

A Hybrid GPR Framework with Feature Selection for Enhanced Prediction and Explainability of Subgrade Soil Strength

Billel Bouguedra

billel.bouguedra@univ-djelfa.dz

Ziane Achour University of Djelfa

Khaled Sandjak

M'hamed Bouguara University of Boumerdes

Mouloud Ouanani

Ziane Achour University of Djelfa

Ahmed Hassan Backar

Umm Al-Qura University

Hegazy Rezk

Prince Sattam bin Abdulaziz University

Article

Keywords: Subgrade soil, Strength, Hybrid GPR, Feature selection, Prediction

Posted Date: March 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8970563/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on April 18th, 2026. See the published version at <https://doi.org/10.1038/s41598-026-48770-2>.

A Hybrid GPR Framework with Feature Selection for Enhanced Prediction and Explainability of Subgrade Soil Strength

Billel Bouguedra ^{1,*}, Khaled Sandjak ², Mouloud Ouanani ¹, Ahmed Hassan Backar ³, Hegazy Rezk ^{4,5}

¹ Civil Engineering Department, Faculty of Sciences and Technology, Development in Mechanics and Materials Laboratory DMML, Ziane Achour University of Djelfa, Algeria.

² Civil Engineering Department, Faculty of Technology, M'hamed Bouguara University of Boumerdes, Algeria.

³ Mechanical Engineering Department, College of Engineering and Architecture, Umm Al-Qura University, P.O Box 5555, Makkah 21955, Saudi Arabia.

⁴ Department of Electrical Engineering, College of Engineering in Wadi Alldawasir, Prince Sattam bin Abdulaziz University, Saudi Arabia

⁵ Electrical Engineering Department, Faculty of Engineering, Minia University, Egypt

* Correspondence: billel.bouguedra@univ-djelfa.dz

Abstract: In order to overcome the drawbacks of the labour-intensive California Bearing Ratio (CBR) test for pavement design, a Hybrid Gaussian Process Regression (GPR) model for predicting subgrade soil strength was developed. The model is optimised using the Bayesian Tree Parzen Estimator and uses Recursive Feature Elimination for feature selection. Several machine learning techniques, such as Multi-Linear Regression (MLR), Multi-Layer Perceptron of Artificial Neural Networks (MLPNN), and Random Forest (RF), were used in the first stage of evaluation on a regional geotechnical database with over 1000 samples. In the second stage, performant hybrid models based on Dense Neural Network and tree-based ensembles models were used. All of the machine learning models utilised in this study were outperformed by the Hybrid GPR model, according to the trial results. With the training and testing subsets' MSE, MAE, MAPE, and R2 ranging from 0.0768 to 0.2974, 0.2107 to 0.4364, 5.32% to 13.79%, and 0.9982 to 0.9935, respectively, it attained the highest accuracy. The SHAP interpretability method is used to analyse the prediction outcomes in order to further investigate the "black box" problem of the prediction model. According to the SHAP study, dry density, gravel content, and optimum moisture content are the main factors influencing subgrade soil strength based on CBR values. The results of the study show that the Hybrid Gaussian Process Regression model effectively captures the intricate nonlinear interactions between the feature parameters and offers a crucial foundation for CBR value estimation.

Keywords: Subgrade soil; Strength; Hybrid GPR; Feature selection; Prediction.

1. Introduction

Evaluating the strength of the subgrade soil is essential for the flexible pavements design. The California Bearing Ratio (CBR) is commonly used to assess the strength or stiffness modulus of subgrade soils. It takes a lot of time, effort, and labor to measure the CBR, and it involves complex testing procedures [1-3]. A variety of efforts have been put in place to establish empirical links between the California Bearing Ratio and the important soil characteristics in light of the difficulties encountered. Multiple studies demonstrate that essential soil parameters affect CBR, with proven correlations between the California Bearing Ratio and different soil characteristics, including grain size distribution, water content (WC), maximum dry density (MDD), and plasticity [4,5]. Many prediction approaches that only use linear and non-linear regression are insufficient to provide accurate CBR forecasts and flexible applications, therefore soft computing techniques are becoming more and more important in a variety of engineering domains [1,6]. Case-Based Intelligence (CBR) prediction models have been developed in the last few decades using a variety of Artificial Intelligence (AI) methodologies.

The R² value showed that the models' accuracy ranged from 0.65 to 0.80. Inadequate training of machine learning models, overfitting, a decreased correlation between the parameters (input and output), and poor input parameters are some of the possible causes of the accuracy decline.

These consist of gradient boosting machines [2], artificial neural networks (ANN) [1,5,7,8], gene expression programming (GEP) [1,9, 10], support vector regression (SVR) [10,11], group method of data handling (GMDH) [12] and more. Typically, these models include input factors such as compaction characteristics, soil index qualities, and particle size distribution. Several models with respectable determination coefficient (R²) values have been proposed using conventional machine learning (CML) models [7,9,13-15]. However, the credibility of these research is limited because most of these models were tested using a limited collection of datasets [16].

A recent research investigation by Bardhan et al. employed 312 experimental findings to assess the soaked (CBR) of fine-grained soils [3]. Using six soil factors, the authors achieved a substantial prediction accuracy of 80% to 92% during the testing phase, as indicated by the R² value. The parameters examined comprised (gravel, sand, silt and clay) content, plasticity index, maximum dry density, and optimum moisture content. According to a study by Al-Busultan et al., the accuracy of the conventional models declined with dataset size [17].

GPR has recently been recognized as a workable solution for a number of geotechnical issues because of its exceptional forecast accuracy and resilience. Using a large database from the northeastern part of Algeria, this study uses a

novel GPR approach to develop predictive models for the CBR, using a number of soil index properties as modeling inputs. GPR is a probabilistic, non-parametric approach used to solve regression problems. Even with tiny datasets, the efficacy of this supervised machine learning technique is demonstrated. In order to tackle complex engineering problems, GPR has been widely used in many domains [18-23]. In geotechnical engineering, GPR has been frequently employed to develop prediction models. These models forecast soil hydraulic properties [24], assess rock slope stability [8], determine pile-bearing capacity [25], and evaluate the triaxial strength of jointed rock mass [26]. Further applications include analyzing tunnel geotechnical parameters [27], predicting ground vibrations from blasting [28], estimating tunnel surface settlements [29], and investigating slope failures and soil compaction [30,31]. This diverse use of GPR fosters confidence in its application for estimating the CBR.

Feature selection methods typically focus on minimizing a model's loss function by making selections of tailored subset of features. Recursive Feature Elimination (RFE) is a feature selection technique that systematically remove the least effective features until only the best subset remains. It works by training a model, ranking features by their significance, pruning the weakest ones, and repeating the process until a predefined number of features remains [32-35]. Complementing this, tools like SHAP (SHapley Additive exPlanations) and TPE (Tree-structured Parzen Estimator) offer advanced analytical capabilities. SHAP leverages game theory to quantify each feature's marginal contribution to model predictions, providing both global and local interpretability through precise importance scores [36]. Due to its versatility, SHAP has found widespread use across diverse domains [37-39].

Hyperparameter optimization, another critical challenge in machine learning, involves tuning model configurations to maximize performance. Manually setting hyperparameters is complex and can drastically influence model efficacy. The TPE algorithm, addresses this by combining Bayesian optimization with surrogate modeling to efficiently identify optimal hyperparameters, thereby boosting model accuracy [40-44].

The incorporation of the RFE method expedites and enhances feature selection, while the TPE framework automates hyperparameter optimization, alleviating problems like overfitting and inadequate generalization. Collectively, they facilitate faster and economical model creation with enhanced predictive capability [45-47]. Additionally, by demonstrating how the input variables impact the model's output, the incorporation of SHAP enhances interpretability.

This thorough method facilitates deployment and analysis in real-world applications while improving model correctness. This study presents three major contributions that advance the technical state of the art. First, it employs the TPE technique for hyperparameter optimization and improving the model architecture to outperform conventional ML techniques in terms of prediction accuracy. Second, it makes the model more efficient by using Recursive Feature Elimination (RFE) to find the most effective variables. Lastly, it combined Hybrid Gaussian Process Regression with SHapley Additive explanations (SHAP). This approach increases openness and provides useful data regarding the ways in which soil properties impact CBR outcomes. The study concludes by demonstrating the effectiveness of hybrid machine learning algorithms in estimating soil performance and motivating sustainable infrastructure system design methodologies when combined with SHAP analysis.

2. Research Methodology

2.1. Recursive Feature Elimination

One feature screening technique that comes from the machine learning domain is RFE [48]. It operates as a methodical approach that prioritizes features through a sorting mechanism, embodying the principles of a greedy algorithm. Starting with the initial feature set, the basic idea is to methodically remove the least important features according to the machine learning model's determination of feature relevance. Following numerous iterations, a number of

feature subsets are produced, and the classifier's prediction accuracy is used to determine which subset is most useful. RFE is based on the classifier's ability to evaluate feature importance, as shown by technique like SVM and RF.

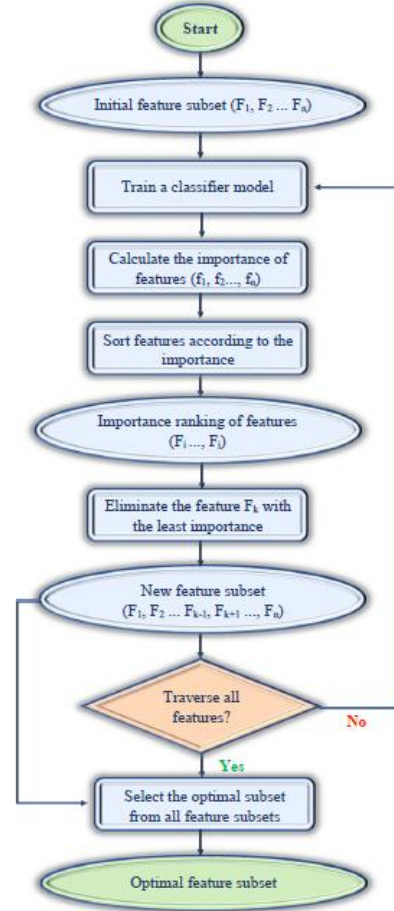


Fig.1. An RFE algorithm flowchart.

Fig.1 illustrates the RFE method flowchart, which consists of five primary steps. Based on the relevance rating, the least relevant feature is eliminated, creating a new feature subset $\{F_1, F_2, F_3, \dots, F_{n-1}, F_n\}$ that has $n - 1$ features. Steps 1 through 3 are repeated using the feature subset that was obtained in Step 3 as a new feature set. Every iteration creates a new feature subset, resulting in n feature subsets overall at the end. The accuracy of the classifier is used to choose the best subset [49-51].

2.2. Gaussian Process Regression

Because it can efficiently analyze small samples and limited-dimensional regression problems, GPR is a kernel-based machine learning technique that is widely used in the study of solar energy prediction and lithium-ion battery research [52,53]. GPR is essentially a non-parametric model that uses Gaussian process priors to do regression analysis on data. [54]. Other regression techniques necessitate precise modelling parameters, whereas GPR trains on sample data using probabilistic techniques. Additionally, the mean and covariance functions determine GPR, and hypotheses for posterior probability are obtained by Bayesian inference [55]. When it comes to handling complex and nonlinear situations, GPR is more applicable [54]. The following functional formula is frequently used to identify a Gaussian process:

$$f(x) \sim GP[m(x), k(x, x')] \quad (1)$$

With: $m(x) = E[f(x)]$ and $k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]$
 $m(x)$ and $k(x, x')$ are the mean function and covariance function, respectively,
 whereas $x, x' \in \mathbb{R}^n$ are random variables.

To make calculations easier, $m(x) = 0$ is typically used [53]. Given the presence of noise in the observed target value y , the general model for determining GPR is:

$$y = f(x) + \varepsilon \quad (2)$$

With: $\varepsilon \sim N(0, \sigma_n^2)$, where ε is noise. Therefore, the prior distribution of the observed value y is:

$$y \sim N(0, k(x, x') + \sigma_n^2 I_n) \quad (3)$$

In this case, I_n is an identity matrix with n dimensions. The joint prior distribution of the observed value X and the predicted value y^* , assuming that the training dataset X and the testing dataset X^* have the same Gaussian distribution, is:

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim N \left(0, \begin{bmatrix} K(X, X) + \sigma_n^2 I_n & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \quad (4)$$

This is the covariance matrix between the training dataset X and the testing dataset X^* , where $K(X, X)$ is the covariance matrix of the training dataset and $K(X^*, X^*)$ is the covariance matrix of the testing dataset. The posterior distribution of the expected value y^* can therefore be computed as follows:

$$y^* | X, y, X^* \sim N(\bar{y}^*, \text{cov}(y^*)) \quad (5)$$

$$\bar{y}^* = K(X^*, X)[K(X, X) + \sigma_n^2 I_n]^{-1} Y$$

Using the formula:

$$\text{cov}(y^*) = K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma_n^2 I_n]^{-1} K(X, X^*)$$

The mean and covariance of the predicted value y^* on the testing dataset X^* are denoted by $\bar{y}^*$ and $\text{cov}(y^*)$, respectively.

