [55] was utilized to identify and interpret the impact of various factors on predicting the CBR value using the NGBoost probability model. SHAP is a method in ML that provides insights into how models make their predictions by utilizing concepts from cooperative game theory. Essentially, it breaks down a prediction to show the contribution of each feature. SHAP calculates this by considering the impact of including or excluding a particular feature on the model's output. It does this for every feature, across all possible combinations of other features, to ensure a fair attribution. This is based on the idea of Shapley values, which were originally developed to distribute payouts in cooperative games fairly. In the context of machine learning, these 'payouts' are the predictions, and SHAP assigns each feature a value that represents its contribution towards making the prediction. This method stands out for its ability to handle complex interactions between features and provide both global (across all predictions) and local (specific to one prediction) interpretations. By offering a detailed breakdown of feature contributions, SHAP helps in understanding, trusting, and improving ML models. SHAP enables a comprehensive analysis of how various soil properties and conditions impact the predicted CBR value, establishing SHAP as an indispensable tool for interpreting complex ML models. The SHAP values, representing the impact of a factor and a sample on the model output, are illustrated in the summary plot depicted in Figure 8. Additionally, in this plot, variables were ordered by their importance, with those having the greatest impact on the model's performance appearing at the top and those with lesser impact positioned towards the bottom. The summary plot serves to elucidate the ranking of factors by importance and their specific effects on the model's predictions. It visually depicts each factor's contribution, where the importance of a factor reflects its overall impact on the model's predictive performance, and its effect indicates the specific role in influencing individual predictions. The colour red signifies a positive influence where the factor increases the prediction value, while blue denotes a negative influence, reducing the prediction value. Furthermore, the positions of points further away from zero indicate stronger influences on the prediction outcome. Each point represents a SHAP value for an individual observation, offering a detailed view of how each factor alters the regression model's predictions. This dual interpretation of colour and distance from zero provides a comprehensive understanding of the factors' directional impact (positive or negative) and their magnitude of effect on the model's predictions.

From the analysis of the summary plot, it can be concluded that G was the most important variable for predicting the CBR value, and furthermore, an increase in G had a positive impact on the CBR value. Next in importance was the variable OMC, although in this case, high values of OMC tended to decrease the model's output. In the third position of importance was the variable F, indicating that high values of F decreased the model's output. The variable of greatest significance following F was PI, and its higher values decreased the model's output. Subsequently, was LL, which behaved similarly to OMC, F, and PI, as high values of LL tended to decrease the model's output. In the second-to-last position of importance was MDD, with a slight tendency for high values of MDD to positively impact the CBR value. Finally, the least important variable in the model was S, where the trend is not clear, with a fairly balanced distribution on both sides of zero. All of these findings were consistent with the established principles in soil mechanics and the known relationships among soil properties (e.g., [56,57]).

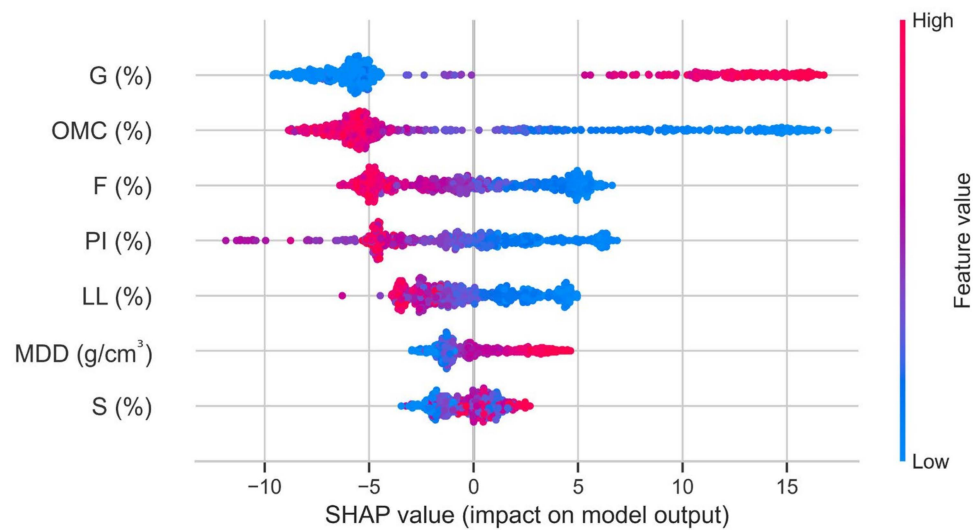


Figure 8. SHAP summary plot for the proposed model.

