



Application of KRR, K-NN and GPR Algorithms for Predicting the Soaked CBR of Fine-Grained Plastic Soils

Gaurav Verma¹ · Brind Kumar² · Chintoo Kumar³ · Arunava Ray⁴ · Manoj Khandelwal⁵

Received: 1 February 2023 / Accepted: 22 May 2023 / Published online: 22 June 2023
© The Author(s) 2023

Abstract A bearing ratio (CBR) test is one of the comprehensive tests used for the last few decades to design the pavement thickness of roadways, railways and airport runways. Laboratory-performed CBR test is considerably rigorous and time-taking. In a quest for an alternative solution, this study utilizes novel computational approaches, including the kernel ridges regression, K-nearest neighbor and Gaussian process regression (GPR), to predict the soaked CBR value of soils. A vast quantity of 1011 in situ soil samples were collected from an ongoing highway project work site. Two data divisional approaches, i.e., K-Fold and fuzzy c-means (FCM) clustering, were used to separate the dataset into training and testing subsets. Apart from the numerous statistical performance measurement indices, ranking and overfitting analysis were used to identify the best-fitted CBR prediction model. Additionally, the literature models were also tried to validate through present study datasets. From the results of Pearson's correlation analysis, Sand, Fine Content, Plastic Limit, Plasticity Index, Maximum Dry Density and Optimum Moisture Content were found to be most influencing input parameters in developing the soaked CBR of fine-grained plastic soils. Experimental results also establish the proficiency of the GPR model developed through FCM and K-Fold data division approaches. The K-Fold data division approach was found to be helpful in removing the overfitting of the models. Furthermore, the predictive ability of any model is considerably influenced by the geological location of the soils/materials used for the model development.

Keywords Kernel ridge regression · K-nearest neighbor · Gaussian process regression · Soaked CBR · Fine-grained soil · K-fold cross-validation · Fuzzy c-means clustering

✉ Manoj Khandelwal
m.khandelwal@federation.edu.au;
mkhandelwal1@gmail.com

Gaurav Verma
gauravv.rs.civ16@itbhu.ac.in

Brind Kumar
kumar_brind.civ@iitbhu.ac.in

Chintoo Kumar
chintookr.rs.cse17@itbhu.ac.in

Arunava Ray
arunava.ray@vit.ac.in

- ¹ Department of Civil Engineering, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh 221005, India
- ² Department of Civil Engineering, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh 221005, India
- ³ Department of Computer Engineering, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh 221005, India
- ⁴ Centre for Disaster Mitigation and Management, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India

Abbreviations

AI	Artificial intelligence
ANN	Artificial neural network
ANSI	Adaptive neuro swarm intelligence
AUC	Area under curve
BIS	Bureau of Indian Standards
CBR	California bearing ratio
ELM	Extreme learning machine
FC	Fine content
FCM	Fuzzy c-means
GEP	Gene expression programming
GMDH	Group method of data handling
GP	Genetic programming
GPR	Gaussian process regression
GPR _F	GPR model at FCM data division approach

- ⁵ Institute of Innovation, Science and Sustainability, Federation University Australia, Ballarat, VIC 3350, Australia



GPR _K	GPR model at K-Fold data division approach
IOA	Index of agreement
IOS	Index of scatter
I _P	Performance index
KRR	Kernel ridge regression
KRR _F	KRR model at FCM data division approach
KRR _K	KRR model at K-Fold data division approach
K-NN	K-nearest neighbor
K-NN _F	K-NN model at FCM data division approach
K-NN _K	K-NN model at K-Fold data division approach
LL	Liquid limit
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MARS	Multivariate adaptive regression splines
MDD	Maximum dry density
ML	Machine learning
MLR	Multi-linear regression
OMC	Optimum moisture content
OR	Overfitting ratio
PI	Plasticity index
PL	Plastic limit
PSO	Particle swarm optimization
R^2	Coefficient of determination
Adj. R^2	Adjusted coefficient of determination
R	Correlation coefficient
RA	Ranking analysis
REC	Regression error characteristics curve
RMSE	Root-mean-square error
SLR	Simple linear regression
S _P	Performance strength
SVM	Support vector machine
TR	Training dataset
TS	Testing dataset
USCS	Unified Soil Classification System
VAF	Variance account for

1 Introduction

Road transportation network facilitates transferring goods from one place to another and door-to-door services for passengers throughout the world. As of this, the road transport infrastructures majorly govern the economy of the country. In this aspect, many new expressways and green highways are being constructed in India by the Ministry of Road Transport and Highways (MoRTH) department through various infrastructure development plans. The pavement thickness design construction of these roads is based on the strength of the material used in the subgrade and subbase layer. Therefore, highway engineers always desire that the material used in the subgrade layer should fulfill some of the engineering and

technical properties such as swell criteria, plasticity properties, soil settlement conditions, subgrade reaction, bearing capacity etc. A method for recognizing the strength of such layers is of utmost requisite in highway engineering.

In general, the California bearing ratio (CBR) test is espoused to measure the stiffness modulus and the shear strength of subgrade material [1, 2] which may be performed on either re-compacted samples in the laboratory or undisturbed samples cut from the field or in situ surface of subgrade formation [3]. The test is an indirect measure that compares the strength of subgrade material (at known density and moisture content) to standard crushed rock material [4, 5]. Both laboratory and in situ tests are based on the principle of penetrating a standard dimension plunger into a soil specimen at a deformation rate of 1.25 mm/min. The laboratory and field engineers always encounter several difficulties in obtaining the CBR value in the laboratory. Laboratory-soaked CBR test requires a large amount of materials (almost 6 kg), more effort to prepare the test specimen, and lastly, 96 h of the soaking period to simulate the field conditions. Consequently, all those activities make the CBR test more tedious, laborious, and time-consuming. Additionally, if the properties of soil change for each small stretch of highway, then preserving such a huge quantity of soil and conducting the CBR test in the laboratory is laborious and time-consuming. Laboratories are also often packed due to the long queue of materials testing, which causes a delay in testing as well as the testing reports, ultimately the design of construction projects. Furthermore, the test method includes the material transportation cost (from construction site to testing laboratory), testing charge, and finally, the dumping of tested materials, which became more exhausted and increased the final cost of the projects.

Owing to the aforementioned problems, many researchers considered that CBR needs to be replaced either partially or entirely. Although not a fundamental material property, it has a long history in pavement design, and it is reasonably correlated with the index and engineering properties of soil by several investigators in the past. To the author's knowledge, the first fame in predicting the CBR value was earned by Kleyn [6]. Earlier, he attempted to address the discrepancy in the CBR test and later prepared a chart based on a nest of straight lines that relate CBR to PI and grading module for over 1000 soaked CBR tests obtained from road and airport work throughout central and southern Africa. Black [7] suggested that the relationship between CBR and ultimate bearing capacity depends on the type of soil and compaction method, i.e., static or dynamic. Agarwal and Ghanekar [8] tried to generate the correlation equation through statistical analysis between CBR and Atterberg limits for 48 soil samples collected from different parts of India. However, they could not find any significant correlation between these parameters. But when LL and OMC

Table 1 Brief overview of the literature study attempted to predict the CBR value of various soil types

Literature study	Type of soil	Number of datasets	Computational approach	R^2/R
National Cooperative Highway Research Program [9]	Non-Plastic Coarse-grained soils	7	SLR	0.84
	Plastic Fine-grained soils	11		0.67
Kin [10]	Fine-grained soil	57	MLR	NA
Taskiran [11]	Fine-grained soil	151	ANN	0.91
			GEP	0.92
Yildirim and Gunaydin [12]	Granular soil	124	SLR	0.86
			MLR	0.88
			ANN	0.93
Alawi and Rajab [25]	Granular soil	19	MLR	0.95
Varghese, Babu [26]	Fine-grained soil	112	MLR	0.83
			ANN	0.85
Erzin and Turkoz [13]	Sandy soil	61	MLR	0.81
			ANN	0.98
Tenpe and Patel [16]	Mixed soil samples	389	GEP	0.82
			ANN	0.89
Katte, Mfoyet [27]	Subgrade soil	33	MLR	0.84
Kurnaz and Kaya [22]	Mixed soil samples	158	GMDH	0.97
Taha, Gabr [15]	Granular soil	218	ANN	0.97
Alam, Mondal [28]	Fine-grained soil	20	GEP	0.94
			ANN	0.96
Tenpe and Patel [17]	Mixed soil samples	389	GEP	0.78
			SVM	0.80
Bardhan, Gokceoglu [19]	Mixed soil samples	312	MARS	0.90
			GP	0.88
Bardhan, Samui [18]	Mixed soil samples	312	ELM-MPSO	0.91
			ELM-TPSO	0.90
			SVM	0.87

were added, they observed an improved correlation with adequate accuracy for the preliminary identification of materials. National Cooperative Highway Research Program [9] attempted to develop the correlation equation for CBR from the index properties for clean and coarse-grained soil. Kin [10] tried to develop the correlation equation for the CBR value of fine-grained and coarse-grained soil through gradational properties. Taskiran [11] attempted to establish the correlation for 151 CBR test data of fine-grained soils, taken from 354 test samples, by ANN and gene expression programming (GEP) methods. Both techniques were found to exhibit promising results. Using 124 datasets, Yildirim and Gunaydin [12] studied the estimation of CBR by regression and ANN approach. They observed that the ANN technique is better than the regression analysis. Erzin and Turkoz [13] tried to predict the CBR value of Aegean sand from the results of mineralogical properties through ANN and regression approach. They also found that ANN is superior to the

regression technique. Farias, Araujo [14] used the local polynomial regression (LPR) and radial basis network (RBN) techniques for developing the predictive equations for the CBR of soil samples. Using 207 CBR test results of granular soil, Taha, Gabr [15] observed that the correlation obtained through ANN is of excellent accuracy and lower bias than the regression analysis. A comparative study conducted by Tenpe and Patel [16] for 389 datasets collected from City and Industrial Development Corporation, Maharashtra state in India, reveals that ANN and GEP are efficient in predicting the CBR value. Later, in another study, Tenpe and Patel [17] found that SVM can better predict the CBR value than GEP. Recently, Bardhan, Samui [18] attempted to predict the soaked CBR value of 312 soil datasets through a particle swarm optimization (PSO) algorithm with adaptive and time-varying acceleration coefficients. The comparative analysis of various extreme learning machine (ELM) based adaptive neuro swarm intelligence (ANSI) such as

ELM coupled-modified PSO (ELM-MPSO), ELM coupled-time-varying acceleration coefficients PSO (ELM-TPSO) and ELM coupled-improved PSO (ELM-IPSO) reveals that the modified and improved version of PSO has high accuracy at early iterations than the standard PSO. In another investigation, Bardhan, Gokceoglu [19] observed that multivariate adaptive regression splines with piecewise linear (MARS-L) demonstrate a higher accuracy in predicting the soaked CBR as compared to MARS with piecewise-cubic (MARS-C), Gaussian process regression and genetic programming. Hassan, Alshameri [20] attempted to predict the CBR value of plastic fine-grained soil from their index properties and compaction parameters through multi-linear regression analysis (MLR). The study was conducted for the standard proctor compaction energy level, whereas the engineers always prefer the modified proctor compactive energy level to construct the highways and expressways. It is observed from the above literature investigations (also shortened in Table 1) that several artificial intelligence (AI)-based models were used to predict the soaked CBR value which demonstrates the precision from 80 to 100% (R^2 values 0.8 to 1.0). However, there are still some advanced computation approaches which have proven their competency in solving many problems of civil engineering. The literature studies also omitted the investigation of statistical analysis over the obtained results of the model. The deep insight view of literature studies reveals that the range of geotechnical parameters and quantity of dataset are limited. Using a large amount of dataset is always considered to be much worthwhile from generalization point of view [18, 19, 21–24].

1.1 Research Significance and Contributions

The main contribution of this study is to develop an efficient model for predicting one of the challenging real-world problems of highway engineers, i.e., estimation of soil California bearing ratio value. For this purpose, Kernel Ridge Regression (KRR), K-Nearest Neighbor (K-NN) and Gaussian Process Regression (GPR) algorithms were adopted. This study utilizes 1011 in situ samples of fine-grained plastic soils with an extensive range of index and engineering properties. Numerous geotechnical parameters were extracted from the laboratory experiments conducted through Bureau of Indian Standards (BIS) specifications. Two data divisional approaches, viz. K-fold and FCM, were adopted to investigate the influence of the training data features on the predictive ability of the developed model. Furthermore, the influence of employed machine learning algorithms as well as the data division approaches was investigated on the predictive ability of the model. Lastly, the literature models were attempted to validate through the present study datasets.

2 Machine Learning Algorithms and Statistical Assessment Indices

The term “Machine Learning (ML)” is a subfield/type of Artificial Intelligence (AI) and is referred to as predictive analytics or predictive modeling. ML is the development of computer systems that can learn and adapt without following explicit instructions by using algorithms and statistical models to analyze and draw inferences from patterns in data. In the recent past, numerous ML algorithms have been adopted by several researchers for solving many significant engineering problems [29–49]. This section briefly introduces the most prominent ML algorithms used to develop the model for predicting the soaked CBR value of fine-grained plastic soils and several indices to measure their performances.

2.1 Applied ML Algorithms

2.1.1 Kernel Ridge Regression (KRR)

KRR is the nonlinear regression approach that is based on the “kernel trick” in which datasets are nonlinearly transformed into some high-dimensional (or even infinite-dimensional) feature space determined by the kernel functions satisfying Mercer’s theorem [50–52]. Consider a TR set of $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, where N represents the number of TR samples. X is a features matrix, $[x_1, x_2, \dots, x_N]$, of size $N \times d$ and Y is a $N \times 1$ vector, $[1, 2, \dots, m]$, class labels.

KRR algorithm is generally based on the ridge regression and Ordinary Least Squares (OLS), a type of Linear Least Square (LLS), method [51, 53, 54]. The OLS minimizes the squared loss function:

$$\min_{\beta} \|Y - X\beta\|^2 \quad (1)$$

where $\|\cdot\|$ indicates the L_2 norm. In order to control the trade-off between bias and variance of the estimate, a shrinkage or ridge parameter λ is added to the above expression which is represented below:

$$\min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \quad (2)$$

Using the “kernel trick,” the KRR extends the linear regression into nonlinear and high-dimensional space. The data x_i in X is replaced with the feature vectors: $x_i \rightarrow \phi = \phi(x_i)$ induced by the kernel where $K_{ij} = k(x_i, x_j) = \phi(x_i) \phi(x_j)$. Therefore, the predicted class label of a new example x is now represented as:

$$Y^T (K + \lambda I)^{-1} k \quad (3)$$

where $k = (k_1, k_2, \dots, k_N)^T$, $k_N = x_N \cdot x$ and $n = 1, 2, \dots, N$.

In KRR, the kernel function is used to increase the computational power by mapping the data into a high-dimensional feature space which makes the data linear separable, and consequently, increases the stability, accuracy and generalization for both regression and classification problems. In general, the kernel functions are categorized into local and global functions. The local function influences the kernel function values within the vicinity like the Gaussian kernel function, whereas the global function allows the data points which are far from each other like the polynomial kernel function. The other kernel functions can also be employed, such as RBF, sigmoid and Laplacian.

