CHAPTER 4 DEVELOPMENT OF CALIFORNIA BEARING RATIO PREDICTION MODEL

4.1 GENERAL

The current chapter is particularly focused on developing the prediction model for the soaked CBR of fine-grained plastic soils. Various ML techniques, which have been discussed in detail in section 2.3, have been adopted for the model development. Additionally, the significance of several data divisional approaches in various ML techniques has also been discussed.

4.2 PREPARATION OF DATASET

In the present investigation, in-situ soil samples were collected from an ongoing highway construction project work site. Brief information about the project work is provided in section 3.2. A total of 1011 datasets were extracted from various chainage point along the length of the road. Some significant geotechnical parameters, such as % gravel (G), sand content (%), fine content (FC), liquid limit (LL), plastic limit (PL), plasticity index (PI), maximum dry density (MDD), optimum moisture content (OMC) and California bearing ratio (CBR) were obtained from the laboratory experiments conducted as per Bureau of Indian Standard (BIS) specifications. The laboratory obtained database of fine-grained soil was further classified into various soil groups. Figure 4.4 presents the percentage of different soil groups of fine-grained soil. Moreover, the range of these geotechnical parameters as per the categorized soil groups is presented in Table 4.2

4.2.1 Frequency distribution plot for all geotechnical parameters

Figure 4.1 presents the histogram plot for all these parameters. It is observed from Figure 4.1 that the percentage of gravel varies from 0 to 30% and the maximum number

of observations exists up to 5%, sand varies from 0 to 50% and the datasets are frequently observed between the range of 0 to 30%, fine content varies from 50 to 100%, the liquid limit was reported uniformly throughout the ranges, plastic limit and plasticity index was reported frequently in between 15% to 30% and 4% to 16%, respectively, MDD and OMC were frequently observed in the range of 1.8g/cc to 2.1g/cc and 10% to 15%, respectively, CBR value was most frequently observed in the range of 6% to 12%.

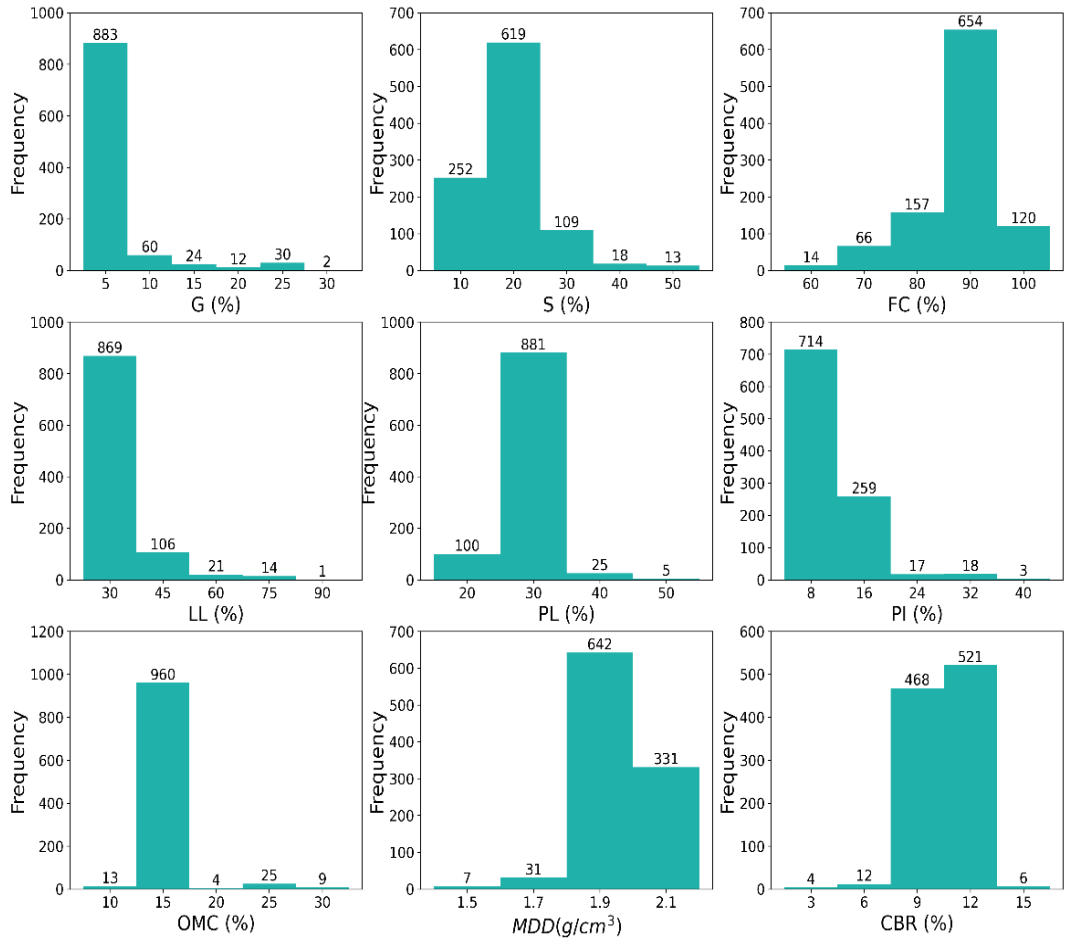


Figure 4.1 Frequency plot for all the geotechnical parameters obtained through laboratory experiments

4.2.2 Descriptive statistic ranges for all geotechnical parameters

Table 4.1 tabulates the values obtained for descriptive statistics of all geotechnical parameters. It can clearly be seen from Table 4.1 that the dataset covers an extensive

range of all the parameters, including gradational parameters, plasticity properties, compaction parameters and shear strength parameters.