One of the main elements influencing model performance is the selection of the covariance function, often known as the kernel function. The covariance function characterizes the correlation between samples as part of the model assumptions [55]. The covariance function, which comprises squared exponential, exponential, Matern, periodic, and rational quadratic covariance, is the foundation of the Gaussian process. [56,57].

Squared exponential covariance function:

$$k(x, x') = \exp \left(-\frac{|d|^2}{2l^2} \right) \quad (6)$$

Exponential covariance function:

$$k(x, x') = \exp \left(-\frac{|d|}{l} \right) \quad (7)$$

Matern covariance function:

$$k(x, x') = \frac{2^{1-v}}{\Gamma(v)} \left(\frac{\sqrt{2v}|d|}{l} \right)^v K_v \left(\frac{\sqrt{2v}|d|}{l} \right) \quad (8)$$

Periodic covariance function:

$$k(x, x') = \exp \left(-\frac{2 \sin^2 \left(\frac{d}{l} \right)}{l^2} \right) \quad (9)$$

Rational quadratic covariance function:

$$k(x, x') = (1 + |d|^2)^{-\alpha} \alpha \geq 0 \quad (10)$$

Where: $d = x - x'$ represents the Euclidean distance between x and x' , α is the shape parameter, l is the length scale and v is the smoothness parameter for the Matern kernel (Generally take values of 1/2, 3/2 or 5/2). In noise modelling, a white kernel is incorporated into the base kernel of the Gaussian Process model. The mathematical representation of the white kernel is as follows:

$$k_{\text{white}}(x_i, x_j) = \sigma^2 \cdot \delta_{ij} \quad (11)$$

With: σ^2 represents the noise variance and δ_{ij} represents the Kronecker delta function.

2.3. Bayesian optimization using Tree-structured Parzen Estimator (TPE)

Because of its effectiveness in handling complex functions with few evaluations, Bayesian optimization has emerged as a key technique for hyperparameter tuning and model optimization in a variety of domains, especially when using the Tree-structured Parzen Estimator (TPE). Recent research demonstrates its use in a variety of fields, including ecological monitoring and machine learning. For example, an article by Lim et al. describes a Bayesian optimization technique that uses Extremely Randomized Trees (TPE) specifically to detect credit card fraud. [58]. This illustrates how TPE achieves notable gains in detection accuracy while efficiently automating hyperparameter adjustment to improve model performance [58]. The use of TPE in an IoT intrusion detection system is also discussed in another paper, which highlights how it helps optimize hyperparameters for the best detection performance even in the face of unbalanced datasets [59]. This trend is corroborated further in research conducted by Zhang et al., where employing Bayesian optimization based on TPE yielded optimal hyperparameter combinations quickly, thus increasing the accuracy of models predicting compressive strength in recycled aggregate concrete [60]. Moreover, the efficacy of TPE extends to more complex models, as illustrated by Aydoğmuş, who incorporate it into reinforcement learning algorithms to enhance bipedal robot locomotion. Their findings underscore TPE's capacity to effectively optimize multiple reinforcement learning algorithms [61]. Similarly, the study by Song et al. highlights TPE's application in soil ammonia spatial prediction, integrating it with machine learning and ordinary kriging to monitor ecological parameters effectively [62].

In a notable advancement, other researchers apply TPE to develop a self-configurable receiver for underwater acoustic communications, marking its potential in novel engineering applications [63]. A study of Westmeier et al. emphasize TPE's role in quantifying uncertainty in neural network predictions, reinforcing the importance of Bayesian methods in managing model predictions amidst complex noise profiles [64]. Cao et al. similarly echo this in their works, where TPE's integration with gradient boosting regression trees facilitates real-time estimation in earthwork excavation, displaying its practical implications in geotechnical engineering [65].

Furthermore, recent publications have consistently illustrated TPE's advantages over traditional hyperparameter tuning methods. For example, researchers in [66] greatly improved the performance of predictions in biomedical contexts by optimizing model parameters for brain epigenetic analysis using TPE. Accordingly, a comparison study by Tørring and Elster demonstrated TPE's robustness in auto-tuning applications by confirming that it provides better performance than its rivals in a variety of sample size scenarios [67]. All things considered, the use of TPE in Bayesian optimization in various domains shows how adaptable it is and emphasizes how Bayesian techniques are increasingly acknowledged for their ability to optimize complex models.

The primary objective of hyperparameter optimization is to maximize the prediction performance of models. This optimization process can be described as Equation (12), where all of the hyperparameters that need to be optimized are contained in the parameter space X . A set of ideal values X^* can be obtained by repeatedly adjusting the model's hyperparameters.

$$X^* = \arg \min f(x) \quad (12)$$

The goal function is denoted by $f(X)$, the parameter space is represented by X and X^* is a collection of hyperparameters from X that minimize the objective function $f(X)$. Bayesian optimization (BO) is an important hyperparameter optimization method that performs well for most problems. Unlike a random search, BO is a history-informed hyperparameter optimization method. Experiences have shown that BO may explore and exploit the unsampled and promising regions to improve the surrogate function between the hyperparameters and losses, lowering the possibility of leaving out the best portion [68].

Bayesian optimization can be achieved through the sequential updating of surrogate models, which allows for the dynamic use of historical data to determine the next sampling point. The goal of this method is to reduce the differences between the surrogate model and the function that actually links hyperparameters to loss. In Bayesian optimization, the three main algorithmic approaches used in sequential model-based optimization (SMBO) are the Tree-structured Parzen estimator (TPE), Random Forest (RF), and Gaussian Process (GP). The black box function that links hyperparameters to the loss function is what SMBO is meant to solve [69]. By using regression and classification approaches, Random Forest (RF) shows increased efficiency in solving simple optimization issues. Predetermined connections between the black box function and hyperparameters are established via Gaussian Processes (GP), whereas Tree-structured Parzen Estimator (TPE) does not require such variable relationships to be defined [70-72]. Therefore, TPE is more appropriate for various machine learning models.

The domain, surrogate model, acquisition function, true loss function, and historical data are some of the components that make up the model-based methodology known as SMBO. In order to clarify the connections between hyperparameters and the losses of the trained machine learning framework, it seeks to optimize the surrogate model to faithfully depict the actual loss function. Beginning with the acquisition function and the surrogate model S as $P(y|x)$. The best value of the present observed results in processes is set: $y^* = \min\{f(x_i), x_i \in X\}$. The Gaussian Process will calculate the predicted improvement as the acquisition function:

$$EI_{y^*}(x) = \int_{-\infty}^{+\infty} \max(y^* - y, 0) P(y|x) dy \quad (13)$$

With: the $EI_{y^*}(x)$ denotes the expected outcomes of the GP model-modeled presumed surrogate function $P(y|x)$. Based on the projected expected improvement EI , GP will decide which exploration or exploitation operation to pursue next. The region that has a bigger uncertainty or a better mean value will be chosen. In order to update the historical data for the next stage, SMBO will then assess this trial to determine the actual loss function value. TPE will use the Bayesian rule, similar to the GP, to calculate the probability of x given the observed condition y , $P(x|y)$, instead of $P(y|x)$, where x is the hyperparameter and y is the value of the actual loss function. According to Bayes' theorem, the relationship between $P(x|y)$ and $P(y|x)$ is as follows:

$$P(y|x) = \frac{P(x|y)p(y)}{p(x)} \quad (14)$$

In order to first classify the remaining samples into bad and good performance, TPE describes $P(x|y)$ as a piecewise function using two density functions, $l(x)$ and $g(x)$ (see equation (15)).

$$P(x|y) = \begin{cases} l(x), & \text{if } y < y^* \\ g(x), & \text{if } y \geq y^* \end{cases} \quad (15)$$

The quantile γ will be used to determine the threshold y^* , so that $p(y < y^*) = \gamma$. Equation (16) can therefore be used to determine the expected improvement EI of the TPE method:

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) P(y|x) dy = \frac{\gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} p(y) dy}{\gamma l(x) + (1-\gamma)g(x)} \alpha \left(\frac{1}{\gamma + \frac{(1-\gamma)g(x)}{l(x)}} \right) \quad (16)$$

According to this final formula, we desire points x that show low probability under $g(x)$ and high probability under $l(x)$ in order to maximize improvement. Based on l , the hierarchical structure of l and g makes it easier to generate a large number of choices and evaluate them using $g(x)/l(x)$. Every time the algorithm is run, the candidate x^* with the highest expected.

2.4. Shapley Additive exPlanations for model interpretability

The SHAP approach, created by Lundberg et al. for tree models, was used in the current study since interpretability is important in machine learning model [73]. Shapley values from cooperative game theory are used in the SHAP computation. Shapley values quantify each characteristic's contribution to the prediction's result. Shapley gave a formula for the Shapley value and described it as the average marginal contribution of a feature f across all potential feature subsets. In practice, the Shapley value is usually estimated, whereas SHAP is an optimization algorithm designed to estimate the Shapley value. Using cumulative feature attribution, the SHAP approach presents an explanatory model based on Shapley values as follows:

$$g(z') = \psi_0 + \sum_{j=1}^M \psi_j z'_j \quad (17)$$

When z' is a binary vector of length M in M -dimensional space, with 1 denoting the presence of a feature and 0 denoting its absence, the function $g(z')$ is the interpretation function of z' . The number of features in the model is indicated by M . The Shapley value of feature j is shown by the constant term ψ_j , whereas the mean prediction value across all samples is indicated by the constant term ψ_0 .

$$\psi_j = \sum_{S \subseteq \{x_1, x_2, \dots, x_p\}} \frac{|S|!(p-|S|-1)!}{p!} (f_x(S \cup \{x_j\}) - f_x(S)) \quad (18)$$

Let $\{x_1, x_2, \dots, x_p\}$ denote the complete collection of input features, where p signifies the total number of input features. The notation $\{x_1, x_2, \dots, x_p\} \setminus \{x_j\}$ indicates all possible subsets of input features excluding the feature $\{x_j\}$. Additionally, the predicted value for the feature subset S is indicated by $f_x(S)$. Over the past few decades, the SHAP approach has become well-known as a leading attribution technique in the field of interpretable machine learning. It has the advantage of being model-agnostic [74] when compared to other approaches, so theoretically it should be able to shed light on any ML model. This increases SHAP's scalability and versatility, which enhances its ability to comprehend machine learning algorithms from a variety of perspectives [39, 75, 76].

2.5. Model Evaluation Criteria

The evaluation metrics for predictive performance in this study were R-squared (R^2), mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). A risk metric called the Mean Squared Error (MSE) calculates the average of the squares of the mistakes. It is always non-negative and represents an estimator's quality with values nearer zero indicate higher quality. The true magnitude of prediction errors is indicated by the MAE, which stands for the average absolute difference between expected and actual

values. A statistical indicator of forecasting accuracy, the Mean Absolute Percentage Error (MAPE) represents the average magnitude of error produced by a predictive model as a percentage of the actual observed values. Lower MAPE values are preferable, as they indicate higher predictive accuracy. The sum of squares of the differences between actual values and their mean is known as the total sum of squares, whereas the residual sum of squares is the sum of squares of the disparities between predicted values and actual values. The R^2 number, which is calculated as 1 minus the ratio of the residual sum of squares to the total sum of squares, represents the model's fit quality. A better model fit is indicated by an R^2 value that is close to 1. The following are the exact formulations for these metrics:

$$MSE = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i)^2 \quad (19)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |X_i - Y_i| \quad (20)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{X_i - Y_i}{Y_i} \right| \times 100\% \quad (21)$$

$$R^2 = \left(\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 \right]}} \right)^2 \quad (22)$$

Where: N is the number of data recordings, X_i and Y_i are the predicted and measured values, and $\bar{X}$ and $\bar{Y}$ are the means of the predicted and measured values of CBR, respectively.

3. Data description and statistical analysis

The ratio CBR of subgrade soil is influenced by various factors. This study will concentrate on the key factors that notably influence the CBR in order to simplify the model. An experimental database is employed to develop the Gaussian hybrid model for the indirect estimation of the soaked CBR of subgrade soils, relying on fundamental engineering and compaction characteristics.

The geotechnical database includes more than 1,000 samples of subgrade soil, which are typically formed from worn sedimentary rocks (marls, shales, and limestone). Alluvial clays and silty clays, which are occasionally soft and compressible, are found in low-lying locations and near river plains. Since 2012, the Northeastern region of Algeria has been the main source of these subgrade samples. The subgrade dataset's classification based on the USCS shows that the soil mass is primarily fine-grained. About 75–80% of all specimens are classified as clays or clayey silts with high plasticity (CH/MH–CH), 8–12% as clays or silts with low to medium plasticity (CL/ML), 5–8% as sandy soils with heavy fines (SM–SC), and only roughly 3–5% as gravelly soils with fines (GM–GC). There are very few clean coarse-grained soils (GW, GP, SW, SP), indicating that the material under study is primarily a fine cohesive subgrade with little gravel and sand.