2.1.2 K-Nearest Neighbor (K-NN)

K-NN is a nonparametric supervised ML algorithm that uses the k -number of most similar outputs from the TR dataset [55–58]. The K-NN algorithm is also known as a lazy learner algorithm, because, instead of immediate prediction, it stores the dataset and categorizes it based on its similarities and then approaches an action on the dataset. K-NN predicts the new records/data for the regression and classification-type problems based on their Euclidean distances, estimated mean, median or model output variable [59–62].

A classic K-NN algorithm for a regression problem can follow the below-given steps [63]:

Step 1: Consider a vector X with m independent variables, in this study $X = \{S, FC, PL, PI, MDD \text{ and } OMC\}$, as predictors and Y dependent variable, i.e., $Y = \{CBR\}$.

Step 2: Consider a TR set consisting of a dataset with t vectors of $X_n = \{S_{1n}, FC_{1n}, PL_{1n}, PI_{1n}, MDD_{1n} \text{ and } OMC_{1n}\}$ and a dependent variable of $Y_n = \{CBR_{1n}\}$ associated with each vector in X_n .

Step 3: The distance of the predictor vector with each of the n training vectors is calculated. The Euclidean distance (E_d) is most commonly used to estimate (see Eq. 4) which of the k output in the TR dataset is similar to the new input. However, other distances such as the Manhattan distance, which computes the distance between real vectors using the sum of absolute distance, Hamming distance, estimating the distance between binary vectors, and Minkowski distance, a generalization of the Euclidean and Manhattan distances [58, 64] could also be adopted according to the availability of the datasets.

$$E_d(x, y) = \sqrt{(X_1 - X_{1n})^2 + (X_2 - X_{2n})^2 + \dots + (X_m - X_{mn})^2} \tag{4}$$

Step 4: Select the k training vector with the least distance to the predictor vector.

Step 5: Calculate the kernel function (using an equation) for each of the k -selected training vectors

$$f_k(\Delta_k) = \frac{1}{\sum_{k=1}^K \frac{1}{\Delta_k}} \tag{5}$$

Step 6: The predicted dependent variable is calculated as

$$Y = \sum_{k=1}^K f_k(\Delta_k) \times Y_{nk} \tag{6}$$

where Y_{nk} is the dependent variable of each selected neighbor.

2.1.3 Gaussian Process Regression (GPR)

GPR is a probabilistic, nonparametric Bayesian approach for generalizing the nonlinear and complex problems related to regression and classification-type data sets [65, 66]. Many supervised ML algorithms learn exact values from the dataset, whereas GPR infers a probability distribution of all admissible functions that could reasonably fit the data space regarding the problems [67]. GPR is very efficient to handle nonlinear data due to the use of kernel functions. A GPR model can make predictions by incorporating the prior knowledge through covariance (kernel) functions and provide uncertainty measures over predictions [66, 68–70]. The algorithm has recently received huge attention from researchers as of having ability to solve many complex engineering problems related to various disciplines [66, 67, 69, 71–74].

For a given TR data set of $S = \{(x_i, y_i) | i = 1, 2, \dots, n\}$, the input data $X \in R^{S \times n}$ is called the designed matrix and $Y \in R^n$ is the vector of the desired output. The main assumption of GPR is that the output can be estimated as follows:

$$y = f(x) + \varepsilon \tag{7}$$

where ε signifies the noise term. GPR assumes that ε follows the Gaussian distribution with a mean of 0 and the variance of σ_n^2 :

$$\varepsilon \sim N(0, \sigma_n^2) \tag{8}$$

In GPR methodology, the n observations in the data set of interest $y = \{y_1, y_2, \dots, y_n\}$ are considered as a single point sampled from a multivariate Gaussian distribution. Moreover, it can be assumed that this Gaussian distribution has a mean of zeros. The covariance function $K(x, x')$ dictates the relation of one observation to another observation.

For a given TR data set, the ultimate goal of the learning process is to predict the output value y_* of a new queried input pattern. To achieve such a goal, it is necessary to establish three covariance matrices as follows:

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix} \quad (9)$$

$$K_* = [K(x_*, x_1) \ K(x_*, x_2) \ \dots \ K(x_*, x_n)] \quad (10)$$

$$K_{**} = [K(x_*, x_*)] \quad (11)$$

Due to the assumption that the data is sampled from multivariate Gaussian distribution, we have the following expression:

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \quad (12)$$

Since, $(y_*|y)$ is proved to be generated from multivariate Gaussian distribution the mean and the variance of the estimated mean and variance of the predicted output are given as:

$$\mu_* = K_* K^{-1} y \quad (13)$$

$$\sigma_*^2 = K_{**} - K_* K^{-1} K_*^T \quad (14)$$

2.1.4 Hyperparameters Tuning Using a Grid Search

Several ML algorithms contain a set of parameters that control many aspects of the algorithm. These parameters are termed as hyperparameters that are fixed before the learning process of the algorithm. These hyperparameters are adjusted to enhance the quality of the model as well as the ability to correctly predict the unseen dataset. In this study, hyperparameters of the above-used algorithms are optimized through a cross-validation (CV)-based grid search method. In the grid search method, the dataset is divided into training sets and test sets using K-Fold cross-validation (CV) [75]. For CV calculation, the grid points are assigned. Out of a total five number of folds, onefold is served as an independent testing set and the remaining k-1 (four) folds are designated as training sets. The grid search method is significantly superior to the random search method where only a few combinations are searched. The most significant advantage of this approach is that it provides better generalization performance for the respective model.

Table 2 Combination of KRR algorithm hyperparameters

Hyperparameters	Selected values
Alpha	3
Kernel type	Polynomial
Coefficient of kernel	0.2
Kernel degree	3

Table 3 Combination of K-NN algorithm hyperparameters

Hyperparameters	Selected values
n_neighbours	5
Weights	Uniform
Algorithm	Auto
Leaf size	30
P	2
Metric	Manhattan

Table 4 Combination of GPR algorithm hyperparameters

Hyperparameters	Selected values
Kernel	Rational quadratic () * Dot Product () + White Kernel ()
Alpha	$1e^{-10}$
Optimizer	Broyden–Fletcher–Goldfarb–Shanno

Initially, some manifestly unreasonable values of the hyperparameters were bypassed. After several trial-and-error approaches, the expected values for each of the hyperparameters were passed through the grid search method. Based on the results of best fit, the values obtained for the selected hyperparameters of KRR, K-NN and GPR algorithms are given in Tables 2, 3 and 4.

2.2 Statistical Performance Measurement Indices

The precision of all the models was assessed through several statistical performance indicators. The widely used performance measurement indicators are coefficient of determination (R^2), adjusted R^2 (adj. R^2), coefficient of correlation (R), mean absolute error (MAE), mean absolute percentage error (MAPE), root-mean-square error (RMSE), variance accounted for (VAF), performance index (I_p), Willmott's index of agreement (IOA), index of scattering (IOS), a20-index and performance strength (S_p) [17, 76–81]. The mathematical expression for these indicators is given in Eqs. (15) to (26) along with their ranges. In order to select an efficient prediction model espousing many more statistical indicators might be useful for assessing the performance of predictive models in terms of error and trend point of view

[19]. Using the least amount of parameters might be challenging, especially when comparing the performance of two or more models in an instant. Often, the models show not much difference in their statistical indices value. In that particular situation considering many more statistical performance indices might be helpful in selecting the best-fitted model.

Parameters	Ideal value	
$R^2 = 1 - \frac{\sum_{i=1}^N (y_i(a) - y_i(p))^2}{\sum_{i=1}^N (y_i(a) - \bar{y}_i(a))^2}$	1	(15)
$Adj.R^2 = \left[1 - \frac{N-1}{N-P-1} (1 - R^2) \right]$	1	(16)
$R = \frac{\sum_{i=1}^N ((y_i(a) - \bar{y}_i(a))(y_i(p) - \bar{y}_i(p)))}{\sqrt{(\sum_{i=1}^N (y_i(a) - \bar{y}_i(a))^2)(\sum_{i=1}^N (y_i(p) - \bar{y}_i(p))^2)}}$	1	(17)
$MAE = \left[\frac{1}{N} \sum_{i=1}^N y_i(a) - y_i(p) \right]$	0	(18)
$MAPE(\%) = \left[\frac{1}{N} \sum_{i=1}^N \left \frac{y_i(p) - y_i(a)}{y_i(a)} \right \right] \times 100$	0	(19)
$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i(a) - y_i(p))^2}$	0	(20)
$VAF(\%) = \left[1 - \frac{Var(y_i(a) - y_i(p))}{Var(y_i(a))} \right] \times 100$	100	(21)
$I_p = Adj.R^2 + 0.01VAF - RMSE$	2	(22)
$IOA = 1 - \frac{\sum_{i=1}^N (y_i(a) - y_i(p))^2}{\sum_{i=1}^N (y_i(p) - \bar{y}_i(a) + y_i(a) - \bar{y}_i(a))^2}$	1	(23)
$IOS = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i(a) - y_i(p))^2}}{\bar{y}_i(p)}$	0	(24)
$a20 - ndex = \frac{n20}{N} \times 100$	1	(25)
$S_p = \frac{(Adj.R^2)_{total} + (0.01VAF)_{total} - (RMSE)_{total}}{\left(\frac{Adj.R^2}{R^2} \right)_{training} + \left(\frac{Adj.R^2}{R^2} \right)_{testing}}$	1	(26)

where $y_i(a)$ = actual value (laboratory-obtained value); $y_i(p)$ = predicted value (value obtained through the developed model); $\bar{y}_i(a)$ = mean of actual value; $\bar{y}_i(p)$ = mean of predicted value; N = number of observations; P = number of input parameters used to develop the model; $n20$ = number of observations lies within error range of $\pm 20\%$; $a20$ = percentage of observations having error $\leq 20\%$ (Fig. 1).

2.2.1 Data Preparation and Analysis

2.3 Data Collection and Geographical Location

The Ministry of Road Transport and Highways (MoRTH) has decided to implement the Engineering Procurement and Construction (EPC) mode for constructing the National Highways (NH) in India. The Twelfth 5-Year Plan envisions the construction of 20,000 km of 4-lane National Highways projects through EPC mode in the various state of India. In this study, the Varanasi Gorakhpur section of NH-29 belongs

to the state of Uttar Pradesh (UP), India, was selected as the study area [42]. The geographical location of the study area is shown in Fig. 2.

2.4 Laboratory Experiments

From the aforementioned project worksite, a total of 1011 soil samples were collected and brought to the laboratory for experimental investigations. The laboratory tests were conducted as per BIS specifications. The test method includes IS 2720 (Part 4) [82], IS 2720 (Part 5) [83], IS 2720 (Part 8) [84] and IS 2720 (Part 16) [85] for the grain size distribution, Atterberg limits, modified Proctor compaction parameters and soaked CBR of fine-grained soil, respectively. Through these laboratory tests, numerous geotechnical parameters were collected such as gravel content (G), sand content (S),

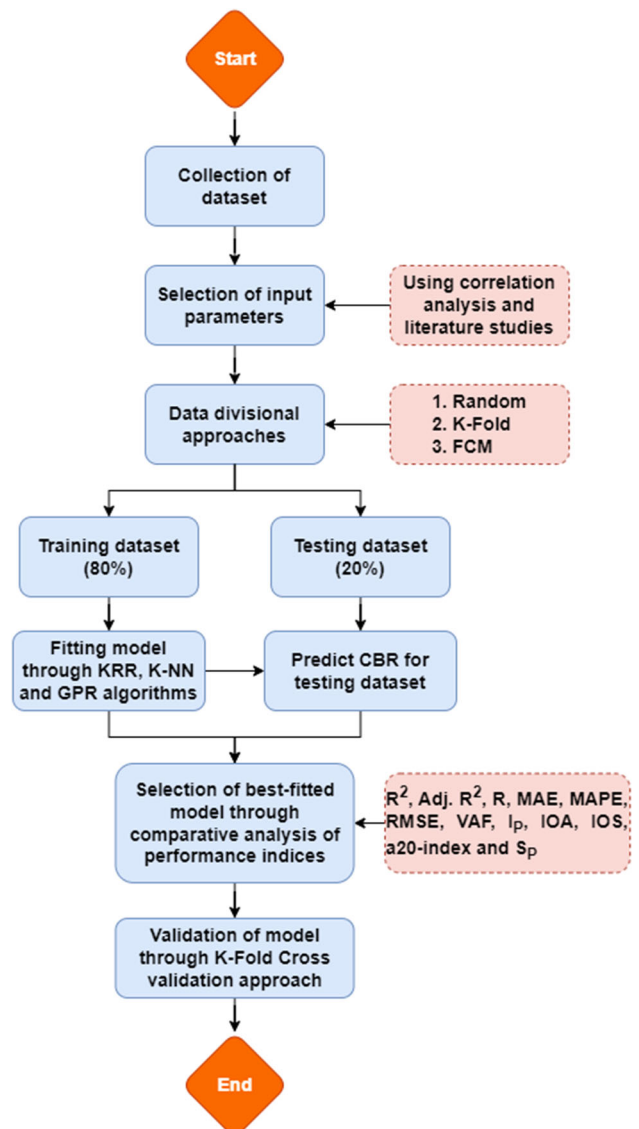


Fig. 1 Flowchart for predicting the soaked CBR of fine-grained plastic soils

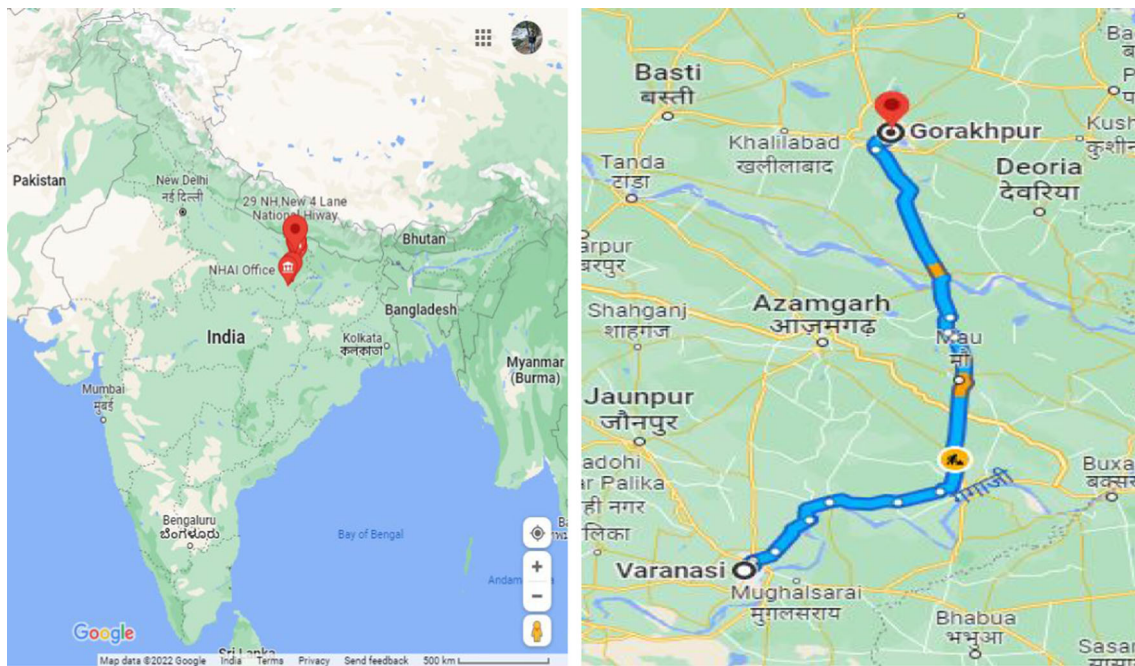


Fig. 2 Geographical location of the study area

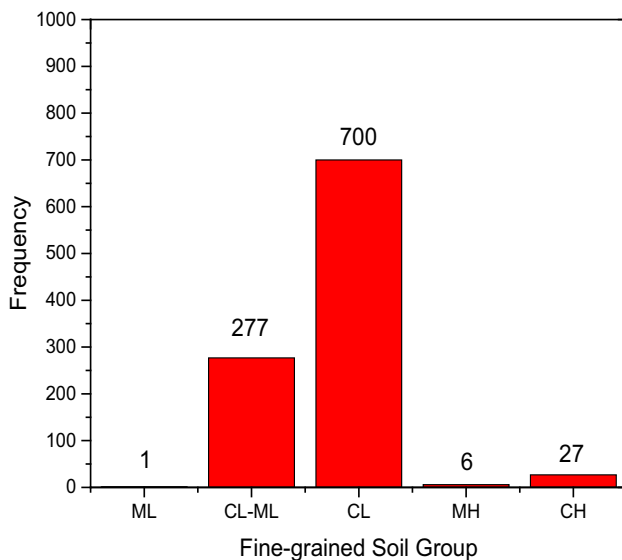


Fig. 3 Histogram plot for the different soil groups of fine-grained soil

silt and clay content termed as fine content (FC), liquid limit (LL), plastic limit (PL), plasticity index (PI), maximum dry density (MDD), optimum moisture content (OMC) and CBR value of fine-grained soil. The laboratory-obtained database of fine-grained soil was further classified into various soil groups using the USCS soil classification system. Figure 3 presents the histogram plot for different soil groups of fine-grained soil. The vertical column represents the amount of particular soil groups present in the fine-grained soil database.