In Figure 9 are depicted the scatter plots of SHAP values, which aid in understanding the contribution of each feature to the model's predictions. Each point on the plots corresponds to an individual prediction from the model, with the position on the Y-axis reflecting the magnitude and direction of a feature's impact on that prediction. A higher SHAP value indicates a higher contribution to the model's prediction, while a lower value indicates a lower or negative contribution.

From the examination of the individual plots provided, it could be observed that G displayed a positive relationship that appeared slightly non-linear. In contrast, S showed widely dispersed points without a clear linear trend, indicating that the S feature may have influenced predictions in a complex, non-proportional way. F revealed a clearly negative and rather linear trend, where an increase in F correlated with a proportional decrease in the impact on the output. Turning to the LL, there was a negative trend that seemed non-linear. For PI, the trend was similar to that for LL. MDD showed a positive and linear relationship, with the impact on the output increasing in a proportional manner as MDD rose. Lastly, OMC indicated that the impact on the output significantly declined as OMC increased, suggesting a negative and slightly non-linear relationship, particularly stable at the higher range of OMC values. Considering the obtained results, it can be concluded that the algorithm has predominantly comprehended the relationship between each independent variable and the CBR value, aligning with the well-known relationships between properties of soils (e.g., [56,57]). In conclusion, a better understanding of the key factors affecting CBR values facilitates the optimization of construction practices for cost-effectiveness. This insight allows engineers to make informed choices regarding material selection, taking into account the variables' significance and impact direction. Such knowledge guides engineers in concentrating their efforts strategically to maximize benefits.