The CBR tests are conducted, according to the EN 13286-47:2021 standard, by the laboratory of the Central Transport Agency located in Algiers, Algeria.

From this experimental database, various input parameters are identified, specifically, the (Gravel (G), Sand (S) and Fines (FC)) content, Coefficient of (Uniformity (CU) and Curvature (CC)), Liquidity Limit (LL), Plasticity Limit (PL), Plasticity Index (PI), Sample Moisture Content (MC), Sample Dry Density (DD), Optimum Moisture Content (OMC), and Maximum Dry Density (MDD). The result of these parameters in modelling is the Ratio (CBR) of the subgrade soil.

Table 1

The used dataset's summary statistics.

Variable	Mean	SD	Min	Q1	Q3	Max
G(%)	4.1	4.65	0.0	1.6	7.9	16.6
S(%)	21.1	10.50	0.1	12.7	28.2	46.9
FC(%)	74.2	12.10	50.0	65.3	80.6	89.9
CU	17.60	14.50	0.0	7.4	22.3	102.5
CC	0.66	0.77	0.0	0.1	1.0	2.8
LL(%)	38.80	12.00	21.7	29.7	44.5	94.4
PL (%)	19.4	4.9	10.7	17.0	22.1	41.0
PI(%)	19.1	7.9	6.6	14.6	21.9	53.3
MC(%)	11.3	3.0	3.1	9.6	16.6	18.6
DD(g/cm ³)	1.83	0.13	1.4	1.7	1.9	2.2
OMC (%)	11.3	3.0	3.8	9.5	12.8	18.8
MDD(g/cm ³)	1.9	0.12	1.5	1.84	1.98	2.1
CBR(%)	7.40	6.44	0.2	2.8	9.7	26.9

From The dataset's statistical analysis in Table 1 provides crucial information on the geotechnical characteristics of the soil samples taken from the area under study. Important statistical measures such as mean, standard deviation (SD), minimum, quartiles, and maximum values are highlighted in the dataset. For geotechnical engineering applications, a comprehensive understanding of the variability of subgrade soils is made easier by these data. The following is the key finding from the statistical examination of the unique properties of the soil:

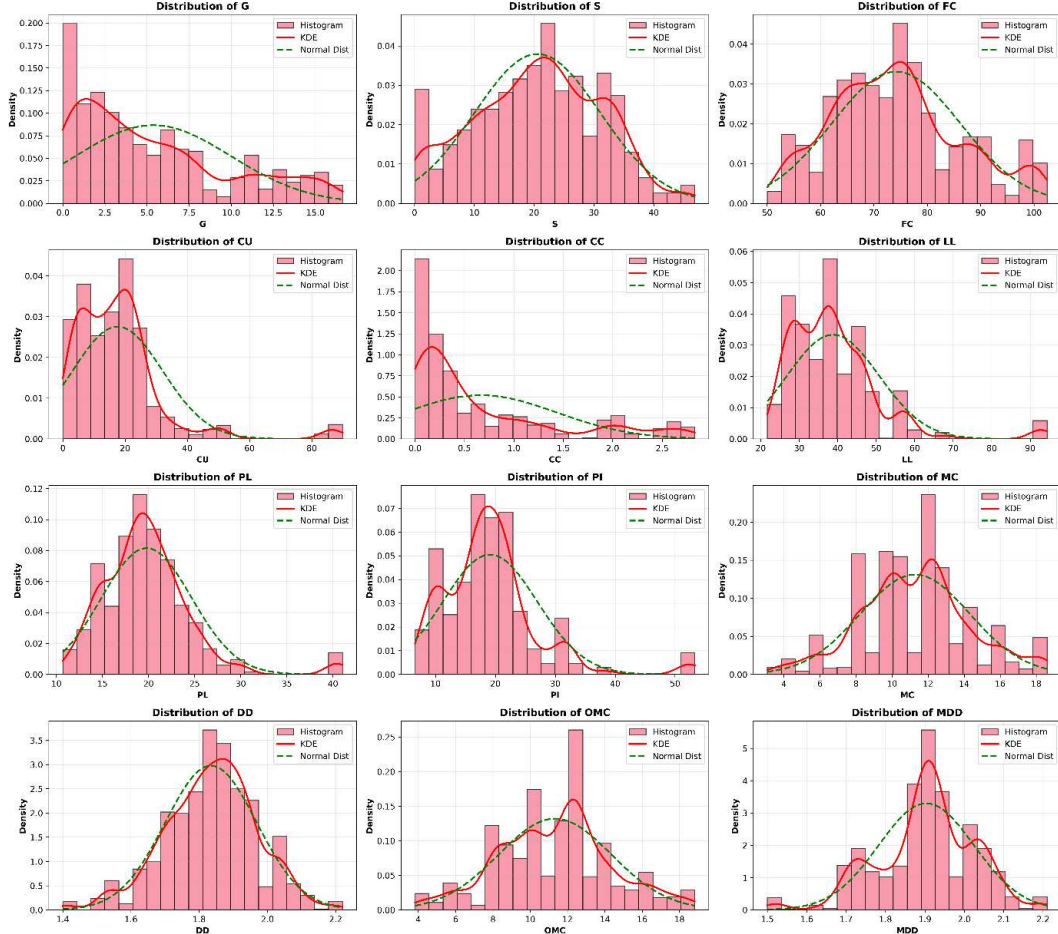
- Sand (%) and Gravel (%) Table 1's contents indicate that sandy-sized soil fractions are prevalent in the dataset, with average values of 4.1% and 21.1%, respectively. The soil's capacity to sustain loads is directly impacted by the amount of coarse-grained fractions, such as sand and gravel. The findings presented here are consistent with what is known about how soil behaves under structural loads [77].
- The soil has moderate cohesiveness and flexibility, as shown by the Fines Content (FC%), which averages 74.2%. This trait has a big impact on how adaptable and effective the soil is in various moisture conditions. The diversity of soil types along the examined segment is reflected in the change in the detected particle content [78].
- The morphology of the particle-size distribution curve is described by the coefficients of uniformity (CU) and curvature (CC), which have respective means of 17.6 and 0.66. This indicates that the soils in the dataset are either uniformly graded or gap-graded, which means that some particle sizes may be absent. This frequently results in increased compressibility and decreased strength.
- The average value of the Plasticity Index (PI), which is the difference between the soil's plastic limit (PL) and liquid limit (LL), is 19.1%. This measure is essential for determining how malleable and capable the soil is of holding onto moisture. The significance of PI values in determining soil suitability for geotechnical applications is shown by Table 1, which shows that the range of PI values is closely related to soil mineralogy and water retention characteristics [78,79].
- With an average of 1.9 g/cm³, the Maximum Dry Density (MDD) provides important information on the soil's proctor compaction properties, which are crucial for figuring out how much weight it can support. The variations in MDD values show how the soil samples' compaction potential varies, which impacts the soil's overall stability. Similarly, the moisture level required to achieve optimal compaction efforts is indicated by the optimum moisture content (OMC), which averages 11.3%. When predicting how the soil would respond to structural loads, both factors are essential [80].

- Through penetration resistance, the CBR which has an average of 7.4% and a range of 0.2% to 26.6% acts as an indirect indicator of subgrade strength. When designing pavements, CBR is a crucial metric for assessing the soil's ability to sustain structural loads. The test data set's variation in CBR values emphasizes the necessity of local soil assessments to guarantee pavement design that is both safe and effective [81].

which could help in future pavement design requirements in the surrounding area. The Histograms, shown in Fig. 2 give more insight in data distribution, while some of soil properties are approximately symmetric (S, FC, MC, DD, OMC, MDD) the remaining properties are right skewed.

A comprehensive view of the linear correlations between the soil engineering characteristics derived from the test dataset is provided by the correlation matrix, which is shown in Fig. 3. The degree and direction of these interactions are measured by the correlation coefficients, which range from -1 to 1. This information is crucial for understanding how elements interact in geotechnical research [82].

a)



b)

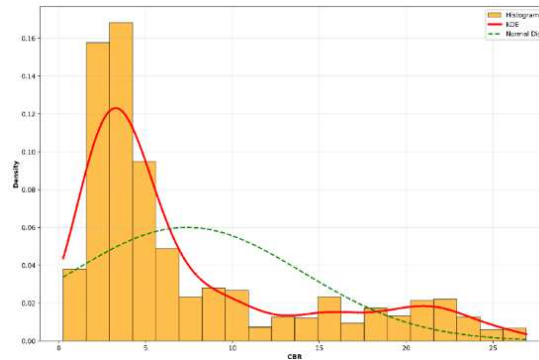


Fig. 2. Histograms with probability distributions a) Input variables; b) Output variable (CBR).

A thorough overview of the geotechnical parameters is provided in Table 1, which also shows the statistical distribution of the input and target variables. The observed variability highlights the importance of site-specific soil characterization in civil engineering applications. The statistics offered enable engineering professionals to evaluate the dataset's validity and accuracy for ML applications in forecasting soil strength, particularly in terms of CBR values,

Significant positive correlations are apparent among specific parameters, illuminating their interrelated behavior. The association between DD, MDD, and CBR ($r = 0.42$ and 0.37), indicates that greater density is associated with elevated CBR values, aligning with previous research [1,83].



Fig. 3. Correlation matrix derived from soil properties.

Furthermore, the correlation between CBR and the coarse aggregate characteristics (G% and S%) emphasizes how important coarse particles are for raising the CBR value. This result is in line with research that highlights the contribution of coarse particles to improving soils' ability to support loads [84]. Similarly, the association between PI and FC ($r = 0.33$) highlights the impact of particle composition on moisture response by showing that higher fine content is linked to increased soil flexibility [85]. Given that finer soils typically exhibit different mechanical properties than coarser soils, the results are crucial for predicting how soils will behave in response to varying moisture levels. Although not as strong, moderate correlations provide additional insight on geotechnical behavior. The correlation between MDD and PI ($r = -0.31$) suggests that more malleable soils typically exhibit lower maximum densities, which is crucial for improving compaction methods. [86].

Although it is less significant than other factors, the association between CBR and OMC ($r = -0.21$) indicates that moisture conditions affect soil strength [87]. As a diagnostic tool, the heatmap chart makes it possible to identify significant connections between crucial geotechnical engineering data [82]. In geotechnical engineering, understanding these linkages is crucial for forecasting soil behavior, enhancing material selection, and facilitating well-informed design choices

4. Hybrid Model development

One hybrid technique that uses Gaussian process regression as the basis learner for recursive feature reduction is the Bayesian Tree-structured Parzen Estimator optimized Gaussian process regression mixed with recursive feature elimination (TPE-GPR-RFE). Based on the relevance measurements generated by the Gaussian process regression model, it ranks and chooses features using the sequential backward selection method.

The GPR model has many hyperparameters, and changing these combinations might cause the model to identify different optimal subsets throughout the feature selection process. Using inappropriate hyperparameters may result in a subset that is not optimal. Gaussian Process Regression is used as the base learner, and Bayesian optimization using the Tree-structured Parzen Estimator is used to improve the recursive elimination model.

After dimensionality reduction, this method seeks to determine the ideal subsets and hyperparameters that reduce cross-validation error. To create a surrogate model framework for optimization rounds, different model parameter combinations function as independent variables, and the objective function is the precision of the ideal feature subset obtained using GPR-RFE with cross-validation. The GPR-RFE model's parameter optimization process is shown in Fig.4.

The steps in the TPE-GPR-RFE process are as follows:

- (1) Set the range of hyperparameters for random sampling and initialize the GPR model parameters. The TPE algorithm uses the training set and initialization parameters as input. The objective function to be reduced is the GPR-RFE cross-validation results for every set of parameters. In order to describe promising and non-promising hyperparameters, the TPE algorithm iteratively creates two probability distributions. This is then used to inform the choice of further parameters for assessment.
- (2) The TPE algorithm uses its probabilistic model to choose new hyperparameter combinations to assess. Based on the ratio of the likelihood that a candidate point is a member of the promising group of observations $l(x)$ to the non-promising group $g(x)$, it calculates the Expected Improvement EI for that point. Since it is most likely to improve the most well-known outcome, the candidate point that maximizes this EI ratio is selected for study. This procedure effectively directs the search for the ideal set of parameters.
- (3) The GPR-RFE model is trained using the chosen hyperparameter combination x , and its performance is assessed to determine the objective function value $f(x)$ (e.g., prediction error). This new observation $(x, f(x))$ is added to the historical dataset. Instead of updating a surrogate model after every step, the TPE algorithm waits until a predefined number of new results (e.g., 10 or 20) are collected. Once this threshold is met, the entire history is used to refit the density distributions $l(x)$ and $g(x)$ for the next round of candidate selection.
- (4) Upon reaching the maximum iteration limit, the model update ceases, and the optimal sampling point together with the corresponding optimal subset is produced.