2.5 Statistical Visualization and Correlation Analysis

Table 5 presents the descriptive statistic values for all the fine-grained soil parameters. As seen from Table 5, the obtained database covers an extensive range of CBR values from 1.0 to 13.20. The gravel content varies from 0 to 28%, sand ranges between 2 and 49%, and fine content exists in the range of 50 to 96%. Similarly, from a plasticity point of view, the selected database shows the liquid limit from 24 to 85%, plastic limit from 11 to 50% and plasticity index from 1 to 39%. MDD and OMC of fine-grained soil range from 1.455 g/cc to 1.959 g/cc and 9% to 30%, respectively.

Pearson's correlation (R) is one of the commonly used measures of association between the parameters. The range of R varies from -1 to 1 , where ± 1 indicates the strong association between the parameters, and a value of 0 (zero) illustrates no relationship between the parameters. The positive or negative sign specifies the respective increase or decrease in the associated parameters simultaneously. The correlation matrix obtained for all the geotechnical parameters is presented in Fig. 4. As observed from Fig. 4, CBR is positively correlated with S and MDD, whereas negatively with FC, LL, PL, PI and OMC. The final selected input parameters for developing the CBR prediction model are S , FC, PL, PI, MDD and OMC.

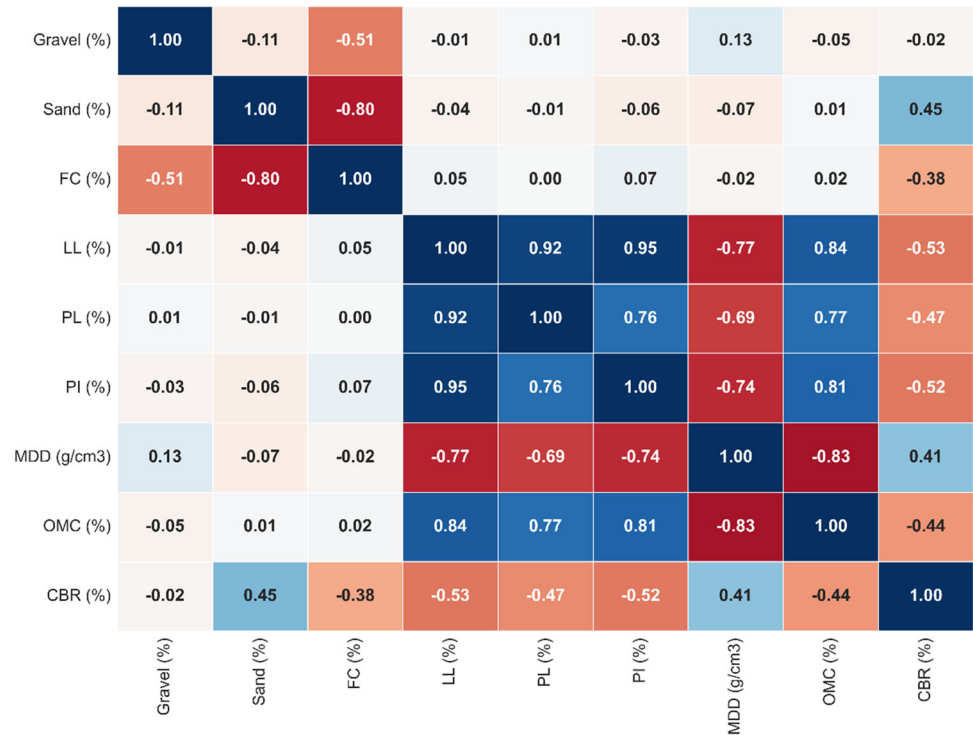
2.6 Data Divisional Approaches

Data division is the process of separating the complete dataset into the TR and TS subsets. In this study, about 80% of

Table 5 Descriptive statistic details for all fine-grained soil parameters

	G (%)	S (%)	FC (%)	LL (%)	PL (%)	PI (%)	MDD (g/cc)	OMC (%)	CBR (%)
Min	0.00	2.25	50.65	24.40	11.81	1.93	1.455	9.50	1.00
Max	27.42	48.85	96.28	85.00	50.00	39.00	1.959	29.50	13.20
Mean	2.83	13.99	83.18	29.85	21.60	8.25	1.866	11.96	9.02
Median	1.23	12.62	85.46	28.70	21.10	7.65	1.885	11.45	9.10
Mode	0.00	12.79	87.00	29.00	21.30	7.75	1.900	10.70	10.00
S. D	4.64	6.64	7.68	6.21	2.89	3.72	0.073	2.52	1.16
Variance	21.54	44.05	59.00	38.59	8.34	13.83	0.005	6.36	1.35

Fig. 4 Correlation matrix for all the geotechnical parameters



the entire dataset was considered to train the model and the remaining 20% was kept for testing the model. The basic problem with machine learning modeling is that we are unknown to the fact that how well a model performs or will perform until it is tested on the unseen dataset. One can build a perfect model on the TR data with 100% accuracy or 0 error, but it may fail to generalize for unseen data. It is not a good model as it over-fits the TR data. Machine learning is all about generalization, meaning that the model’s performance can only be measured with data points that have never been used during the training process. To overcome this problem, two data divisional approaches were adopted, discussed below:

2.6.1 K-Fold Division Approach

Initially, shuffle the dataset randomly and split it into K-number of folds (as seen in Fig. 5). Once the dataset is separated, the first fold is used as the TS dataset and the remaining k-1 folds are used for training purposes. The model with specific hyperparameters is trained with TR data (K-1 folds) and TS data as one fold. The performance of the model is recorded. The above steps are repeated until each k-fold got used for testing purposes.

Using fivefold, the complete datasets were divided into TR and TS sets. The descriptive statistics value obtained for the TR and TS datasets is shown in Tables 6 and 7, respectively.

Fig. 5 Data splitting process in K-fold cross-validation approach

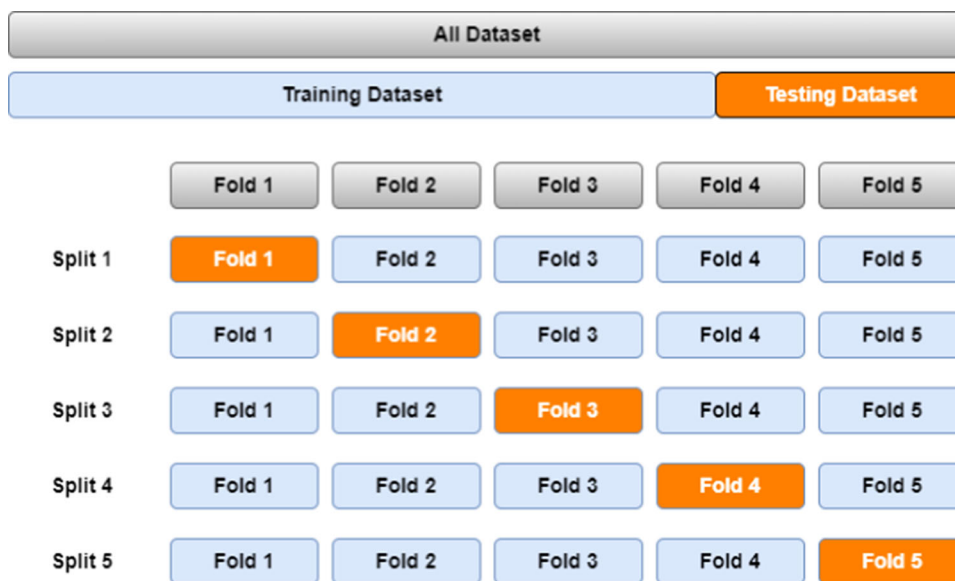


Table 6 Descriptive statistics value for the selected input and output parameters of the TR dataset obtained through the K-fold approach

	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Sand (%)	2.25	48.85	46.60	14.11	12.72	12.79	6.62	43.80
FC (%)	50.65	96.28	45.63	83.06	85.44	77.00	7.71	59.39
PL (%)	11.81	44.00	32.20	21.60	21.10	21.30	2.75	7.58
PI (%)	1.93	39.00	37.07	8.28	7.66	7.85	3.80	14.46
MDD (g/cc)	1.480	1.959	0.479	1.867	1.885	1.900	0.073	0.005
OMC (%)	9.50	29.50	20.00	11.98	11.45	10.70	2.53	6.38
CBR (%)	3.00	13.00	10.00	9.03	9.19	10.00	1.13	1.28

Table 7 Descriptive statistics value for the selected input and output parameters of TS dataset obtained through the K-fold approach

	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Sand (%)	4.08	48.40	44.33	13.53	12.19	10.32	6.71	44.98
FC (%)	51.60	95.61	44.01	83.66	85.54	89.57	7.58	57.41
PL (%)	12.72	50.00	37.29	21.62	21.09	21.45	3.38	11.45
PI (%)	5.25	35.00	29.75	8.11	7.58	8.35	3.37	11.33
MDD (g/cc)	1.455	1.950	0.495	1.863	1.882	1.900	0.073	0.005
OMC (%)	9.50	29.30	19.80	11.91	11.43	11.70	2.52	6.33
CBR (%)	1.00	13.20	12.20	8.98	9.05	8.40	1.28	1.65

2.6.2 Fuzzy C-Means (FCM) Division Approach

Clustering is the process of separating the dataset into a number of groups. Each group represents the observations that are homogeneous to each other and the objects that are dissimilar to each other are clustered into different groups. In fuzzy clustering, each observation can belong to more than one cluster based on its membership value. The total membership value for any observation distributed over the entire

cluster is 1.0. Brief information about this approach is given in Shi [86], Shahin, Maier [87] and Das [88].

Initially, two clusters were taken and the silhouette value was estimated. The number of a cluster was increased gradually and silhouette values for each of the corresponding clusters were calculated. The silhouette score obtained corresponding to the number of clusters is depicted in Fig. 6. It can be clearly observed from Fig. 6 that the maximum silhouette score was obtained when two clusters were used

Fig. 6 Silhouette value obtained for each of the clusters

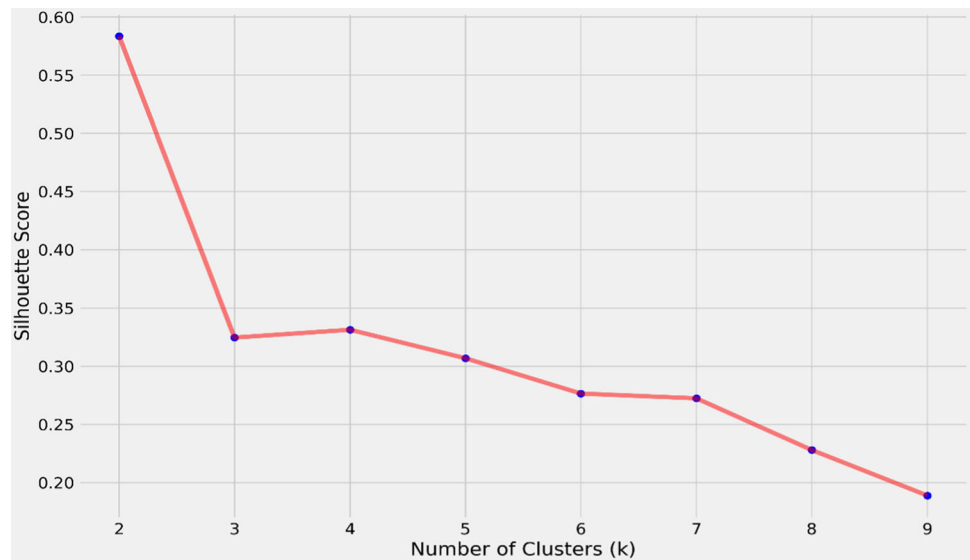


Table 8 Descriptive statistics value for the selected input and output parameters of the TR dataset obtained through the FCM approach

	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Sand (%)	4.01	48.85	44.85	13.96	12.46	11.00	6.76	45.67
FC (%)	50.65	95.61	44.96	83.16	85.45	87.00	7.77	60.38
PL (%)	11.81	50.00	38.20	21.59	21.10	21.30	2.87	8.26
PI (%)	1.93	39.00	37.07	8.29	7.65	7.90	3.86	14.90
MDD (g/cc)	1.455	1.959	0.504	1.867	1.885	1.900	0.073	0.005
OMC (%)	9.50	29.50	20.00	11.98	11.46	11.70	2.58	6.66
CBR (%)	1.00	13.20	12.20	9.01	9.10	10.00	1.18	1.40

Table 9 Descriptive statistics value for the selected input and output parameters of TS dataset obtained through the FCM approach

	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Sand (%)	2.25	44.76	42.51	14.14	12.96	12.96	6.14	37.67
FC (%)	51.70	96.28	44.58	83.27	85.49	89.57	7.33	53.68
PL (%)	18.79	42.00	23.21	21.66	21.06	20.95	2.95	8.70
PI (%)	5.20	29.00	23.80	8.08	7.57	6.50	3.08	9.50
MDD (g/cc)	1.490	1.943	0.453	1.863	1.883	1.900	0.075	0.006
OMC (%)	9.80	29.30	19.50	11.91	11.38	11.05	2.28	5.21
CBR (%)	3.00	10.90	7.90	9.05	9.20	9.60	1.09	1.18

for the analysis. The dataset obtained in the first and second clusters is represented as C_1 and C_2 , respectively. The C_1 dataset was separated into TR and TS sets through the K-fold approach (discussed in Sect. 3.4.1) and labeled as Train₁ and Test₁, respectively. Similarly, C_2 TR and TS dataset was designated as Train₂ and Test₂, respectively. The final TR dataset was obtained by concatenating the Train₁ and Train₂ datasets. Similarly, the TS dataset was achieved by concatenating the Test₁ and Test₂ datasets. The descriptive statistics value of the final TR and TS dataset is given in Tables 8 and 9.