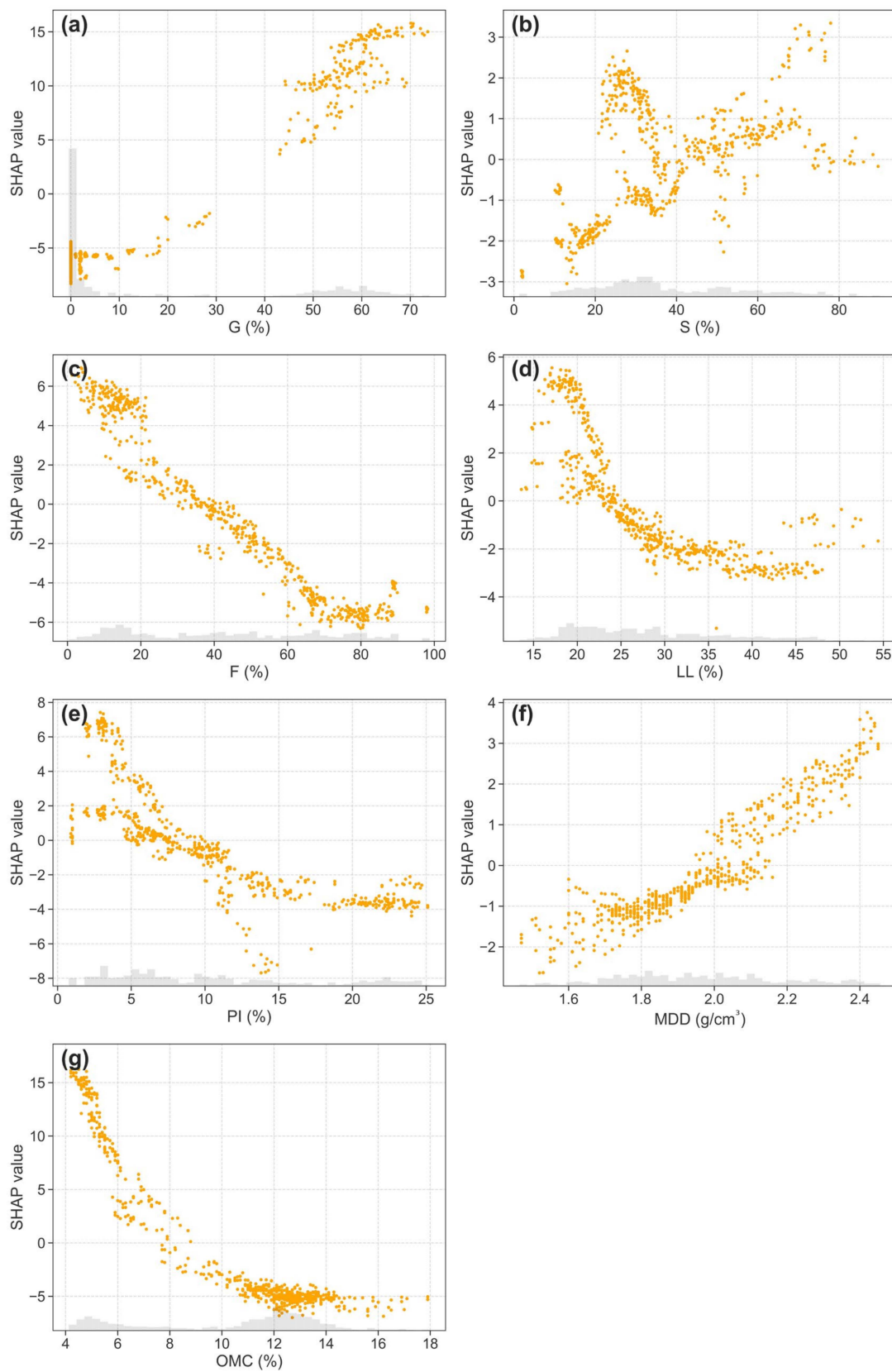


Figure 9. SHAP value scatter plots for the input variables. (a) Gravel content, (b) sand content, (c) fine content, (d) liquid limit, (e) plasticity index, (f) maximum dry density, and (g) optimum moisture content. Note that at the bottom of each plot, the distribution histogram of the data is displayed in light grey.

7. Conclusions

This paper presents an ML model based on the natural gradient boosting algorithm for probabilistic predictions of the soaked CBR value in soils. The model was developed employing a database of 2130 soaked CBR tests, which is considerably larger than the databases used to develop existing design models. A model selection study of probabilistic ML models revealed that NGBoost consistently achieved the best performance. Subsequently, the hyperparameters of the NGBoost model were extensively tuned and optimized using the artificial bee colony algorithm to enhance model predictions. The optimized NGBoost model demonstrated outstanding performance metrics on the testing set, with R^2 , MAE, and RMSE values of 0.967, 2.53, and 4.26, respectively. With regard to the obtained performance metrics, when compared to studies that do not consider small datasets, they are very similar to those obtained by Bardhan, Gokceoglu [17] and above the rest of the works. Furthermore, NGBoost was utilized to estimate predictive uncertainty and demonstrated that 96.15% of the actual soaked CBR values in the test dataset fell within the 95.4% confidence level. Additionally, the predictions of CBR values lower than 20 showed a high degree of reliability and low uncertainty. This demonstrated the model's robust ability to provide precise predictions within the range of CBR values lower than 20, which encompasses the majority of soil types. However, for CBR values exceeding 40, the model exhibits increased uncertainty and lower reliability compared to CBR values below 20. This implementation demonstrated a robust and effective approach, not only providing point predictions but also quantifying the associated uncertainty caused by the quality of the data and the number of samples. This is crucial in data-driven decision-making within contexts of uncertainty. As a result, a more reliable evaluation of the CBR value can be carried out. The optimized NGBoost model was interpreted using the SHAP method, which indicated that the gravel content, optimum moisture content, and fine content are the most important parameters affecting the soaked CBR value. Finally, it is important to emphasize that the algorithm was trained on a large database, but the samples are localized to a specific site, giving it a strong local character. Extrapolation to other areas should be conducted with caution.

Author Contributions: Conceptualization, E.D. and G.S.; methodology, E.D. and G.S.; software, E.D.; validation, E.D. and G.S.; formal analysis, E.D. and G.S.; investigation, E.D. and G.S.; data curation, E.D.; writing—original draft preparation, E.D.; writing—review and editing, E.D. and G.S.; visualization, E.D.; supervision, E.D. and G.S.; project administration, E.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to their current utilization for future works involving the authors of this paper.