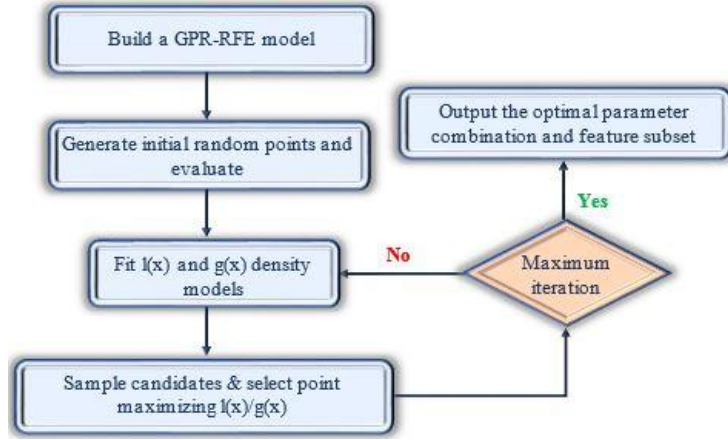


Fig. 4. TPE-GPR-RFE Hybrid algorithm flowchart.

The machine learning algorithms in this study are trained using the training dataset, which makes up 80% of the total database. During the training phase, the model discovers links and patterns in the data to produce accurate predictions. 20% of the dataset is used for testing, which assesses the models created during training. Using this independent dataset, researchers may measure the models' performance in actual events and evaluate how well they generalize to new data. Additionally, it is crucial to employ specific datasets for both training and testing in order to ensure a fair and consistent comparison between different models throughout the modeling process. Furthermore, a subset of training data is segregated as validation data using k-Fold cross-validation ($k=10$) to reduce the model overtraining.

Three other popular techniques are also included in order to objectively assess the Hybrid TPE-GPR-RFE model's performance: Multi-Linear Regression (MLR), Multi-layer Perceptron Neural Network (MLPNN) and Random Forest (RF) as a comparison baseline. Using the initial experimental data, prediction models for CBR were created for comparison and assessment. The technique is depicted in Fig.5.

Random Forest (RF) as a comparison baseline. Using the initial experimental data, prediction models for CBR were created for comparison and assessment. The technique is depicted in Fig.5.

5. Experimental results and discussion

5.1. Prediction performance of the Hybrid Model

The hybrid GPR-RFE uses feature selection algorithm and cross-validation to calculate the chosen feature set at each RFE step in tenfold cross-validation using the mean squared error (MSE) as the evaluation criterion to determine the ideal number of features with the lowest MSE. In parallel, Bayesian optimization with TPE is used to fine-tune the hyperparameters of GPR-RFE in an iterative process over 100 iterations in order to find the ideal feature subset that minimizes the mean squared error (MSE) and maximizes the R^2 value, especially in the testing set (unseen data). The process is depicted in Fig.6.

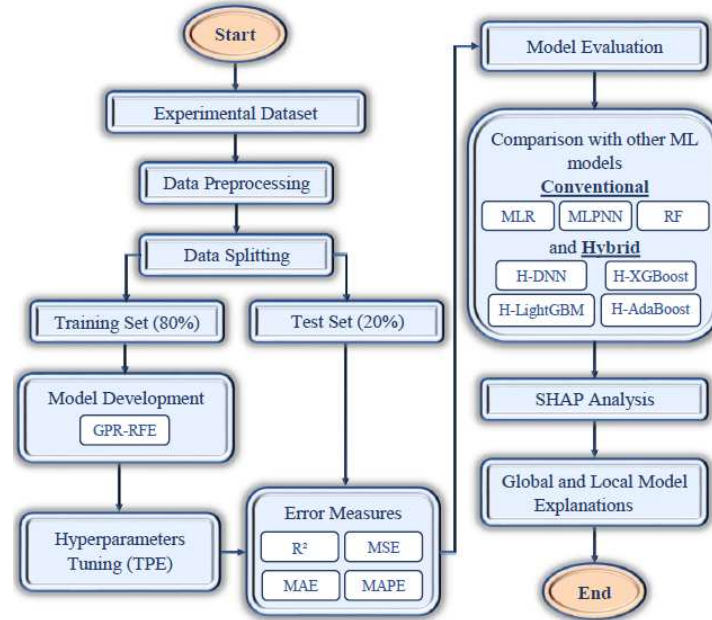


Fig. 5. Flowchart illustrating the current study's methodology.

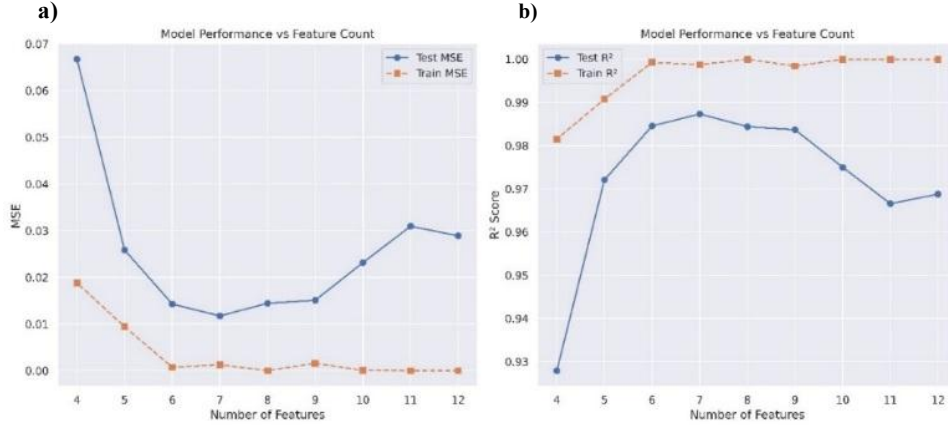


Fig. 6. Optimal feature subset search results a) MSE results; b) R² results.

Fig.6 demonstrates that both test and training performance, measured by MSE and R², dramatically enhance when the number of features increases from 4 to 7. Performance reaches its zenith at around seven features, where the test MSE is minimized and the test R² is maximized. Beyond this juncture, the incorporation of additional features results in overfitting, as training performance (MSE/R²) shows marginal improvement, while test performance deteriorates (test MSE escalates, test R² declines).

The selected optimal set of seven features is as follow: [G, FC, PI, MC, DD, OMC, MDD].The parameter range and the optimized hyperparameters value of the Hybrid TPE-GPR-RFE model are summarized in Table 2.

Table 2

Hyperparameters range and optimized values.

hyperparameters	Range	Optimized value
Kernel	All kernels (eq 2-6)	Rational Quadratic
Length_scale	0.01-10	0.6192
alpha_kernel	0.01-10	1.5966
Noise_level (white kernel)	1e-04 - 1	0.04059

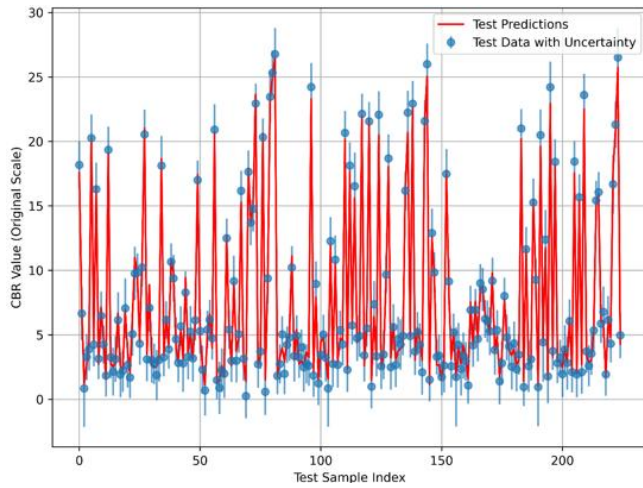


Fig. 7. Test predictions with uncertainty estimates for CBR values.

The hybrid TPE-GPR-RFE model exhibits robust predictive ability, as the test predictions closely align with the actual test data points (see Fig.7). The model accurately represents the general trend and fluctuations in the CBR values. The shaded area denotes the model's forecast uncertainty, which is significantly reduced in areas with a greater density of data points, signifying enhanced confidence.

The uncertainty boundaries effectively include the majority of the actual data, indicating that the model offers a dependable and accurately calibrated assessment of its prediction inaccuracy.

Utilizing the hybrid TPE-GPR-RFE model developed for this investigation, the collected subgrade soils' CBR values were estimated. The 45° line (perfect fit) in Fig.8 represents the fitted line between the experimental and model-predicted CBR values. It is clear that the fitted line and the 45° line are near each other. The coefficient of determination (R²) values of the fitted curve on the train and test sets, respectively, are 99%, indicating an adequate fit.

As a result, the obtained experimental laboratory data closely aligns with the predicted outcomes.

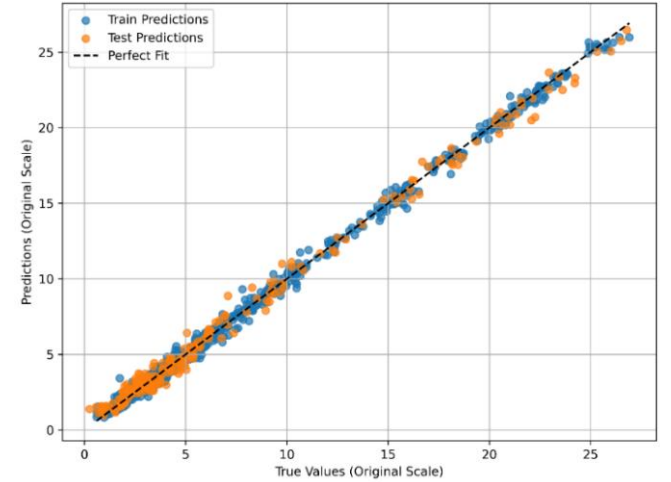


Fig. 8. The graph of predicted and measured CBR values with the hybrid GPR model.

The residual analysis findings are also shown in Fig.9 to evaluate the model's compatibility. If the residual data points are dispersed about the zero line, the model is considered satisfactory. The residual plot is a scatter plot. Every residual data point is located close to the zero line, as seen in Fig.9. This finding supports the validity and dependability of the model.

For both the train and test sets, all error metrics (MSE, MAE, and MAPE) have minimum values. Additionally, the values of the coefficient of determination (R²) are near 1. The prediction ability of the model used in this investigation is really high, according to the data displayed in Table 3. Notably, the predictions of the hybrid model are believed to be accurate.

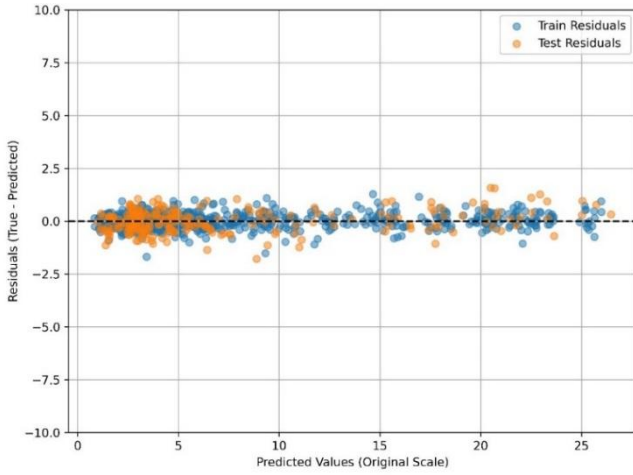


Fig. 9. Plot of residuals for the hybrid GPR model.

Table 3

Performance evaluation metrics for the hybrid GPR model.

Regression metrics	Train set	Test set
MSE	0.0768	0.2974
MAE	0.2107	0.4364
MAPE	5.32 %	13.79 %
R ²	0.9982	0.9935

5.2. Comparison with Conventional and Hybrid AI models

In order to enable a thorough comparison with the hybrid TPE-GPR-RFE model (noted in what follows H-GPR) in terms of their performance and dependability in predicting CBR values, four conventional machine-learning techniques were carefully selected in the first stage: Multi-Linear Regression (MLR), Multi-Layer Perceptron of ANN (MLPNN), Random Forest (RF), and the GPR model with default hyperparameters. Using TPE for hyperparameter tuning and RFE for feature selection, four hybrid models the hybrid such: dense neural network model (H-DNN) with five hidden layers, hybrid XGBoost (H-XGB), hybrid LightGBM (H-LGB), and hybrid AdaBoost (H-Ada) are used for the comparison task in the second stage.