3 Results

3.1 Statistical Performance of Developed Models

Table 10 illustrates the performance of the trained KRR, K-NN and GPR models in terms of several performance measurement indices. In Table 10, KRR_K and KRR_F indicate the KRR model developed through K-fold and FCM approaches, respectively. Similarly, models developed for K-NN and GPR algorithms are represented as $K-NN_K$, $K-NN_F$ and GPR_K , GPR_F , respectively. It is observed from Table 10 that the R^2 values obtained for the KRR algorithm in

Table 10 Statistical performance of KRR, K-NN and GPR model for TR dataset in K-Fold and FCM division approach

	KRR _K	KRR _F	K-NN _K	K-NN _F	GPR _K	GPR _F
R^2	0.647	0.681	0.715	0.727	0.746	0.887
Adj. R^2	0.644	0.679	0.713	0.725	0.744	0.886
R	0.804	0.825	0.847	0.854	0.866	0.943
MAE	0.515	0.512	0.432	0.437	0.438	0.308
MAPE	5.913	5.915	5.181	5.594	5.045	3.478
RMSE	0.672	0.666	0.604	0.617	0.570	0.397
VAF	64.705	68.131	71.530	72.743	74.551	88.690
I_P	0.620	0.694	0.824	0.835	0.919	1.376
IOA	0.882	0.897	0.906	0.911	0.917	0.968
IOS	0.074	0.074	0.067	0.068	0.063	0.044
a20-index	0.979	0.978	0.977	0.975	0.985	1.000

K-Fold and FCM approaches are 0.647 and 0.681, respectively. This means that KRR_K and KRR_F models can explain 64.7%, and 68.1% variability in the soaked CBR value of fine-grained plastic soils. Similarly, the variability explained by K-NN_K and K-NN_F models are 71.5% and 72.7%, respectively, and by GPR_K and GPR_F models are 74.6% and 88.7%, respectively. The MAE values obtained for KRR_K and KRR_F models are 0.515 and 0.512, for K-NN_K and K-NN_F models are 0.432 and 0.437, and for GPR_K and GPR_F models are 0.438 and 0.308, respectively. Results of the a20-index demonstrate that all the models can predict almost 98% of observations within $\pm 20\%$ variations.

The performance of developed models on the TS dataset is presented in Table 11. The R^2 value obtained in the TS dataset of the K-Fold approach is 0.680, 0.706 and 0.758 for KRR, K-NN and GPR models, respectively, whereas for KRR_F, K-NN_F and GPR_F models are 0.407, 0.645 and 0.700, respectively. It is clearly observed from the deep closest view of all the statistical parameters in training (refer to Table 10) and testing (refer to Table 11) dataset that the GPR algorithm achieved the highest prediction in both K-Fold and FCM approaches which is followed by K-NN and KRR algorithms.

3.2 Visual Interpretation of Developed Models

Visual interpretation facilitates the viewer to find the insight features from the model which is represented in a graphical form such as a scatter plot, error plot and regression error characteristics curve, etc.

3.2.1 Trend and Error Plot for the Developed Models

Figures 7 and 8 present the actual versus predicted CBR value of the TR and TS datasets, respectively, for the KRR, K-NN and GPR models at K-Fold and FCM approaches. It is observed from the scatter plot results that the maximum

number of observations follows a specific trend along with the line of equality. However, the closeness of data points toward the line of equality in the TR dataset (see Fig. 7) is maximum for GPR models, and an acceptable inclination is obtained for KRR and K-NN models. Moreover, the utmost precision in the GPR model is obtained when the dataset is trained through the FCM data division approach followed by the K-Fold approach. Similarly, the results obtained for the TS dataset (see Fig. 8) also reveal that GPR models are more efficient in predicting the unseen dataset, followed by K-NN and KRR algorithms.

The error distribution and bar plots for the TR and TS dataset of KRR, K-NN and GPR models at K-Fold and FCM data division approaches are shown in Figs. 9 and 11, respectively. The center horizontal line in the error distribution plot represents the zero error line, the datasets existing on that line have zero error, i.e., the difference in the actual and predicted CBR value is zero. At the same time, the upper and lower line specifies the +20% and -20%, respectively, error or variation band. It is observed from Figs. 9 and 11 that some datasets are below the zero error line (displays negative error) and some are above the zero error line (shows positive error). This random pattern of error indicates that the developed models have a decent fit for the dataset. Furthermore, the existence of a dataset within $\pm 20\%$ variation seems to be the maximum for the GPR_F model (refer to Fig. 11). This can also be confirmed from the comparative analysis of Figs. 10 and 12, showing the error frequency plot for the TR and TS datasets of KRR, K-NN and GPR models at K-Fold and FCM approaches, respectively. As seen from Fig. 12, the GPR_F model can predict almost 96% and 90% observations of the TR and TS dataset, respectively, within $\pm 10\%$, whereas 100% and 99% within $\pm 20\%$ variations which are substantially higher than other models (Figs. 11, 12).



Table 11 Statistical performance of KRR, K-NN and GPR model for TS dataset in K-Fold and FCM division approach

	KRR _K	KRR _F	K-NN _K	K-NN _F	GPR _K	GPR _F
R^2	0.680	0.407	0.706	0.645	0.758	0.700
Adj. R^2	0.670	0.389	0.697	0.634	0.750	0.690
R	0.840	0.739	0.847	0.804	0.880	0.837
MAE	0.542	0.549	0.501	0.491	0.476	0.457
MAPE	7.642	6.964	7.612	5.750	6.605	5.310
RMSE	0.724	0.836	0.694	0.647	0.630	0.595
VAE	68.036	40.758	70.749	64.584	75.822	69.956
I_P	0.626	-0.040	0.710	0.634	0.879	0.795
IOA	0.915	0.851	0.899	0.883	0.918	0.904
IOS	0.081	0.093	0.077	0.071	0.070	0.066
a20-index	0.970	0.950	0.955	0.965	0.985	0.985

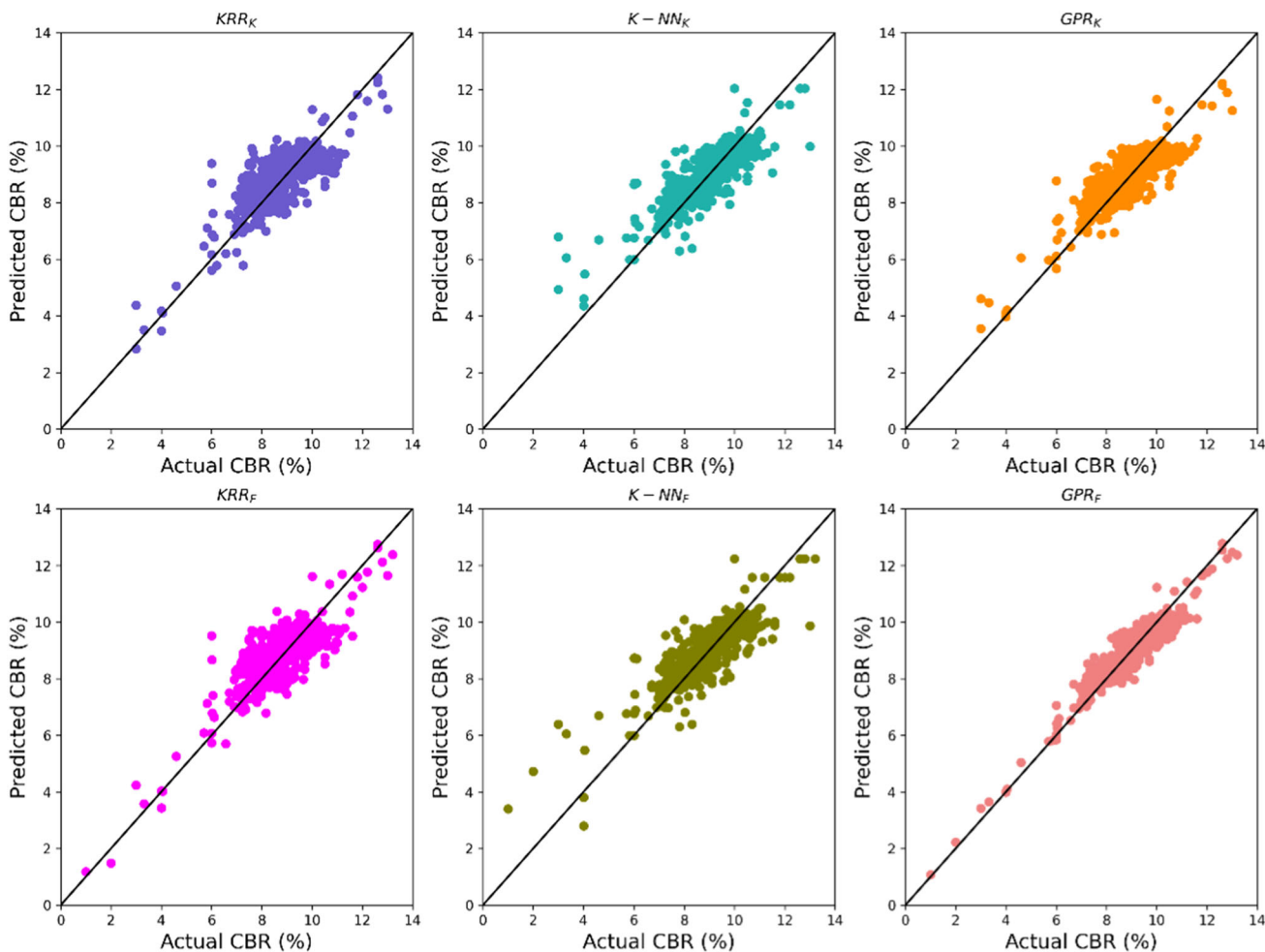


Fig. 7 Scatter plot for the TR dataset of KRR, K-NN and GPR models at K-Fold and FCM approaches

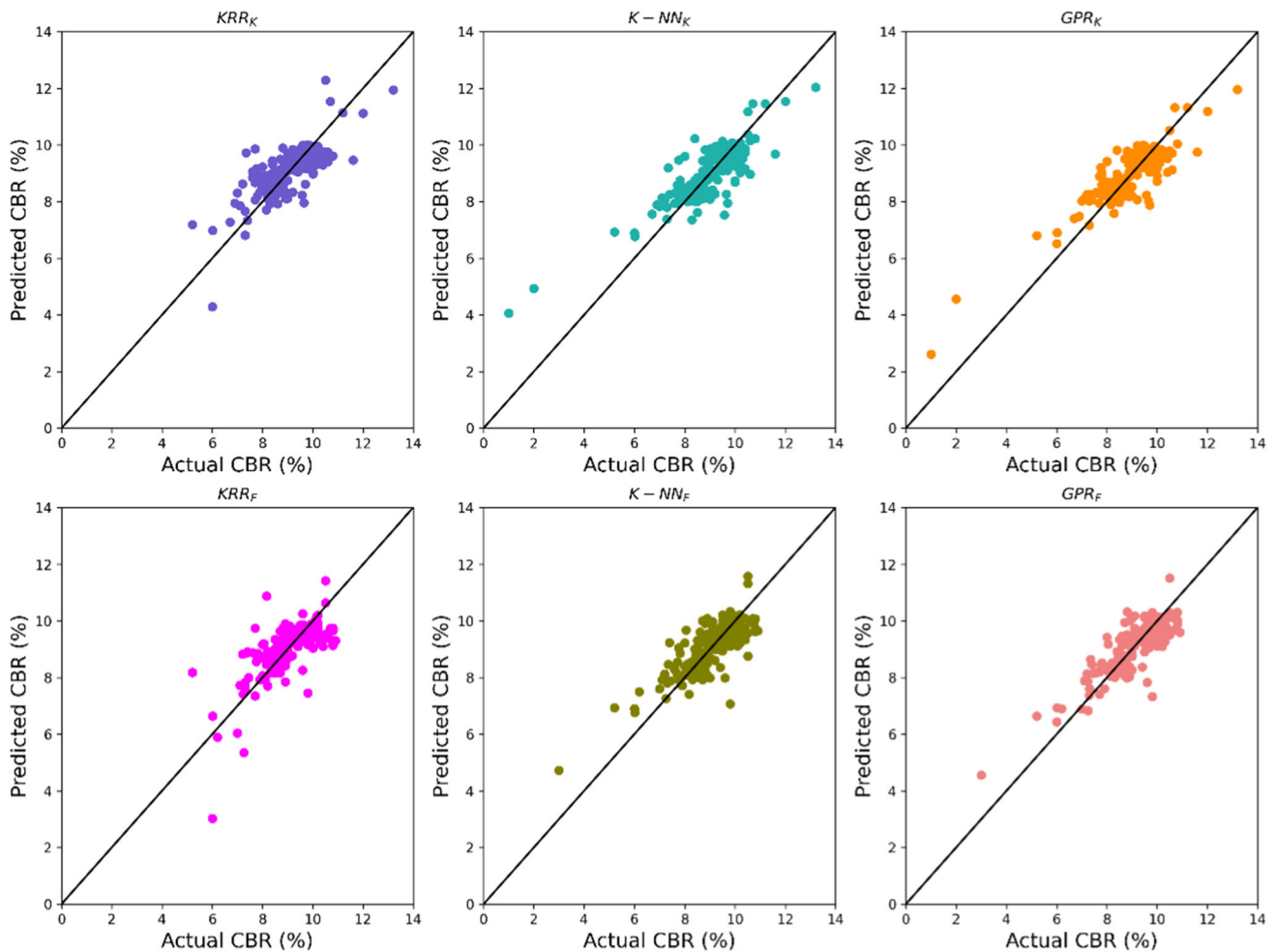


Fig. 8 Scatter plot for the TS dataset of KRR, K-NN and GPR models at K-Fold and FCM approaches

3.2.2 Regression Error Characteristics (REC) curve

In regression problems, REC curves are equivalents to the receiver operating characteristics (ROC) curves in classification problems. The X-axis of the REC curve plot demonstrates the error tolerance, whereas the Y-axis represents the accuracy in terms of the percentage of points predicted within the tolerance [79, 80, 89]. An ideal model's curve should pass through the upper left corner and therefore should have an area under the curve (AUC) value is 1. This means that the model can perfectly discriminate between all the positive and the negative class points. In general, an AUC of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is deemed to be excellent, and more than 0.9 is considered outstanding.