Acknowledgments: The authors would like to express their gratitude to ITC, S.L. for their valuable assistance with the sample data.

Conflicts of Interest: Author Giovanni Spagnoli was employed by the company DMT GmbH & Co., KG. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Davis, E. The California bearing ratio method for the design of flexible roads and runways. *Géotechnique* **1949**, *1*, 249–263. [[CrossRef](#)]
2. BS, B.S. *Methods of Test for Soils for Civil Engineering Purposes*; British Standards Institution: London, UK, 1990.
3. *ASTM D1883-21*; Standard Test Method for California Bearing Ratio (CBR) of Laboratory-Compacted Soils. ASTM: West Conshohocken, PA, USA, 2016.
4. Jumikis, A.R. Geology of soils of the newark (NJ) metropolitan area. *J. Soil Mech. Found. Div.* **1958**, *84*, 1646-1–1646-41. [[CrossRef](#)]
5. Black, W. A method of estimating the California bearing ratio of cohesive soils from plasticity data. *Geotechnique* **1962**, *12*, 271–282. [[CrossRef](#)]
6. Ring, G. Correlation of compaction and classification test data. *Hwy. Res. Bull.* **1962**, *325*, 55–75.

7. De Graft-Johnson, J.; Bhatia, H.; Gidigas, D. The engineering characteristics of the laterite gravels of Ghana. In Proceedings of the 7th International Conference on Soil Mechanics and Foundation Engineering, San Fandila, Mexico, 21 July 1969; pp. 117–128.
8. Katte, V.Y.; Mfoyet, S.M.; Manefouet, B.; Wouatong, A.S.L.; Bezeng, L.A. Correlation of California Bearing Ratio (CBR) Value with Soil Properties of Road Subgrade Soil. *Geotech. Geol. Eng.* **2019**, *37*, 217–234. [[CrossRef](#)]
9. Patel, R.S.; Desai, M. CBR predicted by index properties for alluvial soils of South Gujarat. In Proceedings of the Indian Geotechnical Conference, Mumbai, India, 16–18 December 2010.
10. Hassan, J.; Alshameri, B.; Iqbal, F. Prediction of California Bearing Ratio (CBR) Using Index Soil Properties and Compaction Parameters of Low Plastic Fine-Grained Soil. *Transp. Infrastruct. Geotechnol.* **2022**, *9*, 764–776. [[CrossRef](#)]
11. Taskiran, T. Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. *Adv. Eng. Softw.* **2010**, *41*, 886–892. [[CrossRef](#)]
12. Wang, H.-L.; Yin, Z.-Y. High performance prediction of soil compaction parameters using multi expression programming. *Eng. Geol.* **2020**, *276*, 105758. [[CrossRef](#)]
13. Diaz, E.; Salamanca-Medina, E.L.; Tomás, R. Assessment of compressive strength of jet grouting by machine learning. *J. Rock Mech. Geotech. Eng.* **2023**, *16*, 102–111. [[CrossRef](#)]
14. Díaz, E.; Spagnoli, G. Gradient boosting trees with Bayesian optimization to predict activity from other geotechnical parameters. *Mar. Georesources Geotechnol.* **2023**, 1–11. [[CrossRef](#)]
15. Venkatasubramanian, C.; Dhinakaran, G. ANN model for predicting CBR from index properties of soils. *Int. J. Civil Struct. Eng.* **2011**, *2*, 614–620.
16. Sabat, A.K. Prediction of California bearing ratio of a soil stabilized with lime and quarry dust using artificial neural network. *Electron. J. Geotech. Eng.* **2013**, *18*, 3261–3272.
17. Bardhan, A.; Gokceoglu, C.; Burman, A.; Samui, P.; Asteris, P.G. Efficient computational techniques for predicting the California bearing ratio of soil in soaked conditions. *Eng. Geol.* **2021**, *291*, 106239. [[CrossRef](#)]
18. Khasawneh, M.A.; Al-Akhrass, H.I.; Rabab'ah, S.R.; Al-Sugaier, A.O. Prediction of California Bearing Ratio Using Soil Index Properties by Regression and Machine-Learning Techniques. *Int. J. Pavement Res. Technol.* **2022**. [[CrossRef](#)]