Table 4 presents the summarized results of performance metrics for both the hybrid GPR model and other conventional machine learning models. The Multiple Linear Regression (MLR) model demonstrates suboptimal performance, evidenced by elevated error metrics (MSE: 31.51/38.78, MAPE: 96/103%) and limited explanatory power (R²: 0.15/0.23), suggesting inadequate fit to both training and test datasets. The MLP Neural Network demonstrates notable advancements compared to MLR; however, it continues to display considerable errors. Random Forest (RF) and standard Gaussian Process Regression (GPR) exhibit strong performance, achieving high R² values ranging from 0.95 to 0.99 on the test set; however, RF presents superior error metrics. The H-GPR hybrid model demonstrates exceptional performance, achieving an R² range of 0.9935 to 0.9982 and minimal errors, with MSE values between 0.0768 and 0.2974, and MAPE ranging from 5.32% to 13.79% across both datasets.

Table 4

The performance comparison of hybrid GPR and conventional models using the evaluation metrics.

Model	Train set				Test set			
	MSE	MAE	MAPE	R ²	MSE	MAE	MAPE	R ²
MLR	31.51	4.70	103.40%	0.2334	38.78	4.86	96.53%	0.1495
MLPNN	2.99	1.1927	27.44%	0.9316	3.70	1.24	28.28%	0.9189
RF	0.2520	0.2840	7.01%	0.9942	2.01	0.84	23.94%	0.9558
GPR	0.0	0.0	0.0	1	1.64	0.94	23.00%	0.9640
H-GPR	0.0768	0.2107	5.32 %	0.9982	0.2974	0.4364	13.79 %	0.9935

Results presented in the Table 5, confirm also that the hybrid Gaussian Process Regression (H-GPR) model performs exceptionally well on all performance criteria, exhibiting exceptional predictive power. With a test R² of 0.9550, the H-XGB model likewise produces competitive performance; however, a slight generalisation gap may be indicated by its larger variance between training and testing. Across both datasets, the H-DNN model consistently performs well, with steady R² values of roughly 0.94. The H-LGB model, on the other hand, exhibits a more pronounced decline in generalisation along with a modest predictive strength. With greater error rates and lower R² values, the H-Ada model performs the worst, suggesting that it is not very suitable for this prediction job. Overall, the findings support the notion that H-GPR offers the best trade-off between generalisation and predictive accuracy.

The five different ML models' prediction performance is visually compared with the actual CBR values in Fig.10's scatter plot. The model's prediction accuracy is indicated by how close data points are to the "Perfect Fit" line. The Multi-Linear Regression model exhibits suboptimal performance, characterized by widely dispersed predictions and considerable deviation from actual values, especially in the underestimation of higher CBR values. The ensemble and probabilistic models, including Random Forest, MLP Regressor, and both Gaussian Process models, demonstrate significantly superior performance, as evidenced by their predictions clustering more closely around the ideal line. The Hybrid Gaussian model exhibits superior performance, as its data points align closely with the perfect fit line throughout the value range, indicating that it accurately and successfully reflects the underlying relationship in the data. The visual assessment corroborates the quantitative findings from the evaluation metrics, illustrating a distinct hierarchy of model performance ranging from basic linear regression to advanced hybrid methods.

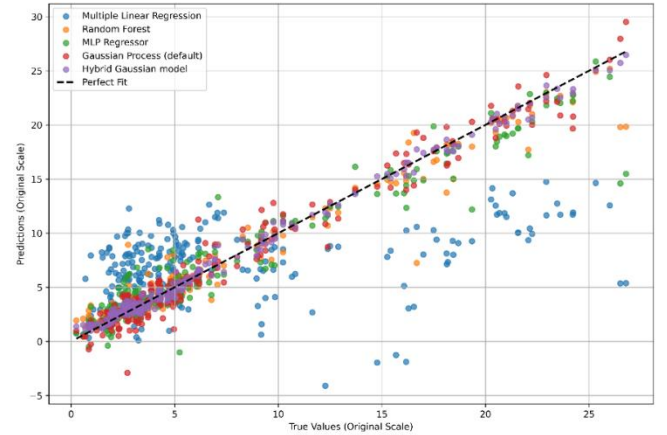


Fig. 10. Conventional models performance comparison with the hybrid GPR model: Predicted vs measured CBR Values.

This residual plot in Fig.11 provides crucial insights into the error patterns and predictive biases of the five different regression models. The distribution of residuals reveals significant differences in model performance. The Multiple Linear Regression model shows the widest dispersion of residuals, indicating substantial prediction errors across the value range, with no clear pattern, suggesting both high variance and potential bias. The Random Forest, MLP reg-

ressor, and Gaussian Process models demonstrate much tighter residual clusters near the zero line, indicating superior predictive accuracy. Particularly noteworthy is the Hybrid Gaussian model, which shows the most concentrated residual distribution around zero, suggesting minimal systematic bias and the highest precision among all models. The relatively symmetric distribution of residuals above and below the zero line for the top-performing models indicates they are not consistently over-predicting or under-predicting the CBR values. This visual analysis complements the quantitative metrics by showing not just the magnitude of errors but also their distribution and pattern, confirming that the more sophisticated algorithms substantially outperform traditional linear regression for this predictive modelling task.

Table 5

The performance comparison of hybrid GPR and other hybrid models using the evaluation metrics.

Model	Train set				Test set			
	MSE	MAE	MAPE	R ²	MSE	MAE	MAPE	R ²
H-DNN	2.40	1.16	29.46%	0.9450	2.61	1.22	32.22%	0.9425
H-XGB	0.15	0.29	7.59%	0.9965	2.05	0.94	24.48%	0.9550
H-LGB	0.84	0.61	14.21%	0.9808	4.44	1.23	30.63%	0.9025
H-Ada	6.33	1.86	49.31%	0.8552	8.11	2.01	61.10%	0.8220
H-GPR	0.0768	0.2107	5.32 %	0.9982	0.2974	0.4364	13.79 %	0.9935

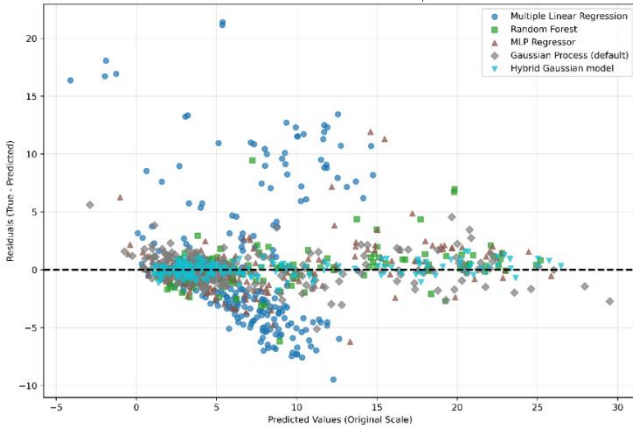


Fig. 11. Residual plot: Conventional models vs hybrid GPR model

Using an inclusive Taylor chart to assess and illustrate the performance of all hybrid models, a thorough evaluation of the Hybrid Gaussian model's prediction reliability was carried out on the testing set (unseen data).

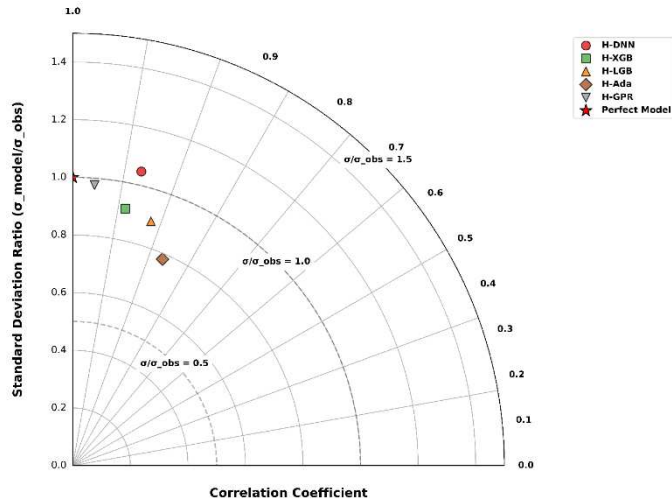


Fig. 12. Taylor diagram of the hybrid models using testing set.

The Taylor diagram contrasting the Hybrid Gaussian model's performance accuracy with the other hybrid models is shown in Fig.12. Because the Hybrid Gaussian model is located closer to the observation point than the other models, the graphic demonstrates its better prediction ability. This placement highlights how the Hybrid Gaussian model outperforms its competitors in terms of prediction accuracy and dependability.

5.3.Shap analysis Results

Because fundamental soil characteristics like CBR impact the durability and resource efficiency of engineering projects, Geotechnics is crucial for the

creation of sustainable infrastructure. In line with international sustainability objectives, this study improves the interpretability of CBR projections, enabling resource-efficient design and construction [76]. In order to help engineers and decision-makers make informed decisions about material selection, soil stabilization, and environmentally friendly construction techniques, explainable AI provides interpretability that enables them to understand the basic principles impacting CBR values.

By fusing explainable artificial intelligence (XAI) with precise machine learning models to improve transparency in predictive modeling, this study substantially broadens the field of soft computing. By using XAI approaches like Shapley Additive Explanations (SHAP), the "black-box" issue that is occasionally associated with complex models is mitigated and both global and local interpretability are provided. By offering insights into feature significance and model behavior, this approach enhances the reliability and credibility of ML applications in important domains, expanding its usefulness beyond the particular use of California Bearing Ratio (CBR) prediction. This section makes it easier to understand the observed patterns in the context of CBR prediction for subgrade soil by providing a succinct explanation of each feature's impact, model-specific insights, practical implications, and technical justifications.

For global interpretations of machine learning models estimating subgrade soil behavior, SHAP (Shapley Additive exPlanations) summary plots are shown in Fig. 13. The optimized dataset consists of only the seven characteristics that were chosen. The most important feature is at the top of the plot's y-axis, which lists the input parameters (features) in order of significance. The x-axis illustrates SHAP values, which measure the influence of each feature on the model's output; positive SHAP values elevate predictions, while negative ones diminish them. The input parameters are emphasized in the graphs. Features positioned higher exert the most significant impact on the model's predictions. The color-coding (pink for elevated values and blue for diminished values) illustrates how fluctuations in feature values influence predictions. MDD (Maximum Dry Density) and DD (sample Dry Density during test conditions) are the most significant characteristics, exhibiting consistently positive SHAP values, which suggest that elevated density values reliably enhance the projected CBR values [76]. G (Gravel content) exhibits the second most significant impact. The Optimum Moisture Content (OMC) and Plasticity Index (PI) have a moderate negative correlation, suggesting that increased values of both parameters typically result in lower predicted California Bearing Ratio (CBR) values. FC and MC (Fine Content and Moisture Content) exhibit the least influence, with values concentrated around zero, indicating a negligible effect on the CBR estimates within this model. The findings offer critical insights into the soil parameters that most significantly influence California Bearing Ratio estimations, with Maximum Dry Density identified as the

predominant positive component, consistent with geotechnical engineering standards that denote denser soils as having greater bearing capacity.

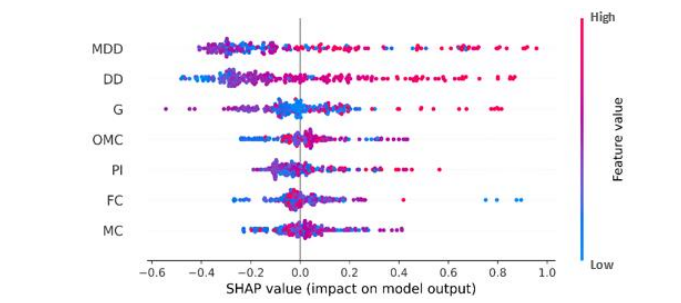


Fig. 13. SHAP global explanations results using the hybrid Gaussian model.

Fig. 14 illustrates the average SHAP values of the features employed to predict the subgrade soil CBR values applying the hybrid Gaussian model. The research underscores the significance of all input parameters, with MDD identified as the most relevant element, substantially affecting predictions.

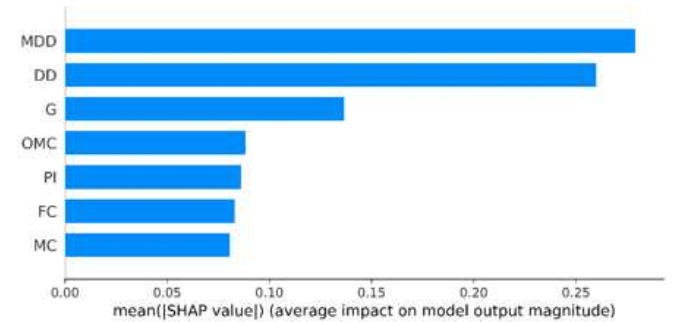


Fig. 14. Mean SHAP global explanations.

Fig.15 demonstrates a robust positive association between MDD and CBR values, showing that expected soil strength is significantly increased by raising the maximum dry density. The connection demonstrates non-linear characteristics with a significant threshold effect at MDD values of 1.7-1.8 g/cm³, beyond which the favorable influence on CBR markedly intensifies. MDD values beyond 1.8 exhibit the most significant advantageous effect, whereas values below 1.7 indicate no or slightly adverse impact, underscoring the essential role of proper compaction in attaining optimal soil bearing capacity.