Figures 13, 14, 15 and 16 depict the REC curve obtained corresponding to the TR and TS dataset of KRR, K-NN and GPR models at K-Fold and FCM data division approaches. As seen from these figures, the AUC value obtained for all the models in TR and TS datasets is higher than 0.9, which

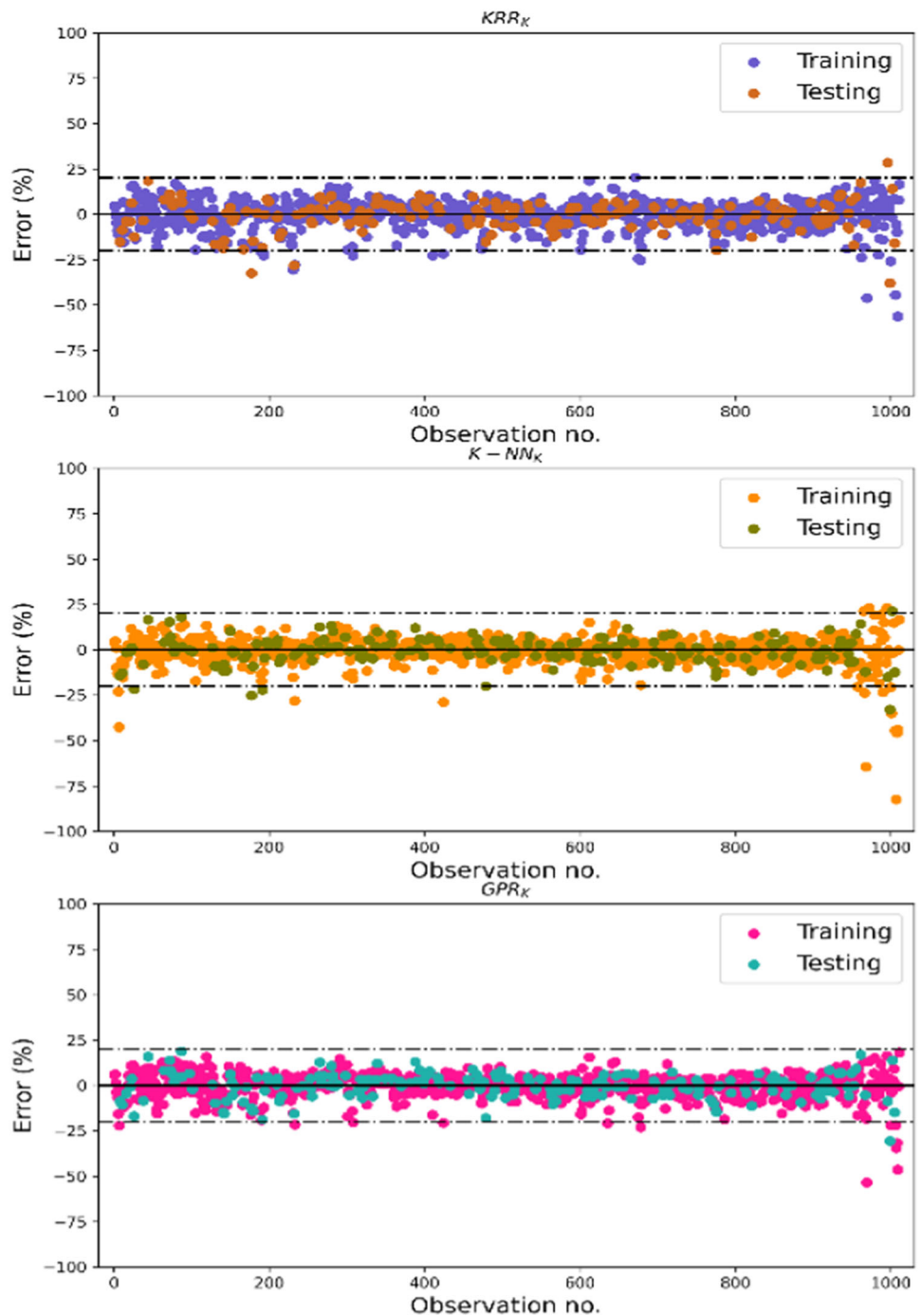
means the developed models outperform very well and are stated to be reliable in predicting the soaked CBR value of fine-grained plastic soils. Furthermore, it is observed from the comparative analysis of both models in the training and testing set that the REC curve for the GPR_F model exists more closely to the upper left corner compared to the other models as well as the AUC value achieved for the GPR_F model is also higher than those models.

3.2.3 Accuracy Analysis

The accuracy analysis is a novel assessment used to evaluate the efficiency of the models. The analysis demonstrates the accuracy (%) of a model, which is obtained through the comparative analysis of the values obtained for different performance measurement parameters to their ideal values (as shown in Sect. 2.2) using Eqs. (27) and (28).

$$A_e = |(1 - |m_e|)| \times 100 \quad (27)$$

Fig. 9 Error distribution plot for the TR and TS dataset of KRR, K-NN and GPR models at K-Fold approach



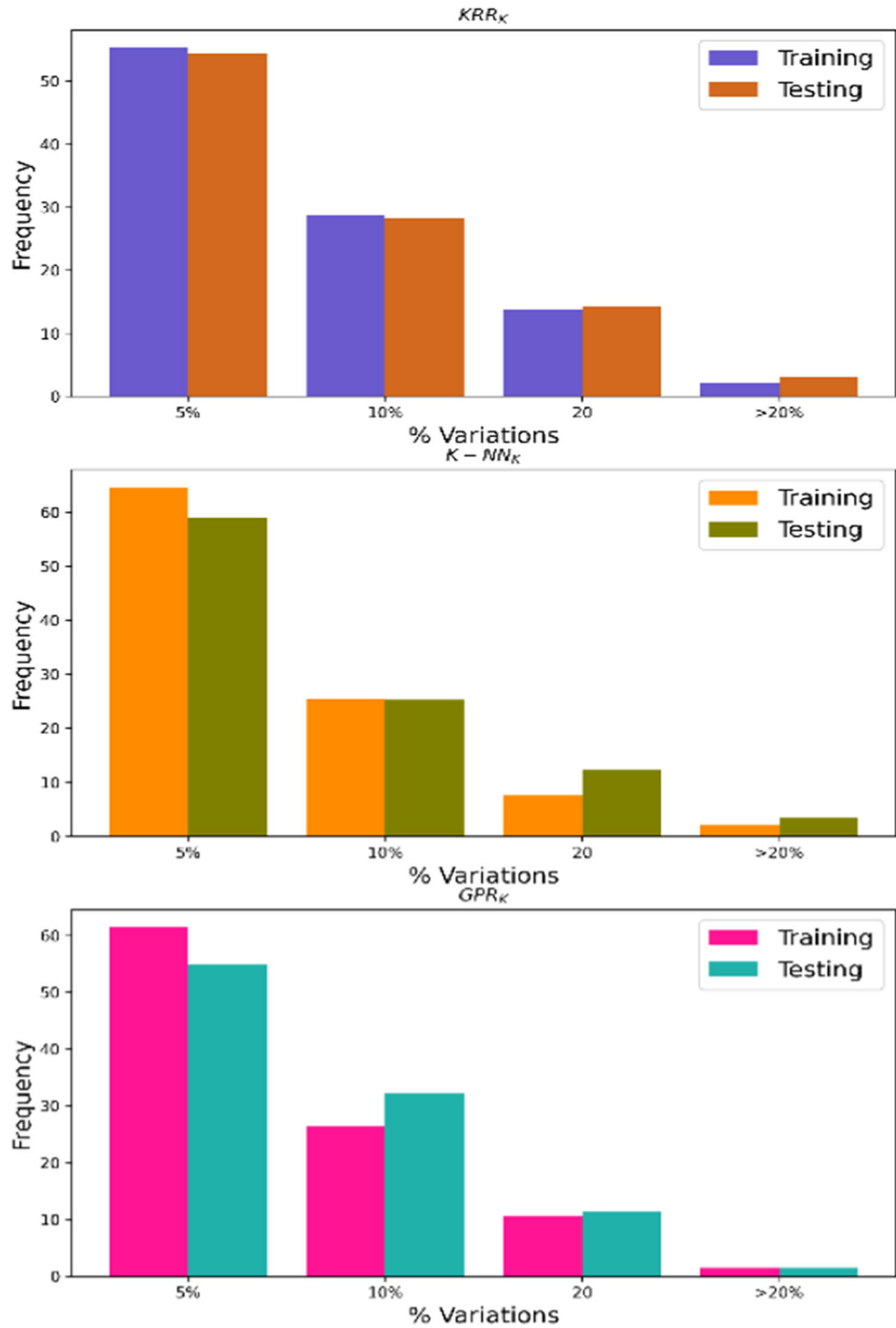
$$A_t = \frac{|m_t|}{i_t} \times 100 \tag{28}$$

where A_e and A_t denote the error and trend-measuring performance parameters. m_e and m_t indicates the measured values of the error and trend-measuring performance parameters. The performance measurement parameters MAE, MAPE, RMSE and IOS belong to the class of error, whereas R^2 , Adj. R^2 , R , VAF, I_p , IOA and a20-index belong to the

trend. i_t represents the ideal value of the respective trend parameters.

Tables 12 and 13 present the accuracy of all the developed models in the TR and TS dataset, respectively. As seen from Tables 12 and 13, the accuracy of the GPR_F model throughout numerous statistical performance measurement indices is significantly higher than other models.

Fig. 10 Error frequency plot for the TR and TS dataset of KRR, K-NN and GPR models at K-Fold approach

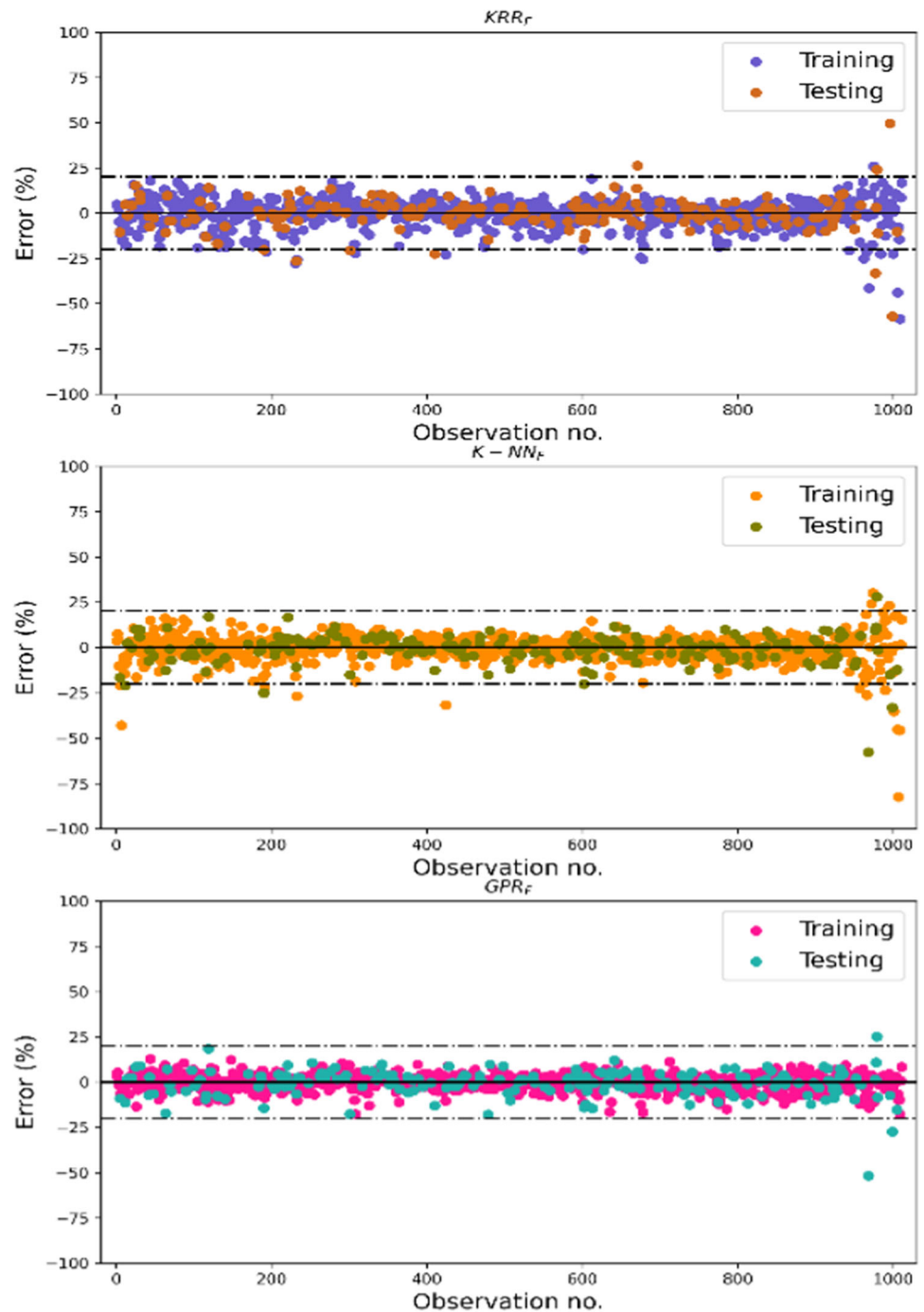


3.3 Selection of Best-Fitted CBR Prediction Model

Numerous statistical performance indices were adopted to evaluate the model performances and the conclusions can easily be drawn by comparing their values. However, the situation becomes more complicated when the model demonstrates adequate accuracy in the TR dataset and failed to reach

a good amount of accuracy in the TS dataset as well. Moreover, it also becomes complicated when the value of different statistical performance indices describes their own best models. In that particular situation, the overfitting analysis and ranking analysis might be useful for identifying the best-fitted model as they envisage the overall analysis of all the parameters of a particular model.