19. Othman, K.; Abdelwahab, H. The application of deep neural networks for the prediction of California Bearing Ratio of road subgrade soil. *Ain Shams Eng. J.* **2023**, *14*, 101988. [[CrossRef](#)]
20. Verma, G.; Kumar, B.; Kumar, C.; Ray, A.; Khandelwal, M. Application of KRR, K-NN and GPR Algorithms for Predicting the Soaked CBR of Fine-Grained Plastic Soils. *Arab. J. Sci. Eng.* **2023**, *48*, 13901–13927. [[CrossRef](#)]
21. Tenpe, A.R.; Patel, A. Utilization of Support Vector Models and Gene Expression Programming for Soil Strength Modeling. *Arab. J. Sci. Eng.* **2020**, *45*, 4301–4319. [[CrossRef](#)]
22. Al-Busultan, S.; Aswed, G.K.; Almuhan, R.R.A.; Rasheed, S.E. Application of artificial neural networks in predicting subbase CBR values using soil indices data. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *671*, 12106. [[CrossRef](#)]
23. Yildirim, B.; Gunaydin, O. Estimation of California bearing ratio by using soft computing systems. *Expert Syst. Appl.* **2011**, *38*, 6381–6391. [[CrossRef](#)]
24. Kumar, S.A.; Kumar, J.P.; Rajeev, J. Application of machine learning techniques to predict soaked CBR of remolded soils. *IJERT* **2013**, *2*, 3019–3024.
25. Bhatt, S.; Jain, P.K.; Pradesh, M. Prediction of California bearing ratio of soils using artificial neural network. *Am. Int. J. Res. Sci. Technol. Eng. Math* **2014**, *8*, 156–161.
26. Varghese, V.K.; Babu, S.S.; Bujukumar, R.; Cyrus, S.; Abraham, B.M. Artificial Neural Networks: A Solution to the Ambiguity in Prediction of Engineering Properties of Fine-Grained Soils. *Geotech. Geol. Eng.* **2013**, *31*, 1187–1205. [[CrossRef](#)]
27. Sabat, A.K. Prediction of California bearing ratio of a stabilized expansive soil using artificial neural network and support vector machine. *Electron. J. Geotech. Eng.* **2015**, *20*, 981–991.
28. Erzin, Y.; Turkoz, D. Use of neural networks for the prediction of the CBR value of some Aegean sands. *Neural Comput. Appl.* **2016**, *27*, 1415–1426. [[CrossRef](#)]
29. Ghorbani, A.; Hasanzadehshooili, H. Prediction of UCS and CBR of microsilica-lime stabilized sulfate silty sand using ANN and EPR models; application to the deep soil mixing. *Soils Found.* **2018**, *58*, 34–49. [[CrossRef](#)]
30. Suthar, M.; Aggarwal, P. Predicting CBR Value of Stabilized Pond Ash with Lime and Lime Sludge Using ANN and MR Models. *Int. J. Geosynth. Ground Eng.* **2018**, *4*, 6. [[CrossRef](#)]
31. Farias, I.G.; Araujo, W.; Ruiz, G. Prediction of California bearing ratio from index properties of soils using parametric and non-parametric models. *Geotech. Geol. Eng.* **2018**, *36*, 3485–3498. [[CrossRef](#)]
32. Kurnaz, T.F.; Kaya, Y. Prediction of the California bearing ratio (CBR) of compacted soils by using GMDH-type neural network. *Eur. Phys. J. Plus* **2019**, *134*, 326. [[CrossRef](#)]
33. Alam, S.K.; Mondal, A.; Shiuly, A. Prediction of CBR Value of Fine Grained Soils of Bengal Basin by Genetic Expression Programming, Artificial Neural Network and Krigging Method. *J. Geol. Soc. India* **2020**, *95*, 190–196. [[CrossRef](#)]
34. Islam, M.R.; Roy, A.C. Prediction of California bearing ratio of fine-grained soil stabilized with admixtures using soft computing systems. *J. Civil Eng. Sci. Technol.* **2020**, *11*, 28–44. [[CrossRef](#)]
35. Taha, S.; Gabr, A.; El-Badawy, S. Regression and Neural Network Models for California Bearing Ratio Prediction of Typical Granular Materials in Egypt. *Arab. J. Sci. Eng.* **2019**, *44*, 8691–8705. [[CrossRef](#)]