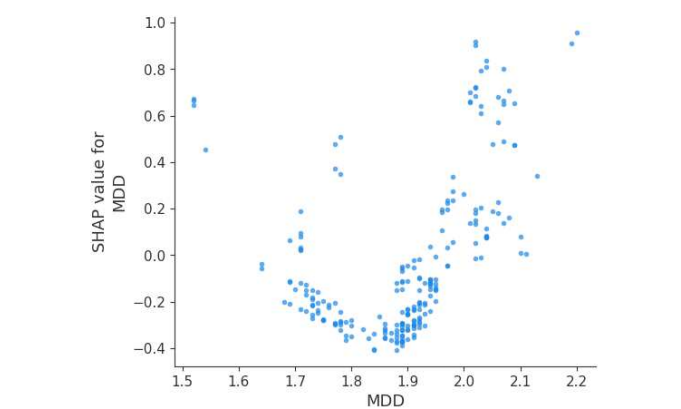


Fig. 15. Dependence plot of MDD.

The SHAP dependence plot in Fig. 16 illustrates a nuanced relationship in which MDD values below approximately 1.7 g/cm³ exert negligible influence on CBR predictions, values ranging from 1.7 to 1.9 display a progressively positive impact, and MDD values above 1.9 yield the most significant advantageous effect on predicted soil strength. The color gradient signifies an interaction effect, suggesting that another soil parameter influences the relationship between MDD and CBR, likely moisture content or compaction energy, with darker shades denoting higher values of this moderating feature that amplify MDD's positive effect on bearing capacity.

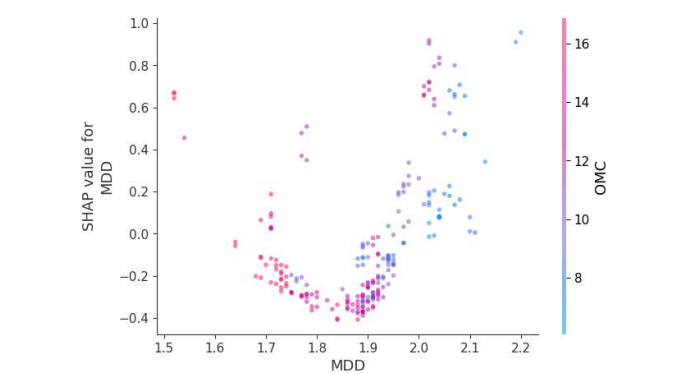


Fig. 16. Dependence plot of MDD with OMC interaction.

Fig. 17 illustrates that the most substantial positive SHAP values (maximum CBR advantage) occur at elevated MDD levels (>1.9 g/cm³) coupled with low PI, while even well compacted soils (MDD >1.8) exhibit diminished strength improvements when linked to high plasticity. This suggests that attaining appropriate density might partially alleviate but not entirely surmount the inherent strength constraints imposed by plastic clayey soils, underscoring the relevance of plasticity index as a limiting factor in subgrade performance.

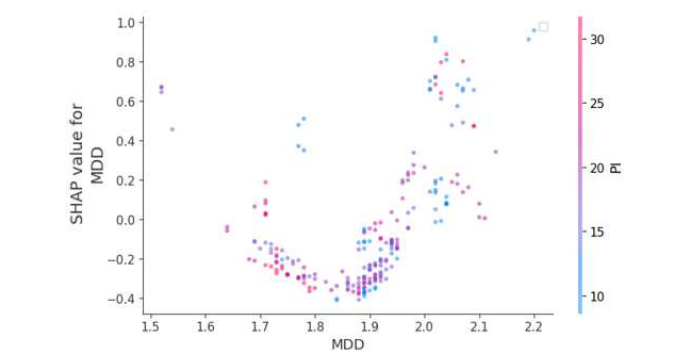


Fig. 17. Dependence plot of MDD with PI interaction.

A number of 3D graphs were generated in order to examine the second order effect of the gravel content (G), PI, MDD, and OMC in the prediction of CBR values.

Fig. 18's 3D SHAP plot illustrates how gravel content (G), optimum moisture content (OMC), and maximum dry density (MDD) interact to determine subgrade soil CBR. The contribution of MDD to model output is shown on the vertical axis (MDD SHAP value), which varies with OMC (horizontal axis) and gravel concentration (colour gradient). Higher positive SHAP contributions from MDD are often correlated with warmer gravel levels and mid-range OMC values, indicating that balanced moisture and more gravel improve density-related CBR performance. This demonstrates how compaction control and soil gradation work together to increase subgrade strength.

Fig. 19's 3D SHAP plot shows how the interaction between gravel content (G) and maximum dry density (MDD) affects how crucial optimum moisture content (OMC) is for CBR prediction. Gravel content (colour) and MDD levels

(horizontal axis) affect the OMC SHAP value (vertical axis). OMC's beneficial effect is typically enhanced by higher gravel content, particularly at moderate to high MDD values. This suggests that accurate moisture regulation is crucial for attaining the best subgrade strength in well-compacted, gravel-rich soils. The 3D Interaction SHAP plot in Figure 20 most likely shows how the SHAP contribution of maximum dry density (MDD) in forecasting CBR is influenced by the optimal moisture content (OMC) and plasticity index (PI). The relevance of obtaining high density (MDD) for subgrade strength is significantly influenced by soil plasticity and moisture levels, as can be seen by plotting the MDD SHAP value versus PI and OMC. According to this interaction figure, MDD values have a favourable impact on CBR values for low PI and OMC values, whereas high PI and OMC values show the opposite pattern.

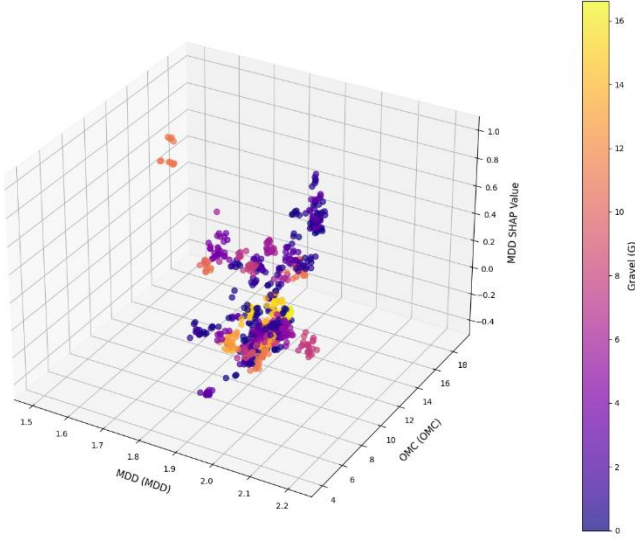


Fig. 18. 3D interaction plot MDD SHAP value with G, OMC, MDD.

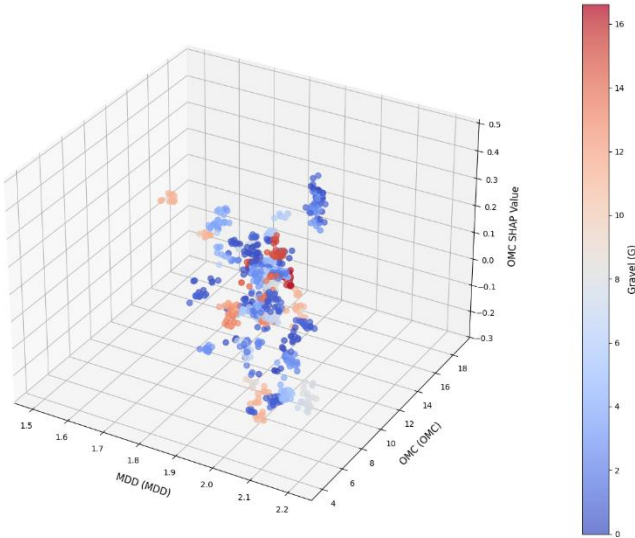


Fig. 19. 3D interaction plot OMC SHAP value with G, OMC, MDD.

The 3D SHAP interaction plot from Fig. 21 examine the impact of optimal moisture content (OMC) on CBR prediction in relation to the plasticity index (PI) and maximum dry density (MDD). When the OMC SHAP value is plotted against PI and MDD, it becomes clear that soil plasticity and compaction level have a significant impact on OMC's contribution. OMC has a significant positive impact on high-plasticity soils at lower densities, highlighting the

importance of moisture control in enhancing the strength of cohesive, less compacted subgrades.

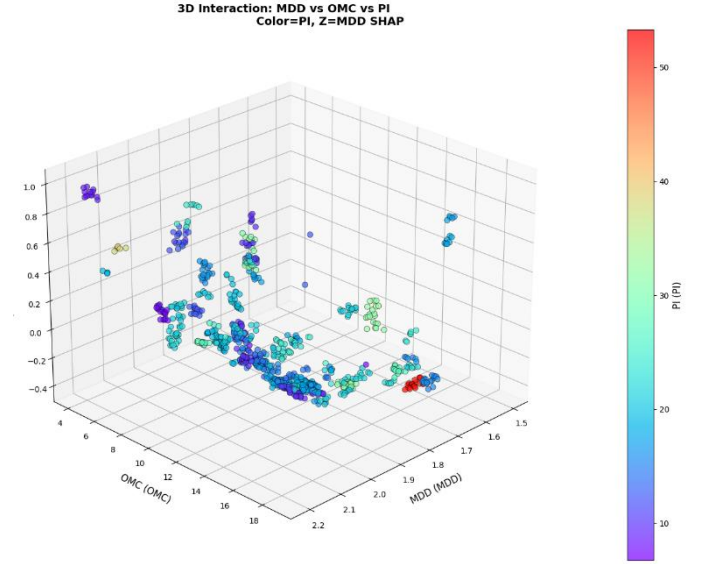


Fig. 20. 3D interaction plot MDD SHAP value with PI, OMC, MDD.

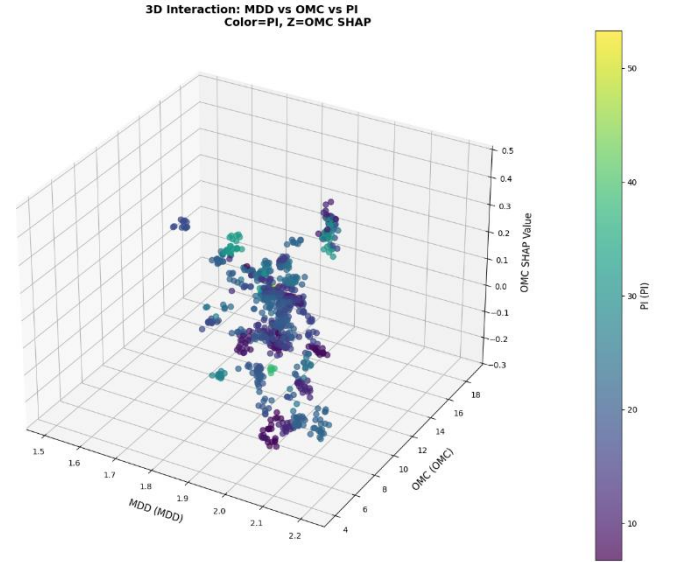


Fig. 21. 3D interaction plot OMC SHAP value with PI, OMC, MDD.

The SHAP waterfall plot presented in Fig. 22 provides a local explanation derived from SHAP analysis results on a randomly selected sample. The model's base prediction value (expected value) is 6.561, whereas the actual predicted CBR value for this specific sample is 4.074. Each feature's contribution to changing the forecast from the base value to the final output is shown in the graphic. Feature DD, with a value of 2.01, yields the most significant positive contribution of +1.72, whereas feature G, with a value of 11, results in the largest negative contribution of -1.72. Additional features such as OMC, FC, PI, MC, and MDD contribute incrementally, with their positive and negative impacts largely offsetting one another, leading to a net reduction of approximately 2.487 from the base prediction.

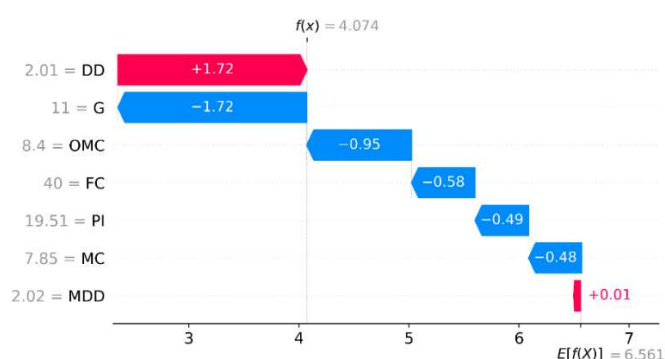


Fig. 22. Waterfall plot for local explanation.