Fig. 11 Error distribution plot for the TR and TS dataset of KRR, K-NN and GPR models at FCM approach



3.3.1 Ranking Analysis (RA)

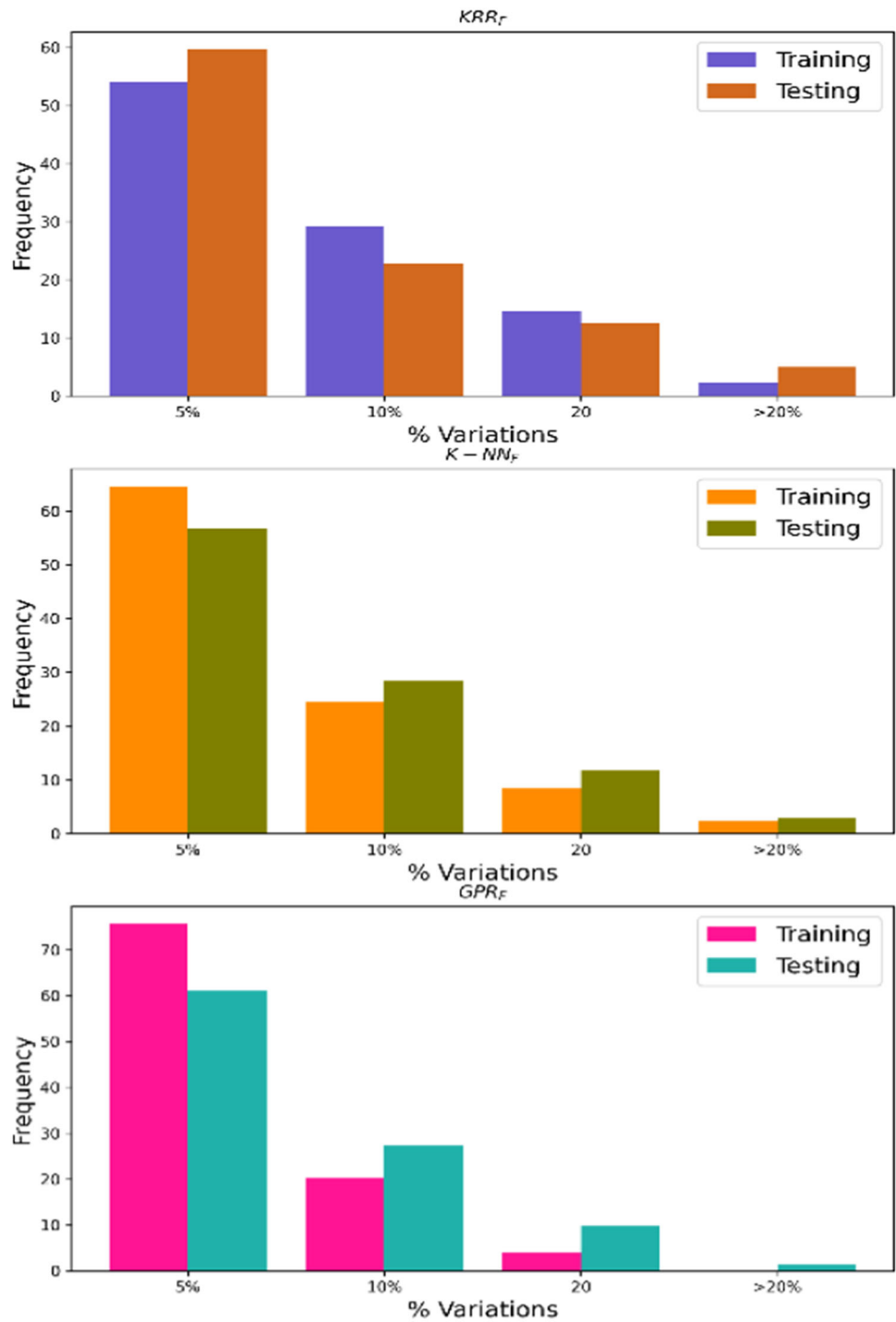
According to RA, used by many researchers in the past [79, 80, 89, 90], a maximum score of s (equal to the total number of corresponding models) is assigned to the model having the highest value in particular performance indices, minimum to the model with the lowest value and the score to the other intermediate models are assigned either in the ascending or descending order. Table 14 presents the RA results obtained

for the KRR, K-NN and GPR models at K-Fold and FCM approaches.

3.3.2 Overfitting Ratio (OR)

OR is the ratio of the RMSE of the TS dataset to the TR dataset as shown in Eq. (29). According to this formula, a model with a lesser OR value or close to one is considered

Fig. 12 Error frequency plot for the TR and TS dataset of KRR, K-NN and GPR models at FCM approach



less prone to overfitting, which implicitly enhances the generalization capacity of the models and provides the maximum realistic approach for the real-world applications [16]. The obtained OR value for all developed models is shown in Table 14.

$$OR = \frac{RMSE_{of\ TS\ dataset}}{RMSE_{of\ TR\ dataset}} \tag{29}$$

Initially, a score was assigned to each developed model corresponding to their statistical performance measurement parameters, as shown in Table 14. The total score for a particular model was achieved by combining the score obtained in its TR and TS dataset. The overall accuracy of the developed models was determined through the performance strength parameters (using Eq. (26)). Based on the S_p value obtained for each model, a score was assigned to them as per the

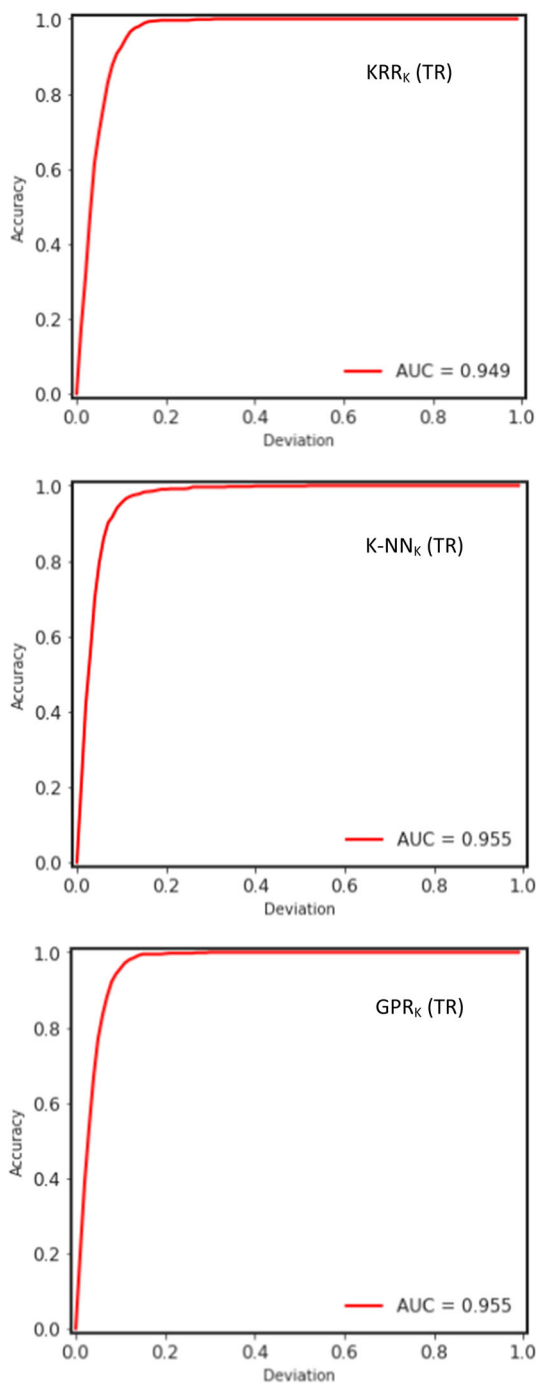


Fig. 13 REC curve for the TR dataset of KRR, K-NN and GPR models at K-Fold approach

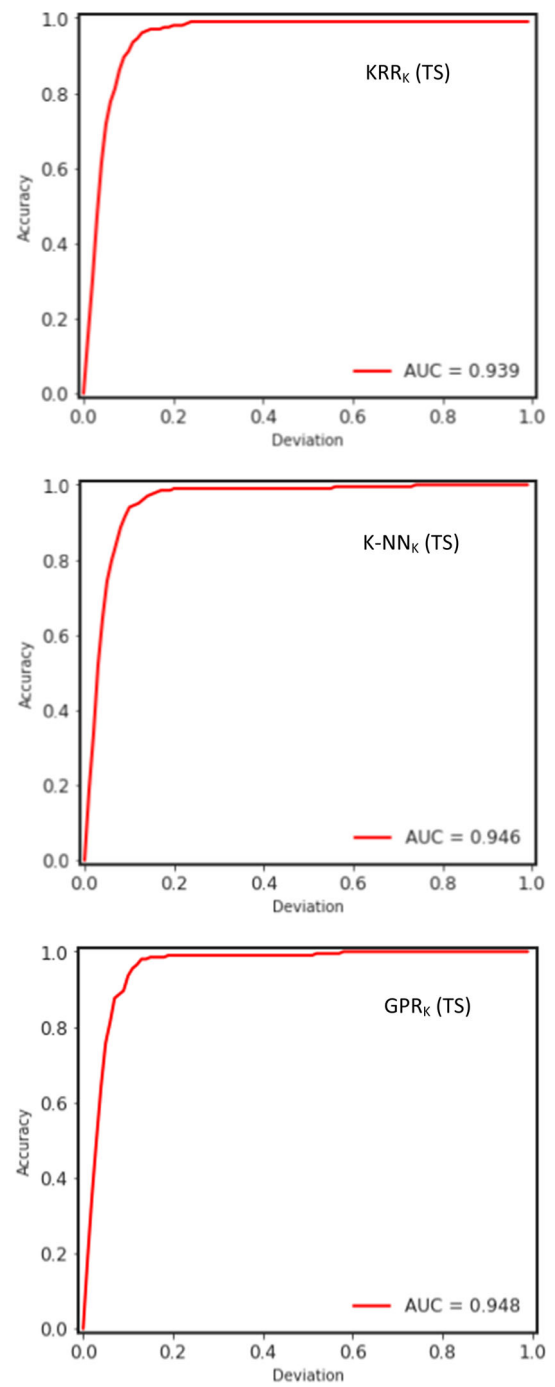


Fig. 14 REC curve for the TS dataset of KRR, K-NN and GPR models at K-Fold approach

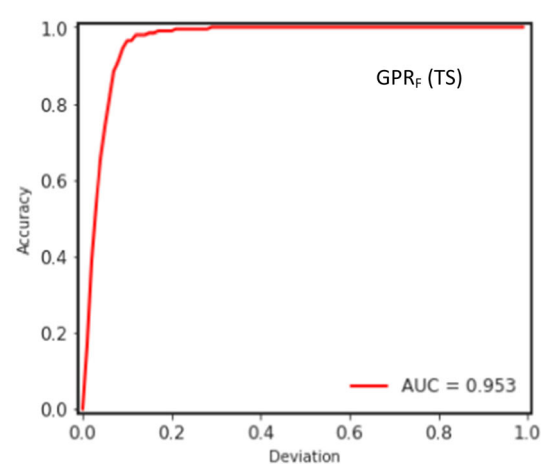
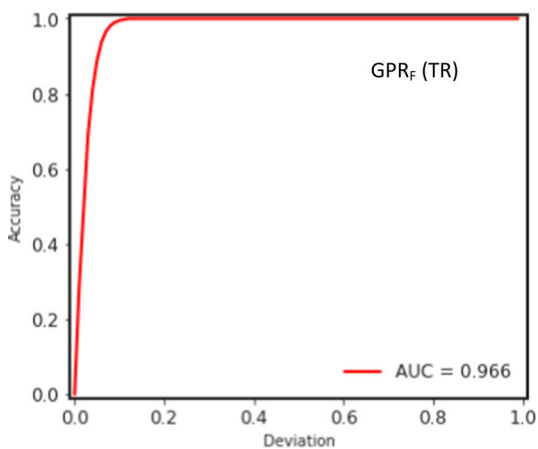
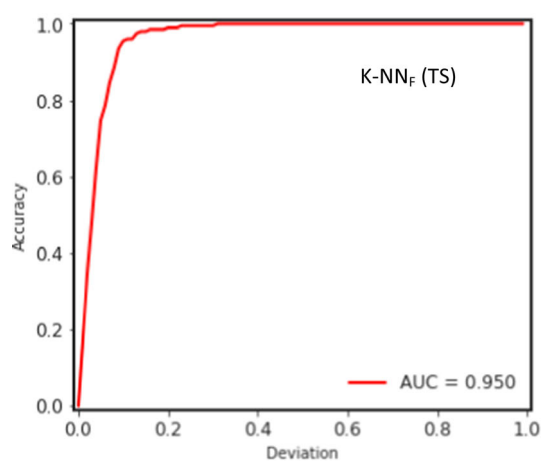
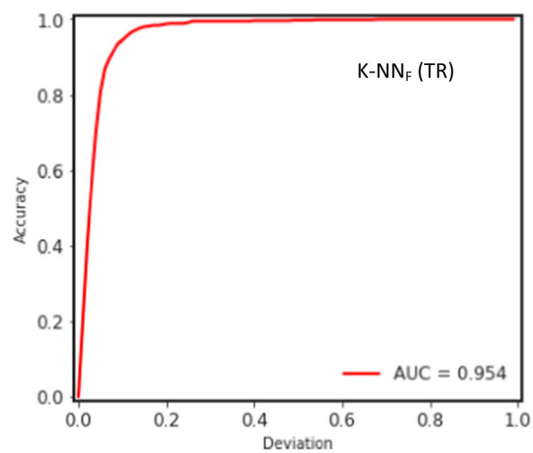
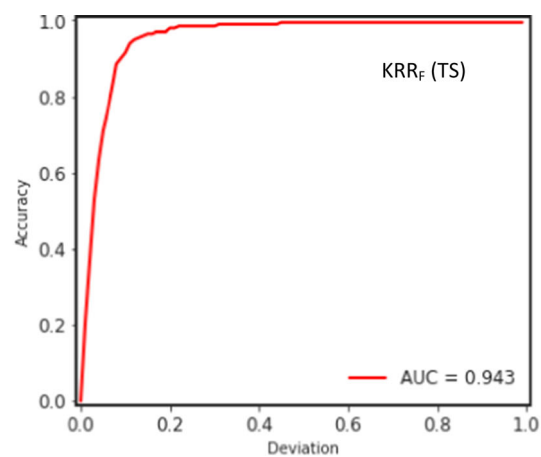
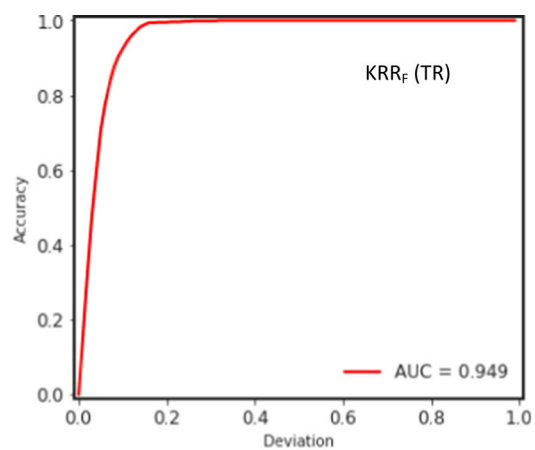


Fig. 15 REC curve for the TR dataset of KRR, K-NN and GPR models at FCM approach

Fig. 16 REC curve for the TS dataset of KRR, K-NN and GPR models at FCM approach

ranking analysis method. The final score of the model was accomplished by concatenating its total score value and the score obtained for the S_p value. Accordingly, the rank was assigned to the models with respect to their final scoring value. As seen from Table 14, the maximum score was achieved by the GPR_F model, therefore, designated as the first rank, followed by GPR_K , $K-NN_K$, $K-NN_F$, KRR_K , and KRR_F models. The ranked models were passed through overfitting analysis. It is clearly observed from the results of the overfitting analysis, shown in Table 14, that the GPR_F model exhibits a little amount of overfitting as the OR value is considerably higher than 1. Again, the analysis was performed on a second-ranked model, i.e., the GPR_K model. The OR value obtained for the GPR_K model is almost close to one, which means that the model is more generalized to the unseen dataset. However, the GPR_F model could also be used as a predictive model as the R -value obtained in the TS dataset is 0.83. According to Smith [91], Taskiran [11], Yildirim and Gunaydin [12], Verma [92] and Tenpe and Patel [17], if the R -value of a model is ≥ 0.8 , then it is established that the predicted and actual values are good in agreement and strongly correlated. Conclusively, GPR_F and GPR_K models were finally selected as the best-fitted model for predicting the soaked CBR value of fine-grained plastic soils.