36. Hao, S.; Pabst, T. Prediction of CBR and resilient modulus of crushed waste rocks using machine learning models. *Acta Geotech.* **2022**, *17*, 1383–1402. [[CrossRef](#)]
37. Khatti, J.; Grover, K.S. Prediction of soaked CBR of fine-grained soils using soft computing techniques. *Multiscale Multidiscip. Model. Exp. Des.* **2023**, *6*, 97–121. [[CrossRef](#)]
38. Kamrul Alam, S.; Shiuly, A. Soft Computing-Based Prediction of CBR Values. *Indian Geotech. J.* **2023**, 1–15. [[CrossRef](#)]
39. Bolón-Canedo, V.; Remeseiro, B. Feature selection in image analysis: A survey. *Artif. Intell. Rev.* **2020**, *53*, 2905–2931. [[CrossRef](#)]
40. Kabir, H.; Garg, N. Machine learning enabled orthogonal camera goniometry for accurate and robust contact angle measurements. *Sci. Rep.* **2023**, *13*, 1497. [[CrossRef](#)] [[PubMed](#)]
41. UNE UNE-EN ISO 17892-4:2019; Investigación y Ensayos Geotécnicos. Ensayos de Laboratorio de Suelos. Parte 4: Determinación de la Distribución Granulométrica. Asociación Española de Normalización y Certificación: Madrid, Spain, 2019.
42. UNE, UNE-EN ISO 17892-12:2019; Investigación y Ensayos Geotécnicos. Ensayos de Laboratorio de Suelos. Parte 12: Determinación del Límite Líquido y del Límite Plástico. (ISO 17892-12:2018). Asociación Española de Normalización y Certificación: Madrid, Spain, 2019.
43. UNE, UNE 103501:1994; Ensayo de Compactación. Proctor Modificado. Asociación Española de Normalización y Certificación: Madrid, Spain, 1994.
44. UNE, UNE 103502:1995; Método de Ensayo para Determinar en Laboratorio el Índice C.B.R. de un Suelo. Asociación Española de Normalización y Certificación: Madrid, Spain, 1995.
45. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE: New York, NY, USA, 2008.
46. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 20–25 August 1995; Volume 2, pp. 1137–1143.
47. Duan, T.; Anand, A.; Ding, D.Y.; Thai, K.K.; Basu, S.; Ng, A.; Schuler, A. Ngboost: Natural gradient boosting for probabilistic prediction. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020.
48. Meinshausen, N.; Ridgeway, G. Quantile regression forests. *J. Mach. Learn. Res.* **2006**, *7*, 983–999.
49. Koenker, R.; Hallock, K.F. Quantile regression. *J. Econ. Perspect.* **2001**, *15*, 143–156. [[CrossRef](#)]
50. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
51. Rasmussen, C.E. *Gaussian Processes in Machine Learning, in Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 63–71.
52. Karaboga, D. *An Idea Based on Honey Bee Swarm for Numerical Optimization*; Technical Report-tr06; Erciyes University, Engineering Faculty, Computer Engineering Department: Kayseri, Turkey, 2005.
53. Asteris, P.G.; Nikoo, M. Artificial bee colony-based neural network for the prediction of the fundamental period of infilled frame structures. *Neural Comput. Appl.* **2019**, *31*, 4837–4847. [[CrossRef](#)]
54. Asteris, P.G.; Koopialipoor, M.; Armaghani, D.J.; Kotsonis, E.A.; Lourenço, P.B. Prediction of cement-based mortars compressive strength using machine learning techniques. *Neural Comput. Appl.* **2021**, *33*, 13089–13121. [[CrossRef](#)]
55. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
56. Hunt, R.E. *Geotechnical Engineering Investigation Handbook*; CRC Press: Boca Raton, FL, USA, 2005.
57. Bowles, J.E. *Physical and Geotechnical Properties of Soils*; McGraw-Hill: New York, NY, USA, 1979.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.