6. Conclusions

Assessing the strength of subgrade soils is crucial and constitutes a major focus of research in geotechnical engineering related to road construction. This study incorporates an extensive dataset of subgrade soil strength derived from CBR testing, consisting of more than one thousand samples. To enhance CBR value prediction, a hybrid Gaussian model (H-GPR) was developed, which couples a TPE-optimized Gaussian Process Regressor (GPR) with Recursive Feature Elimination (RFE) for effective feature selection. Five conventional machine learning models and four hybrid models are employed for comparative study. The "black box" problem with the prediction model is addressed by analyzing prediction results using the SHAP interpretability method. The visualization of SHAP values for the input feature parameters is used to analyze the significance and impact of factors on prediction outcomes. Several regression measures are used to assess the developed model's performance. The subsequent conclusions were primarily derived:

- 1- Prior to the creation of the model, statistical analysis was done on a comprehensive database that included more than 1000 samples from the northeastern part of Algeria that were related to subgrade soil characteristics. The input and output properties showed a moderate association, according to the Pearson correlation analysis.
- 2- The iterative method implementing the TPE-GPR-RFE hybrid model effectively identifies the ideal subset from the complete original dataset (feature selection) and subsequently fine-tunes the hyperparameters utilizing the Bayesian TPE algorithm. This method facilitates both dimensionality reduction and enhancement of predictive accuracy.
- 3- The R^2 value between the predicted and measured values is remarkably close to one, which is responsible for the concordance between the calculated CBR values from the hybrid model and laboratory data. Furthermore, the suggested model's residual data points are located close to the zero line, confirming the model's dependability.
- 4- Of the machine learning models examined in this work, the suggested hybrid GPR (H-GPR) model has the lowest regression error metrics; nevertheless, its coefficient of determination shows the highest value. The findings show that the hybrid Gaussian model outperformed every machine learning model examined in this investigation.
- 5- The SHAP interpretability method is employed to examine the prediction outcomes, addressing the "black box" issue of the prediction model. By visualizing the SHAP values of the input feature parameters, the importance and influence of the parameters on the CBR value prediction results are investigated. According to the SHAP analysis, the main determinants of subgrade soil strength as determined by CBR values are dry density, gravel content, and optimal moisture

content. This model serves as a crucial framework for estimating CBR values and skillfully illustrates the many nonlinear relationships within the feature parameters.

- 6- The anticipated result of this study will yield a dependable technique for assessing the CBR values of subgrade soil for pavement design. Therefore, accurate CBR values are essential for road infrastructure planning and construction. The hybrid Gaussian model established throughout this study can be regarded as a dependable tool.
- 7- A notable constraint of the research is the extent and diversity of the experimental dataset, which may limit its usefulness in broader geotechnical contexts. To increase robustness, future studies should focus on cutting-edge ML methodologies and larger, regionally diverse datasets.

Consent for publication: "The author of this manuscript have provided their consent for the publication of this research."

Declaration of competing interest: "No conflicts of interest."

Acknowledgment: "The authors extended their appreciation to Umm Al-Qura University for funding this research work through the project number: 25UQU4280323GSSR04"

Funding: "This research work was funded by Umm Al-Qura University, Saudi Arabia under grant number: 25UQU4280323GSSR04"

Data availability: "The datasets used and/or analysed during the current study available from the corresponding author on reasonable request."

References

- [1] T. Taskiran, Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. *Advances in Engineering Software*. 41(6) (2010) 886–892. <https://doi.org/10.1016/j.advengsoft.2010.01.003>
- [2] V. Quan, H.Q. Do, Prediction of California bearing ratio (CBR) of stabilized expansive soils with agricultural and industrial waste using light gradient boosting machine. *Journal of Science and Transport Technology*. 1(1) (2022). <https://doi.org/10.58845/jstt.utt.2021.en3>
- [3] A. Bardhan, A. K. Alzo'ubi, S. Palanivelu, P. Hamidian, A. GuhaRay, G. Kumar, M.Z. Tsoukalas, P.G. Asteris, A hybrid approach of ANN and improved PSO for estimating soaked CBR of subgrade soils of heavy-haul railway corridor. *International Journal of Pavement Engineering*. 24(1)(2023). <https://doi.org/10.1080/10298436.2023.2176494>
- [4] T. Al-Refaei, A. Al-Suhaibani. Prediction of CBR using dynamic cone penetrometer. *Journal of King Saud University - Engineering Sciences*. 9 (1997) pp. 191–203. [https://doi.org/10.1016/S1018-3639\(18\)30676-7](https://doi.org/10.1016/S1018-3639(18)30676-7)
- [5] S. Alzabeebe, S.A. Mohamad, R.K.S. Al-Hamd. Surrogate models to predict maximum dry unit weight, optimum moisture content and California bearing ratio from grain size distribution curve. *Road Materials and Pavement Design*. 23 (12)(2021) 2733–2750. <https://doi.org/10.1080/14680629.2021.1995471>
- [6] D.T. Bui, V.H. Nhu, N.D. Hoang. Prediction of soil compression coefficient for urban housing project using novel integration machine learning approach of swarm intelligence and multi-layer perceptron neural network. *Advanced Engineering Informatic*. 38 (2018) 593–604. <https://doi.org/10.1016/j.aei.2018.09.005>
- [7] B. Yildirim, O. Gunaydin. Estimation of California bearing ratio by using soft computing systems. *Expert Systems with Applications*. 38 (2011) 6381–6391. <https://doi.org/10.1016/j.eswa.2010.12.054>
- [8] K.J. Kumar Pradeep, Y.M. Patel Harish. Soft computing technique for prediction of CBR from index properties of subgrade soil. (2016) 13852–13860. <https://doi.org/10.15680/IJIRSET.2016.0507214>
- [9] S.K. Alam, A. Mondal, A. Shiuly. Prediction of CBR value of fine grained soils of Bengal basin by genetic expression programming, artificial neural network and kriging method. *Journal of the Geological Society*. 95 (2020) pp. 190–196. <https://doi.org/10.1007/s12594-020-1409-0>
- [10] A.R. Tenpe, A. Patel. Utilization of support vector models and gene expression programming for soil strength modeling. *Arabian Journal for Science and Engineering*. 45 (2020) 4301–4319. <https://doi.org/10.1007/s13369-020-04441-6>
- [11] A.K. Sabat. Prediction of California bearing ratio of a stabilized expansive soil using artificial neural network and support vector machine. *Electronic Journal of Geotechnical Engineering*. 20 (2015) 981–991.
- [12] T.F. Kumaz, Y. Kaya. Prediction of the California bearing ratio (CBR) of compacted soils by using GMDH-type neural network. *The European Physical Journal Plus*. 134 (7)326 (2019). <https://doi.org/10.1140/epjp/i2019-12692-0>