3.4 Influence of ML Algorithms and Data Division Approaches on the Model Performance

The overall influence of adopted ML algorithms and data divisional approaches were identified using the final score value (refer to Table 14) obtained for each model. Figures 17 and 18 present the overall influence of ML algorithms and data division approaches, respectively, on the predictive ability of the developed models. It is perceived from Fig. 17 that the GPR algorithm proves its proficiency more than K-NN and KRR algorithms as the combined final score value is significantly higher than those. The combined final score value obtained for the K-Fold approach is slightly more than the FCM approach. However, the K-Fold approach has been found substantially useful in removing the overfitting of the models.

3.5 Validation of Literature Study Models

For this purpose, only those models were selected which were having input parameters similar to the present study's geotechnical parameters. Kin [10], Taskiran [11], Yildirim and Gunaydin [12] and Bardhan, Gokceoglu [19] models, belonging to various countries (as shown in Table 15), were

attempted to validate through the present study datasets. Datasets from the present study were selected as per their minimum and maximum input and output parameter values.

Table 16 depicts the comparative performance of literature study models on the present study datasets. It is observed from Table 16 that the R^2 value obtained for all the models is negative, which means that the selected models don't follow the specific trend of the dataset, leading to a worse fit than the horizontal line. The Bardhan, Gokceoglu [19] model can predict more than 65% observations within $\pm 20\%$ variations, which is relatively higher than Kin [10], Taskiran [11] and Yildirim and Gunaydin [12] models. This is because the datasets used for the model development and validation are geological and almost nearby, i.e., India. Nagaraj and Suresh [93] also state that soils are likely to be quite variable depending on their geological locations.

4 Discussion of Results

In the above subsections, the prediction of the soaked CBR of fine-grained plastic soils was assessed through some machine learning algorithms. The collected dataset is comprised of various groups of soils such as ML, CL-ML, CL, MH and CH. Pearson's correlation analysis reveals that S, FC, PL, PI, MDD and OMC were the substantial input parameters in predicting the soaked CBR value of fine-grained plastic soils. Using these input parameters, six models for GPR, K-NN and KRR algorithms were developed through K-Fold and FCM data divisional approaches. The quantitative performance of these models was identified using twelve statistical measurement parameters. Based on the ranking analysis and overfitting ratio, GPR_K and GPR_F models were considered the best-fitted model, followed by $K-NN_K$, $K-NN_F$, KRR_K , and KRR_F models. The OR value obtained for GPR_K , $K-NN_K$ and KRR_K models was less than the OR value obtained for GPR_F , $K-NN_F$ and KRR_F models (as shown in Table 14). This means that the RMSE value in the TR and TS datasets of GPR_K , $K-NN_K$ and KRR_K models are almost close to each other. Therefore, the K-Fold approach was considered to be more significant in removing the overfitting of the models. Furthermore, the comparative results of ML algorithms exhibit the maximum proficiency of the GPR algorithm, followed by K-NN and KRR algorithms. The validation results of the literature models on the present study dataset ascertain that the prediction ability of any model in the field of geotechnical/highway engineering is significantly influenced by the geological location of the dataset.

Table 12 Accuracy of all the developed models for the TR dataset

Statistical performance measurement parameters	Models accuracy (%)					
	KRR _K	K-NN _K	GPR _K	KRR _F	K-NN _F	GPR _F
R^2	64.7	71.5	74.6	68.1	72.7	88.7
Adj. R^2	64.4	71.3	74.4	67.9	72.5	88.6
R	80.4	84.7	86.6	82.5	85.4	94.3
MAE	48.5	56.8	56.2	48.8	56.3	69.2
MAPE	94.1	94.8	95.0	94.1	94.4	96.5
RMSE	32.8	39.6	43.0	33.4	38.3	60.3
VAF	64.7	71.5	74.6	68.1	72.7	88.7
I_P	31.0	41.2	46.0	34.7	41.8	68.8
IOA	88.2	90.6	91.7	89.7	91.1	96.8
IOS	92.6	93.3	93.7	92.6	93.2	95.6
a20-index	97.9	97.7	98.5	97.8	97.5	100.0

Table 13 Accuracy of all the developed models for the TS dataset

Statistical performance measurement parameters	Models accuracy (%)					
	KRR _K	K-NN _K	GPR _K	KRR _F	K-NN _F	GPR _F
R^2	68.0	70.6	75.8	40.7	64.5	70.0
Adj. R^2	67.0	69.7	75.0	38.9	63.4	69.0
R	84.0	84.7	88.0	73.9	80.4	83.7
MAE	45.8	49.9	52.4	45.1	50.9	54.3
MAPE	92.4	92.4	93.4	93.0	94.3	94.7
RMSE	27.6	30.6	37.0	16.4	35.3	40.5
VAF	68.0	70.8	75.8	40.8	64.6	70.0
I_P	31.3	35.5	44.0	2.0	31.7	39.8
IOA	91.5	89.9	91.8	85.1	88.3	90.4
IOS	91.9	92.3	93.0	90.7	92.9	93.4
a20-index	97.0	95.5	98.5	95.0	96.5	98.5

Table 14 Rank analysis and overfitting ratio for all the developed models

	R ²	Adj. R ²	R	MAE	MAPE	RMSE	VAF	I _p	IOA	IOS	a20-index	Score	Total score	Sp value	Score for Sp value	Final score	Rank	OR
KRR _K	TR	1	1	1	2	1	1	1	1	1	4	15	46	0.316	2	48	5	1.077
	TS	3	3	4	1	2	3	2	5	2	4	31						
KRR _F	TR	2	2	2	1	2	2	2	2	2	3	22	35	0.404	3	38	6	1.149
	TS	1	1	1	3	1	1	1	1	1	1	13						
K-NN _K	TR	3	3	3	4	4	3	3	3	4	2	37	77	0.459	5	82	3	1.105
	TS	5	5	5	2	3	5	4	3	3	2	40						
K-NN _F	TR	4	4	4	3	3	4	4	4	3	1	38	71	0.288	1	72	4	1.255
	TS	2	2	2	4	4	2	3	2	4	3	33						
GPR _K	TR	5	5	5	5	5	5	5	5	5	5	53	113	0.405	4	117	2	1.049
	TS	6	6	6	4	5	6	6	6	5	5	60						
GPR _F	TR	6	6	6	6	6	6	6	6	6	6	66	120	0.637	6	126	1	1.499
	TS	4	4	3	6	6	4	5	4	6	6	54						

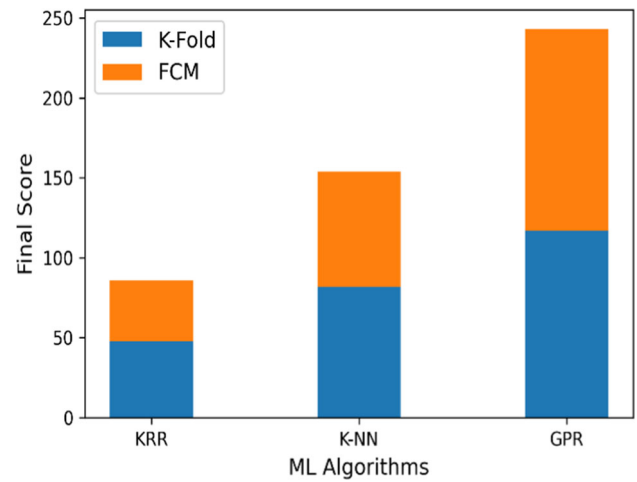


Fig. 17 Influence of ML algorithms on the model performance

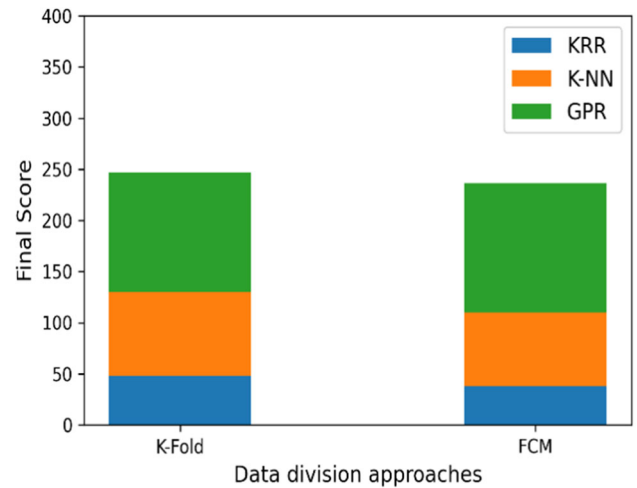


Fig. 18 Influence of data division approaches on the model performance

Table 15 Literature models validation on the present study dataset

S. No.	Literature model	Soil origin	Dataset from the present study
1	Kin [10]	Malaysia	997
2	Taskiran [11]	Turkey	1011
3	Yildirim and Gunaydin [12]	Turkey	1011
4	Bardhan, Gokceoglu [19]	Indian	610

5 Conclusions

The current study offers novel applications of KRR, K-NN and GPR algorithms in predicting the soaked CBR value of fine-grained plastic soils. The analysis was performed on in situ soil samples collected from an ongoing NHAI project

Table 16 Comparative performance of literature models on the present study datasets

Statistical performance indices	Literature models				
	Kin [10]	Taskiran [11]	Yildirim and Gunaydin [12]	Bardhan, Gokceoglu [19] MARS-L	Bardhan, Gokceoglu [19] GP
R^2	– 14.58	– 99.38	– 5.41	– 5.09	– 4.79
Adj. R^2	– 14.58	– 99.38	– 5.41	– 5.09	– 4.79
R	0.20	0.52	0.11	0.00	0.10
MAE	3.84	9.66	2.54	1.51	1.56
MAPE	42.42	106.08	31.83	15.91	16.40
RMSE	4.35	11.65	2.94	1.75	1.71
VAF	– 284.14	– 3035.08	– 64.88	– 94.49	– 8.79
I_p	– 21.76	– 141.37	– 9.00	– 7.79	– 6.59
IOA	0.30	0.16	0.35	0.36	0.39
IOS	0.82	0.62	0.25	0.22	0.22
a20-index	0.16	0.01	0.35	0.65	0.67

work site. Large datasets of 1011 soil samples were collected and laboratory tests were performed as per BIS specifications. The prepared datasets were divided into TR and TS sets using the K-Fold and FCM data divisional approaches. The competency of models was compared through various statistical performance measurement indices, accuracy analysis and REC curve analysis. Apart from them, the final selection of the best-fitted model was accomplished through ranking analysis and overfitting analysis. The obtained results of statistical performance indices reveal that the developed models can explain a maximum variability of 88.7% and a minimum of 64.7% in the CBR value of the TR dataset through S, FC, PL, PI, MDD and OMC as input parameters. The performance of models in the TS dataset was considerably less than that obtained in the TR dataset. Based on the overall analysis of the results, the final selected best-fitted models for predicting the soaked CBR value of fine-grained plastic soils were GPR_F and GPR_K models. The proposed GPR_F and GPR_K models can predict 99% and 98% observations, respectively, within $\pm 20\%$ variations. Additionally, the comparative analysis of the adopted algorithm demonstrates that the GPR algorithm gives the highest proficiency compared to the K-NN and KRR algorithms. Results also indicate that K-Fold and FCM approaches are suitable methods for data division. Eventually, the validation results establish that the predictive ability of any model is substantially influenced by the geological location of the soils/materials used for model development.

Acknowledgements The authors are immensely thankful to the National Highway Authority of India (NHAI), India, as well as Jaypee Constructions for providing the laboratory facilities during the research

work. The corresponding author is also thankful to the Ministry of Education, India, for providing the supporting fund under the grant of a PhD fellowship.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Code and Data Availability Developed code and data are available upon reasonable request to the corresponding author of the manuscript.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Davis, E.: The California bearing ratio method for the design of flexible roads and runways. *Géotechnique* **1**(4), 249–263 (1949)