- [13] V.K. Varghese, et al. Artificial neural networks: a solution to the ambiguity in prediction of engineering properties of fine-grained soils. *Geotechnical and Geological Engineering*. 31 (2013) 1187–1205. <https://doi.org/10.1007/s10706-013-9643-5>
- [14] Y. Erzin, D. Turkoz. Use of neural networks for the prediction of the CBR value of some Aegean sands. *Neural Computing and Applications*. 27 (2016) 1415–1426.
- [15] A. Ghorbani, H. Hasanazadehshooili. Prediction of UCS and CBR of micro-silica-lime stabilized sulfate silty sand using ANN and EPR models; application to the deep soil mixing. *Soils and Foundations*. 58 (2018)pp. 34–49. <https://doi.org/10.1016/j.sandf.2017.11.002>
- [16] H.L. Wang, Z.Y. Yin. High performance prediction of soil compaction parameters using multi expression programming. *Engineering Geology*. 276 (2020) 105758. <https://doi.org/10.1016/j.enggeo.2020.105758>
- [17] S. Al-Busultan, et al. Application of artificial neural networks in predicting sub-base CBR values using soil indices data. *IOP Conference Series: Materials Science and Engineering*. 671 (2020). <https://doi.org/10.1088/1757-899X/671/1/012106>
- [18] A. Assolie. Advanced modeling techniques using hierarchical Gaussian process regression in civil engineering. *Asian Journal of Civil Engineering*. 25(7) (2024) 5599–5612. <https://doi.org/10.1007/s42107-024-01132-7>
- [19] M. Q. Peng, Z. C. Qiu, S. Shen, Y. Li, J. Zhou, H. Xu. Geotechnical Site Characterizations Using a Bayesian-Optimized Multi-Output Gaussian Process. *Sustainability*. 16(13) (2024) 5759. <https://doi.org/10.3390/su16135759>
- [20] J. Wang, P. Filippi, S. Haan, L. Pozza, B.M. Whelan, T. Bishop. Gaussian process regression for three-dimensional soil mapping over multiple spatial supports. *Geoderma*. 444 (2024) 116899. <https://doi.org/10.1016/j.geoderma.2024.116899>
- [21] Z. Li, P. Yi. Probabilistic design optimization of rectangular skirted mud-mats under combined loading in spatially variable soil. *Computers and Geotechnics*. 169 (2024) 106260. <https://doi.org/10.1016/j.compgeo.2024.106260>
- [22] C. Varady, J. Tenório, E. Silva, E.L. Junior, J. Santos, R. Dias, F. Cutrim. Bayesian-based approach in soil characterization for top-hole design. *Offshore Technology Conference*. (2024). <https://doi.org/10.4043/35039-ms>
- [23] M. Ahmad, S. Keawsawasvong, M.R. Bin Ibrahim, M. Waseem, K.R. Kashyadeh, M.M. S. Sabri. Novel approach to predicting soil permeability coefficient using Gaussian process regression. *Sustainability*. 14(14) (2022) 8781. <https://doi.org/10.3390/su14148781>
- [24] M. Rastgou, H. Bayat, M. Mansoorzadeh, A.S. Gregory. Prediction of soil hydraulic properties by Gaussian process regression algorithm in arid and semiarid zones in Iran. *Soil Tillage Res.* 210 (2021) 104980.
- [25] E. Momeni, M.B. Dowlatshahi, F. Omidinasab, H. Maizir, D.J. Armaghani. Gaussian process regression technique to estimate the pile bearing capacity. *Arab. J. Sci. Eng.* 45 (2020) 8255–8267.
- [26] H. Fathipour-Azar. Hybrid machine learning-based triaxial jointed rock mass strength. *Environ. Earth Sci.* 81 (2022) 118.
- [27] A. Mahmoodzadeh, M. Mohammadi, H.H. Ibrahim, T.A. Rashid, A.H.M. Aldalwie, H.F.H. Ali, A. Daraei. Tunnel geomechanical parameters prediction using Gaussian process regression. *Mach. Learn. with Appl.* 3 (2021) 100020.
- [28] C.K. Arthur, V.A. Temeng, Y.Y. Ziggah. Novel approach to predicting blast-induced ground vibration using Gaussian process regression. *Eng. Comput.* 36 (2020) pp. 29–42.
- [29] I. Ocak, S.E. Seker. Calculation of surface settlements caused by EPBM tunneling using artificial neural network, SVM, and Gaussian processes. *Environ. Earth Sci.* 70 (2013) 1263–1276.
- [30] P. He, S. Li, J. Xiao, Q. Zhang, F. Xu, J. Zhang. Shallow sliding failure prediction model of expansive soil slope based on Gaussian process theory and its engineering application. *KSCE J.Civ. Eng.* 22 (2018) 1709–1719.
- [31] M.H. A.Khan, T.H. Jafri, S. Ud-Din, H.S. Ullah, M.N. Nawaz. Prediction of soil compaction parameters through the development and experimental validation of Gaussian process regression models. *Environmental Earth Sciences*. 83(4)(2024). <https://doi.org/10.1007/s12665-024-11433-4>
- [32] L. Zhang, X. Zheng, Q. Pang, W. Zhou. Fast Gaussian kernel support vector machine recursive feature elimination algorithm. *Applied Intelligence*. 51(12) (2021) 9001–9014. <https://doi.org/10.1007/s10489-021-02298-2>
- [33] A.M. Priyatno, T. Widiyaningtyas. A systematic literature review: Recursive feature elimination algorithms. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*. 9(2)(2024) 196–207. <https://doi.org/10.33480/jitk.v9i2.5015>
- [34] C. Chen, J. Liang, W. Sun, G. Yang, X. Meng. An automatically recursive feature elimination method based on threshold decision in random forest classification. *Digui Kongjian Xinxu Kexue Xuebao [Geo-Spatial Information Science]*. (2024)pp .1–26. <https://doi.org/10.1080/10095020.2024.2387457>
- [35] O. Bulut, B. Tan, E. Mazzullo, A. Syed. Benchmarking variants of Recursive Feature Elimination: Insights from predictive tasks in education and healthcare. *Information (Basel)*. 16(6) (2025) 476. <https://doi.org/10.3390/info16060476>
- [36] J. Frankel, F. Tahmoorian. Improving Gravel Material Specifications for Unpaved Roads: Australian Case Study. *Int J Pavement Res Technol.* (2023) pp. 1–14. <https://doi.org/10.1007/s42947-023-00355-2>
- [37] Y.I. Murthy, S. Pandey, S. Gandhi. Prognostication of Half Cell Potential for Slabs Cathodically Protected with A291D using eXplainable and Interpretable Machine Learning Anti Corrosion. *Methods and Materials*. (2024). <https://doi.org/10.1108/ACMM-09-2024-3094>
- [38] I. Thapa, S. Ghani, K.A. Waris, B.M. Basha. Prediction of California bearing ratio of nano-silica and bio-char stabilized soft subgrade soils using explainable machine learning. *Transportation Geotechnics*. 49 (2024) 101387. <https://doi.org/10.1016/j.trgeo.2024.101387>
- [39] M.K.S. Gowda, Y.I. Murthy, A. Gupta. Explainable AI based support vector machine models for soaked CBR prediction. *International Journal of Pavement Research and Technology*. (2025). <https://doi.org/10.1007/s42947-025-00558-9>
- [40] B. Chen, H. Zheng, G. Luo, C. Chen, A. Bao, T. Liu, X. Chen. Adaptive estimation of multi-regional soil salinization using extreme gradient boosting with Bayesian TPE optimization. *International Journal of Remote Sensing*. 43(3) (2022) 778–811. <http://dx.doi.org/10.1080/01431161.2021.2009589>
- [41] T. Kavzoglu, A. Teke. Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). *Bulletin of Engineering Geology and the Environment*. 81(5) (2022). <https://doi.org/10.1007/s10064-022-02708-w>
- [42] J. Mockus. *Bayesian Approach to Global Optimization*. Springer. Dordrecht. (2012) pp. 270. <http://dx.doi.org/10.1007/978-94-009-0909-0>
- [43] J. Yu, W. Zheng, L. Xu, F. Meng, J. Li, L. Zhangzhong. TPE-catboost: An adaptive model for soil moisture spatial estimation in the main maize-producing areas of China with multiple environment covariates. *Journal of Hydrology*. 613 (2022) 128465. <https://doi.org/10.1016/j.jhydrol.2022.128465>
- [44] B. Behkamal, A. Entezami, C. De Michele, A. N. Arslan. Investigation of temperature effects into long-span bridges via hybrid sensing and supervised regression models. *Remote Sensing*. 15(14)(2023) 3503. <https://doi.org/10.3390/rs15143503>
- [45] S. Manzhos, M. Ihara. Optimization of hyperparameters of Gaussian process regression with the help of a low-order high-dimensional model representation: application to a potential energy surface. *Journal of Mathematical Chemistry*. 60(10) (2022) 1797–1815. <https://doi.org/10.1007/s10910-022-01407-x>
- [46] P. Ashrafi, Y. Sun, N. Davey, R.G. Adams, S.C. Wilkinson, G.P. Moss. Model fitting for small skin permeability data sets: hyperparameter optimisation in Gaussian Process Regression. *Journal of Pharmacy and Pharmacology*. 70(3)(2018) 361–373. <https://doi.org/10.1111/jphp.12863>
- [47] P.L. Lualdi, R. Sturm, A. Camero, T. Siefkes, An uncertainty-based objective function for hyperparameter optimization in Gaussian processes applied to expensive black-box problems. *Applied Soft Computing*. 154 (2024) 111325. <https://doi.org/10.1016/j.asoc.2024.111325>
- [48] I. Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* 46 (2002) 389–422.
- [49] C. Cheng, Y. Yang, F. Zhong, C. Song, Y. Zhen. An optimization of statistical index method based on Gaussian process regression and GeoDetector, for higher accurate landslide susceptibility modeling. *Applied Sciences*. 12(20)(2022) 10196. <https://doi.org/10.3390/app122010196>
- [50] G. Rong, et al. Comparison of Tree-Structured Parzen estimator optimization in three typical neural network models for landslide susceptibility assessment. *Remote Sensing*. 13(22) (2021) 4694. <https://doi.org/10.3390/rs13224694>
- [51] W. Xiao, C. Wang, J. Liu, M. Gao, J. Wu. Optimizing faulting prediction for rigid pavements using a hybrid SHAP-TPE-CatBoost model. *Applied Sciences*. 13(23) (2023) 12862. <https://doi.org/10.3390/app132312862>
- [52] H.M. Sheng, J. Xiao, Y.H. Cheng, Q. Ni, S. Wang. Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression. *IEEE Trans. Ind. Electron.* 65 (2018) 300–308.
- [53] K.L. Liu, X.S. Hu, Z.B. Wei, Y. Li, Y. Jiang. Modified Gaussian Process Regression Models for Cyclic Capacity Prediction of Lithium-Ion Batteries. *IEEE Trans. Transp. Electr.* 5 (2019) 1225–1236.
- [54] Y. Zhou, Y.F. Liu, D.J. Wang, G. De, Y. Li, X.J. Liu, Y.Y. Wang. A novel combined multi-task learning and Gaussian process regression model for the prediction of multi-timescale and multi-component of solar radiation. *J. Clean. Prod.* 284 (2021) 124710.
- [55] X.Y. Li, C.G. Yuan, X.H. Li, Z.P. Wang. State of health estimation for Li-Ion battery using incremental capacity analysis and Gaussian process regression. *Energy*. 190 (2020) 116467.
- [56] J. Khattai, K.S. Grover. Determination of the optimum performance AI model and methodology to predict the compaction parameters of soils. *ICTACT Journal of Soft Computing*. 12(3)(2022) 2640–2650. <https://doi.org/10.21917/ijsc.2022.0368>
- [57] M. Ludkovski, J. Risk. Gaussian Process models for Quantitative Finance. In *Springer Briefs in quantitative finance*. (2025). <https://doi.org/10.1007/978-3-031-80874-6>
- [58] Z. Lim, Y. Pang, K. Kamarudin, O. Yin, F. Hiew. Bayesian optimization driven strategy for detecting credit card fraud with extremely randomized trees. *Methodsx*. 13 (2024) 103055. <https://doi.org/10.1016/j.mex.2024.103055>
- [59] B. Xu, L. Sun, X. Mao, R. Ding, C. Liu. IoT intrusion detection system based on machine learning. *Electronics*. 12(20) (2023) 4289. <https://doi.org/10.3390/electronics12204289>
- [60] X. Zhang, C. Dai, W. Li, Y. Chen. Prediction of compressive strength of recycled aggregate concrete using machine learning and Bayesian optimization methods. *Frontiers in Earth Science*. 11 (2023). <https://doi.org/10.3389/feart.2023.1112105>
- [61] Ö. Aydoğmuş, M. Yılmaz. Comparative analysis of reinforcement learning algorithms for bipedal robot locomotion. *Ieee Access*. 12 (2024) 7490-7499. <https://doi.org/10.1109/access.2023.3344393>
- [62] Y. Song, et al. Tree-structured parzan estimator-machine learning-ordinary kriging: an integration method for soil ammonia spatial prediction in the typical cropland of Chinese yellow river delta with sentinel-2 remote sensing image and air quality data. *Remote Sensing*. 15(17)(2023) 4268. <https://doi.org/10.3390/rs15174268>
- [63] C. Melainine, F. Socheleau, A. Jarrot. Self-configurable receiver for underwater acoustic communications using bayesian optimization. (2024) pp. 1-5. <https://doi.org/10.1109/ucomms64662.2024.10847986>
- [64] T. Westmeier, H. Hetzler, D. Kreuter, P. Seitz, G. Chalvatzaki. Quantifying uncertainty in neural network predictions of forced vibrations, *Pamm*. 24(4) (2024) <https://doi.org/10.1002/pamm.202400192>

- [65] Y. Cao, H. Xiao, M. He, F. Liao, Q. Xu. Estimation method of earthwork excavation using shield tunneling data a case study of chengdu metro. *Frontiers in Earth Science*. 11 (2023). <https://doi.org/10.3389/feart.2023.1295672>
- [66] Y. Huang, et al. A machine learning approach to brain epigenetic analysis reveals kinases associated with alzheimer's disease. *Nature Communications*. 12(1) (2021). <https://doi.org/10.1038/s41467-021-24710-8>
- [67] J. Tørring, A. Elster. Analyzing search techniques for auto-tuning image-based gpu kernels: the impact of sample sizes. (2022). <https://doi.org/10.48550/arxiv.2203.13577>
- [68] J. Snoek, H. Larochelle, R.P. Adams. Practical Bayesian optimization of machine learning algorithms. *Adv Neural Inf Process Syst*. 25 (2012).
- [69] Y. Ozaki, et al. Multi-objective tree-structured parzen estimator for computationally expensive optimization problems. In *Proceedings of the genetic and evolutionary computation conference*.(2020).
- [70] J. Wilson, F. Hutter, M. Deisenroth. Maximizing acquisition functions for Bayesian optimization. *Adv Neural Inf Process Syst*. 31 (2018).
- [71] X. He, K. Zhao, X. Chu. Auto-ML: a survey of the state of the art. *Knowl-Based Syst*. 212 (2021) 106622.
- [72] X. Tian, J. Wang, C. Kim, X. Deng, Y. Zhu. Automated machine learning exact dirichlet boundary physics-informed neural networks for solid mechanics. *Engineering Structures*. 330 (2025) 119884. <https://doi.org/10.1016/j.engstruct.2025.119884>
- [73] S.M. Lundberg, S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst*. 30 (2017).
- [74] S.M. Lundberg, G. Erion, H. Chen, A. De Grave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell*. 2 (2020) pp. 56–67.
- [75] A. Moncada-Torres, M.C. Van Maaren, M.P. Hendriks, S. Siesling, G. Geleijnse. Explainable Machine Learning Can Outperform Cox Regression Predictions and Provide Insights in Breast Cancer Survival. *Sci. Rep*. 11 (2021) 6968.
- [76] I. Thapa, S. Ghani. Explainable AI-Enhanced machine learning for CBR prediction in stabilized and unstabilized subgrade soils. *Applied Soft Computing*. (2025) 113275. <https://doi.org/10.1016/j.asoc.2025.113275>
- [77] C. Nwa-David, F. Onyeka. Variability Analysis of the Behaviors of a Lateritic Soil Profile as a Structural Material. *Eng Technol J*. 41 (2023) pp. 1–7. <https://doi.org/10.30684/etj.2023.137779.1359>
- [78] V. Rubinić, et al. Plasticity, Mineralogy, and WRB Classification of Some Typical Clay Soils along the Two Major Rivers in Croatia. *Eurasian Soil Sci*. 53 (2020) 922–940. <https://doi.org/10.1134/S1064.22932.00701.21>
- [79] A.S. Zaimoglu, O. Tan, R.K. Akbulut. Optimization of consistency limits and plasticity index of fine-grained soils modified with polypropylene fibers and additive materials. *KSCE J Civ Eng*. 20 (2016) pp. 662–669. <https://doi.org/10.1007/s12205-015-0540-8>
- [80] S.A. Ola, O.A. Usifoh, E.S. Nnochiri. The Effects of Potassium Permanganate on the Geotechnical Properties of Soils. *Acta Polytech*. 63 (2023) pp. 260–267. <https://doi.org/10.14311/AP.2023.63.0260>
- [81] I.I. Akinwumi, et al. Investigation of calcium carbide residue as a stabilizer for tropical sand used as pavement material. *WIT Trans Built Environ*. 182 (2019) pp. 285–294. <https://doi.org/10.2495/UT180.261>
- [82] S. Marathe, A.P. Rodrigues. Intelligent Models for Prediction of Compressive Strength of Geopolymer Pervious Concrete Hybridized with Agro-Industrial and Construction-Demolition Wastes. *Stud Geotech Mech*. 10 (2024) pp. 1–28. <https://doi.org/10.2478/sgem-2024-0020>
- [83] B. Omer, A.S. Mohammed. Large-scale soaked CBR prediction using machine learning: insights from standard and modified proctor test data. *Modeling Earth Systems and Environment*. 11(5) (2025). <https://doi.org/10.1007/s40808-025-02512-x>
- [84] V.Y. Katte, et al. Correlation of California Bearing Ratio (CBR) Value with Soil Properties of Road Subgrade Soil. *Geotech Geol Eng*. 37 (2019) pp. 217–234. <https://doi.org/10.1007/s10706-018-0604-x>
- [85] M.M.M. Taha, C.P. Feng, S.H.S. Ahmed. Influence of Polypropylene Fibre (PF) Reinforcement on Mechanical Properties of Clay Soil. *Adv Polym Technol*. (2020) pp. 1–15. <https://doi.org/10.1155/2020/95128.39>
- [86] S. Gokova. Investigation of the correlation between California bearing ratio and shear strength of pavement subgrade material with different water contents. *J Innov Transp*. 3 (2022) pp. 1–7. <https://doi.org/10.53635/jit.10977.57>
- [87] J. Duque, W. Fuentes, S. Rey, E. Molina. Effect of Grain Size Distribution on California Bearing Ratio (CBR) and Modified Proctor Parameters for Granular Materials. *Arab J Sci Eng*. 45 (2020) 8231–8239. <https://doi.org/10.1007/s13369-020-04673-6>