2. Sreelekshmypillai, G.; Vinod, P.: Prediction of CBR value of fine grained soils at any rational compactive effort. *Int. J. Geotech. Eng.*, p. 1–6 (2017)
3. Black, W.: The calculation of laboratory and in-situ values of California bearing ratio from bearing capacity data. *Geotechnique* **11**(1), 14–21 (1961)
4. Stephens, D.: The prediction of the California bearing ratio. *Civil Eng. Siviele Ingenieurswese* **32**(12), 523–528 (1990)
5. Bayamack, J.F.N.; Onana, V.L.; Mvindi, A.T.N.; Ze, A.N.O.; Ohandja, H.N.; Eko, R.M.: Assessment of the determination of Californian Bearing Ratio of laterites with contrasted geotechnical properties from simple physical parameters. *Transp. Geotech.* **19**, 84–95 (2019)
6. Kleyn, S.: Possible developments in pavement foundation design. *Civil Eng. Siviele Ingenieurswese* **5**(9), 286–292 (1955)
7. Black, W.: A method of estimating the California bearing ratio of cohesive soils from plasticity data. *Geotechnique* **12**(4), 271–282 (1962)
8. Agarwal, K.; Ghanekar, K.: Prediction of CBR from plasticity characteristics of soil. In: *Proceeding of 2nd South-east Asian Conference on Soil Engineering*, Singapore (1970)
9. National Cooperative Highway Research Program, N.: Guide for mechanistic and empirical-design for new and rehabilitated pavement structures, final document. In: *Appendix CC-1: Correlation of CBR values with soil index properties*. **2001**, (2001)
10. Kin, M.: California bearing ratio correlation with soil index properties, p. 2006. University technology, Malaysia, Master of engineering project (2006)
11. Taskiran, T.: Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. *Adv. Eng. Softw.* **41**(6), 886–892 (2010)
12. Yildirim, B.; Gunaydin, O.: Estimation of California bearing ratio by using soft computing systems. *Expert Syst. Appl.* **38**(5), 6381–6391 (2011)
13. Erzin, Y.; Turkoz, D.: Use of neural networks for the prediction of the CBR value of some Aegean sands. *Neural Comput. Appl.* **27**(5), 1415–1426 (2016)
14. 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.* **36**(6), 3485–3498 (2018)
15. 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.* **44**(10), 8691–8705 (2019)
16. Tenpe, A.R.; Patel, A.: Application of genetic expression programming and artificial neural network for prediction of CBR. *Road Mater. Pavement Des.* **21**(5), 1183–1200 (2018)
17. Tenpe, A.R.; Patel, A.: Utilization of support vector models and gene expression programming for soil strength modeling. *Arab. J. Sci. Eng.* **45**(5), 4301–4319 (2020)
18. Bardhan, A.; Samui, P.; Ghosh, K.; Gandomi, A.H.; Bhattacharyya, S.: ELM-based adaptive neuro swarm intelligence techniques for predicting the California bearing ratio of soils in soaked conditions. *Appl. Soft Comput.* **110**, 107595 (2021)
19. 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.* **291**, 106239 (2021)
20. 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.*, p. 1–13 (2021)
21. Karimpour-Fard, M.; Machado, S.L.; Falamaki, A.; Carvalho, M.F.; Tizpa, P.: Prediction of compaction characteristics of soils from index test's results. *Iran. J. Sci. Technol. Trans. Civil Eng.* **43**(1), 231–248 (2019)
22. 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* **134**(7), 326 (2019)
23. Wang, H.L.; Yin, Z.Y.: High performance prediction of soil compaction parameters using multi expression programming. *Eng. Geol.* **276**, 105758 (2020)
24. Zou, W.-L.; Han, Z.; Ding, L.-Q.; Wang, X.-Q.: Predicting resilient modulus of compacted subgrade soils under influences of freeze–thaw cycles and moisture using gene expression programming and artificial neural network approaches. *Transp. Geotech.* **28**, 100520 (2021)
25. Alawi, M.; Rajab, M.: Prediction of California bearing ratio of subbase layer using multiple linear regression models. *Road Mater. Pavement Des.* **14**(1), 211–219 (2013)
26. Varghese, V.K.; Babu, S.S.; Bijukumar, 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.* **31**(4), 1187–1205 (2013)
27. 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.* **37**(1), 217–234 (2019)
28. 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* **95**(2), 190–196 (2020)
29. Verma, G.; Kumar, B.: Prediction of compaction parameters for fine-grained and coarse-grained soils: a review. *Int. J. Geotech. Eng.*, p. 1–8 (2019)
30. Ray, A.; Kumar, V.; Kumar, A.; Rai, R.; Khandelwal, M.; Singh, T.: Stability prediction of Himalayan residual soil slope using artificial neural network. *Nat. Hazards* **103**, 3523–3540 (2020)
31. Cuong-Le, T.; Nghia-Nguyen, T.; Khatir, S.; Trong-Nguyen, P.; Mirjalili, S.; Nguyen, K.D.: An efficient approach for damage identification based on improved machine learning using PSO-SVM. *Eng. Comput. P.* 1–16 (2021)
32. Czarniecki, S.; Shariq, M.; Nikoo, M.; Sadowski, L.: An intelligent model for the prediction of the compressive strength of cementitious composites with ground granulated blast furnace slag based on ultrasonic pulse velocity measurements. *Measurement* **172**, 108951 (2021)
33. Trong, D.K.; Pham, B.T.; Jalal, F.E.; Iqbal, M.; Roussis, P.C.; Mamou, A.; Ferentinou, M.; Vu, D.Q.; Duc Dam, N.; Tran, Q.A.: On random subspace optimization-based hybrid computing models predicting the California bearing ratio of soils. *Materials* **14**(21), 6516 (2021)
34. Bharati, A.K.; Ray, A.; Khandelwal, M.; Rai, R.; Jaiswal, A.: Stability evaluation of dump slope using artificial neural network and multiple regression. *Eng. Comput.* **38**(Suppl 3), 1835–1843 (2022)
35. Cakiroglu, C.; Islam, K.; Bekdas, G.; Isikdag, U.; Mangalathu, S.: Explainable machine learning models for predicting the axial compression capacity of concrete filled steel tubular columns. *Constr. Build. Mater.* **356**, 129227 (2022)
36. Cuong-Le, T.; Minh, H.-L.; Sang-To, T.; Khatir, S.; Mirjalili, S.; Wahab, M.A.: A novel version of grey wolf optimizer based on a balance function and its application for hyperparameters optimization in deep neural network (DNN) for structural damage identification. *Eng. Fail. Anal.* **142**, 106829 (2022)
37. Ho, L.S.; Tran, V.Q.: Machine learning approach for predicting and evaluating California bearing ratio of stabilized soil containing industrial waste. *J. Clean. Prod.* **370**, 133587 (2022)
38. Karir, D.; Ray, A.; Bharati, A.K.; Chaturvedi, U.; Rai, R.; Khandelwal, M.: Stability prediction of a natural and man-made slope using various machine learning algorithms. *Transp. Geotech.* **34**, 100745 (2022)



39. Paliwal, M.; Goswami, H.; Ray, A.; Bharati, A.K.; Rai, R.; Khandelwal, M.: Stability prediction of residual soil and rock slope using artificial neural network. *Adv. Civil Eng.*, **2022** (2022)
40. Shamsabadi, E.A.; Roshan, N.; Hadigheh, S.A.; Nehdi, M.L.; Khodabakhshian, A.; Ghalehnovi, M.: Machine learning-based compressive strength modelling of concrete incorporating waste marble powder. *Constr. Build. Mater.* **324**, 126592 (2022)
41. Verma, G.; Kumar, B.: Artificial neural network equations for predicting the modified proctor compaction parameters of fine-grained soil. *Transp. Infrastruct. Geotechnol.*, p. 1–24 (2022)
42. Verma, G.; Kumar, B.: Application of multi-expression programming (MEP) in predicting the soaked California bearing ratio (CBR) value of fine-grained soil. *Innov. Infrastruct. Solut.* **7**(4), 1–16 (2022)
43. Zhang, W.; Gu, X.; Tang, L.; Yin, Y.; Liu, D.; Zhang, Y.: Application of machine learning, deep learning and optimization algorithms in geoenvironment and geoscience: Comprehensive review and future challenge. *Gondwana Research*, (2022)
44. Khatti, J.; Grover, K.S.: CBR prediction of pavement materials in unsoaked condition using LSSVM, LSTM-RNN, and ANN approaches. *Int. J. Pavement Res. Technol.* (2023). <https://doi.org/10.1007/s42947-022-00268-6>
45. Liu, S.; Wang, L.; Zhang, W.; He, Y.; Pijush, S.: A comprehensive review of machine learning-based methods in landslide susceptibility mapping. *Geol. J.* (2023). <https://doi.org/10.1002/gj.4666>
46. Nayak, D.K.; Verma, G.; Dimri, A.; Kumar, R.; Kumar, V.: Predicting the Twenty-eight day compressive strength of OPC-and PPC-prepared concrete through hybrid GA-XGB model. *Pract. Period. Struct. Des. Constr.* **28**(3), 04023020 (2023)
47. Nghia-Nguyen, T.; Kikumoto, M.; Nguyen-Xuan, H.; Khatir, S.; Wahab, M.A.; Cuong-Le, T.: Optimization of artificial neural networks architecture for predicting compression parameters using piezocene penetration test. *Expert Syst. Appl.* **223**, 119832 (2023)
48. 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.* **14**(7), 101988 (2023)
49. Zhang, W.; Gu, X.; Hong, L.; Han, L.; Wang, L.: Comprehensive review of machine learning in geotechnical reliability analysis: algorithms, applications and further challenges. *Appl. Soft Comput.*, p. 110066 (2023)
50. Safari, M.J.S.; Rahimzadeh Arashloo, S.: Kernel ridge regression model for sediment transport in open channel flow. *Neural Comput. Appl.* **33**(17), 11255–11271 (2021)
51. Saunders, C.; Gamerman, A.; Vovk, V.: Ridge regression learning algorithm in dual variables. (1998)
52. Naik, J.; Satapathy, P.; Dash, P.: Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression. *Appl. Soft Comput.* **70**, 1167–1188 (2018)
53. Zhang, S.; Hu, Q.; Xie, Z.; Mi, J.: Kernel ridge regression for general noise model with its application. *Neurocomputing* **149**, 836–846 (2015)
54. Rakesh, K.; Suganthan, P.N.: An ensemble of kernel ridge regression for multi-class classification. *Procedia Comput. Sci.* **108**, 375–383 (2017)
55. Peterson, L.E.: K-nearest neighbor. *Scholarpedia* **4**(2), 1883 (2009)
56. Amiri, M.; Bakhshandeh Amnieh, H.; Hasanipناه, M.; Mohammad Khanli, L.: A new combination of artificial neural network and K-nearest neighbors models to predict blast-induced ground vibration and air-overpressure. *Eng. Comput.* **32**(4), 631–644 (2016)
57. Chen, Y.; Hao, Y.: A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Syst. Appl.* **80**, 340–355 (2017)
58. Hsieh, S.-C.: Prediction of compressive strength of concrete and rock using an elementary instance-based learning algorithm. *Adv. Civil Eng.* **1–10**, 2021 (2021)
59. Yu, B.; Song, X.; Guan, F.; Yang, Z.; Yao, B.: k-Nearest neighbor model for multiple-time-step prediction of short-term traffic condition. *J. Transp. Eng.* **142**(6), 04016018 (2016)
60. Cheng, M.-Y.; Hoang, N.-D.: Slope collapse prediction using Bayesian framework with k-nearest neighbor density estimation: case study in Taiwan. *J. Comput. Civ. Eng.* **30**(1), 04014116 (2016)
61. Oh, S.; Byon, Y.-J.; Yeo, H.: Improvement of search strategy with k-nearest neighbors approach for traffic state prediction. *IEEE Trans. Intell. Transp. Syst.* **17**(4), 1146–1156 (2015)
62. Kang, M.-C.; Yoo, D.-Y.; Gupta, R.: Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete. *Constr. Build. Mater.* **266**, 121117 (2021)
63. Ebrahimi, E.; Shourian, M.: River flow prediction using dynamic method for selecting and prioritizing K-nearest neighbors based on data features. *J. Hydrol. Eng.* **25**(5), 04020010 (2020)
64. Inkoom, S.; Sobanjo, J.; Barbu, A.; Niu, X.: Pavement crack rating using machine learning frameworks: Partitioning, bootstrap forest, boosted trees, Naïve bayes, and K-Nearest neighbors. *J. Transp. Eng. Part B Pavements* **145**(3), 04019031 (2019)
65. Wang, J.: An intuitive tutorial to Gaussian processes regression. *arXiv preprint arXiv:2009.10862*, (2020)
66. Dutta, S.; Samui, P.; Kim, D.: Comparison of machine learning techniques to predict compressive strength of concrete. *Comput. Concr.* **21**(4), 463–470 (2018)
67. Ly, H.-B.; Nguyen, T.-A.; Pham, B.T.: Investigation on factors affecting early strength of high-performance concrete by Gaussian Process Regression. *PLoS ONE* **17**(1), e0262930 (2022)
68. Dao, D.V.; Adeli, H.; Ly, H.-B.; Le, L.M.; Le, V.M.; Le, T.-T.; Pham, B.T.: A sensitivity and robustness analysis of GPR and ANN for high-performance concrete compressive strength prediction using a Monte Carlo simulation. *Sustainability* **12**(3), 830 (2020)
69. Ghanizadeh, A.R.; Heidarabadzadeh, N.; Heravi, F.: Gaussian process regression (Gpr) for auto-estimation of resilient modulus of stabilized base materials. *J. Soft Comput. Civil Eng.* **5**(1), 80–94 (2021)
70. Williams, C.K.; Rasmussen, C.E.: *Gaussian processes for machine learning*, Vol. 2. MIT press Cambridge, MA (2006)
71. Cai, H.; Jia, X.; Feng, J.; Li, W.; Hsu, Y.-M.; Lee, J.: Gaussian Process regression for numerical wind speed prediction enhancement. *Renew. Energy* **146**, 2112–2123 (2020)
72. Ceylan, Z.: Estimation of municipal waste generation of Turkey using socio-economic indicators by Bayesian optimization tuned Gaussian process regression. *Waste Manage. Res.* **38**(8), 840–850 (2020)
73. Zeng, A.; Ho, H.; Yu, Y.: Prediction of building electricity usage using Gaussian process regression. *J. Build. Eng.* **28**, 101054 (2020)
74. García-Nieto, P.J.; García-Gonzalo, E.; Paredes-Sánchez, J.P.; Bernardo Sánchez, A.: A new hybrid model to foretell thermal power efficiency from energy performance certificates at residential dwellings applying a Gaussian process regression. *Neural Comput. Appl.* **33**(12), 6627–6640 (2021)
75. Goodfellow, I.; Bengio, Y.; Courville, A.: *Deep learning*. MIT press (2016)
76. Alzabeebee, S.: Application of EPR-MOGA in computing the liquefaction-induced settlement of a building subjected to seismic shake. *Eng. Comput.* p. 1–12 (2020)
77. Hanandeh, S.; Ardah, A.; Abu-Farsakh, M.: Using artificial neural network and genetics algorithm to estimate the resilient modulus for stabilized subgrade and propose new empirical formula. *Transp. Geotech.* **24**, 100358 (2020)
78. Alzabeebee, S.; Alshkane, Y.M.; Al-Taie, A.J.; Rashed, K.A.: Soft computing of the recompression index of fine-grained soils. *Soft Comput.* **25**, 15297–15312 (2021)

79. Kardani, N.; Bardhan, A.; Kim, D.; Samui, P.; Zhou, A.: Modelling the energy performance of residential buildings using advanced computational frameworks based on RVM, GMDH, ANFIS-BBO and ANFIS-IPSO. *J. Build. Eng.* **35**, 102105 (2021)
80. Kardani, N.; Bardhan, A. ; Samui, P. ; Nazem, M. ; Zhou, A. ; Armaghani, D.J.: A novel technique based on the improved firefly algorithm coupled with extreme learning machine (ELM-IFF) for predicting the thermal conductivity of soil. *Eng. Comput.*, p. 1–20 (2021)
81. Bardhan, A.; Kardani, N.; Alzouibi, A.; Roy, B.; Samui, P.; Gandomi, A.H.: Novel integration of extreme learning machine and improved Harris hawks optimization with particle swarm optimization-based mutation for predicting soil consolidation parameter. *J. Rock Mech. Geotech. Eng.* **14**(5), 1588–1608 (2022)
82. IS 2720 (Part 4): Methods of test for soils–Grain size analysis. 1985, Bureau of Indian Standards New Delhi, India
83. IS 2720 (Part 5): Determination of liquid limit and plastic limit (second revision) (1985)
84. IS 2720 (Part 8): Determination of water content, dry density relation using heavy compaction (second revision) (1994)
85. IS 2720 (Part 16): Laboratory determination of CBR (second revision) (1987)
86. Shi, J.J.: Clustering technique for evaluating and validating neural network performance. *J. Comput. Civ. Eng.* **16**(2), 152–155 (2002)
87. Shahin, M.A.; Maier, H.R.; Jaksa, M.B.: Data division for developing neural networks applied to geotechnical engineering. *J. Comput. Civ. Eng.* **18**(2), 105–114 (2004)
88. Das, S.K.: Application of genetic algorithm and artificial neural network to some geotechnical engineering problems. Ph.D Thesis, IIT Kanpur (India) (2005)
89. Asteris, P.G.; Skentou, A.D.; Bardhan, A.; Samui, P.; Pilakoutas, K.: Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cem. Concr. Res.* **145**, 106449 (2021)
90. Zhang, H.; Zhou, J.; Jahed Armaghani, D.; Tahir, M.; Pham, B.T.; Huynh, V.V.: A combination of feature selection and random forest techniques to solve a problem related to blast-induced ground vibration. *Appl. Sci.* **10**(3), 869 (2020)
91. Smith, G.N.: Probability and statistics in civil engineering. Collins professional and technical books, **244** (1986)
92. Verma, J.: Data analysis in management with SPSS software. Springer Science & Business Media (2012)
93. Nagaraj, H.; Suresh, M.: Influence of clay mineralogy on the relationship of CBR of fine-grained soils with their index and engineering properties. *Transp. Geotech.* **15**, 29–38 (2018)