Table 4.1 Descriptive statistics value for the geotechnical parameters of all dataset

	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Gravel (%)	0.00	27.42	27.42	2.83	1.23	0.00	4.64	21.54
Sand (%)	2.25	48.85	46.60	13.99	12.62	12.79	6.64	44.05
FC (%)	50.65	96.28	45.63	83.18	85.46	87.00	7.68	59.00
LL (%)	24.40	85.00	60.60	29.85	28.70	29.00	6.21	38.59
PL (%)	11.81	50.00	38.20	21.60	21.10	21.30	2.89	8.34
PI (%)	1.93	39.00	37.07	8.25	7.65	7.75	3.72	13.83
MDD (g/cc)	1.455	1.959	0.504	1.866	1.885	1.900	0.073	0.005
OMC (%)	9.50	29.50	20.00	11.96	11.45	10.70	2.52	6.36
CBR (%)	1.00	13.20	12.20	9.02	9.10	10.00	1.16	1.35

4.2.3 Selection of input parameters and correlation analysis

The development of any predictive model is performed based on the selected input parameters and output parameters. In the present study, the output was CBR value and the selection of input parameters was done by analyzing the previously attempted research in this particular domain (Bardhan, Gokceoglu, et al., 2021; Bardhan, Samui, et al., 2021; Duque et al., 2020; Katte et al., 2019; K. P. Kumar et al., 2014; S. A. Kumar et al., 2013; Kurnaz and Kaya, 2019; Patel and Desai, 2010; Rakaraddi and Gomarsi, 2015; Ramasubbarao and Sankar, 2013; Sabat, 2015; Suthar and Aggarwal, 2018; Taha et al., 2019; Talukdar, 2014; Taskiran, 2010; Tenpe and Patel, 2018, 2020; Venkatasubramanian and Dhinakaran, 2011; Vinod and Reena, 2008; Yildirim and Gunaydin, 2011; Zumrawi, 2012) and the correlation analysis performed in the current study. The correlation (R) of CBR value with other geotechnical parameters is presented in Figure 4.2. It is clearly observed from Figure 4.2 that the CBR value exhibits good strength of association with many of the gradational parameters, plasticity properties and compaction parameters of the fine-grained soil which can also be confirmed from the correlation matrix obtained for these parameters shown in Figure 4.3.

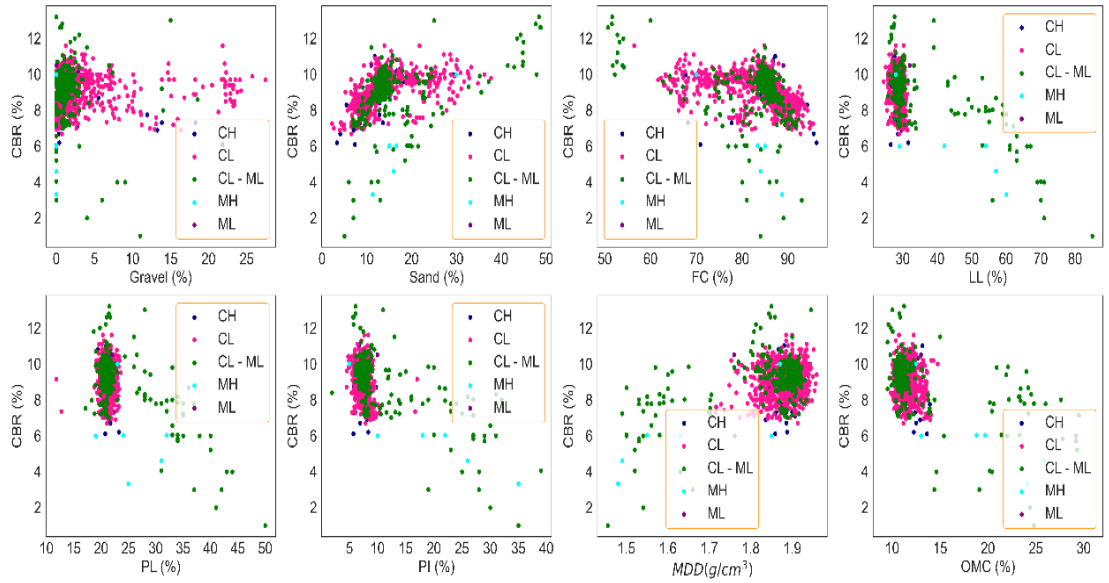


Figure 4.2 Correlation of CBR value with gradational properties, plasticity properties and compaction parameters of all dataset

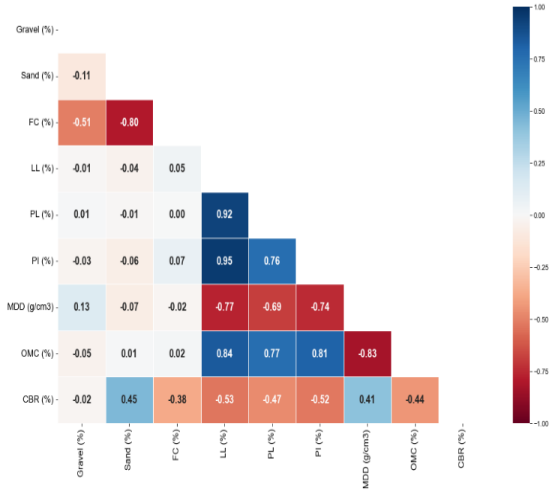


Figure 4.3 Correlation matrix for the geotechnical parameters of all dataset

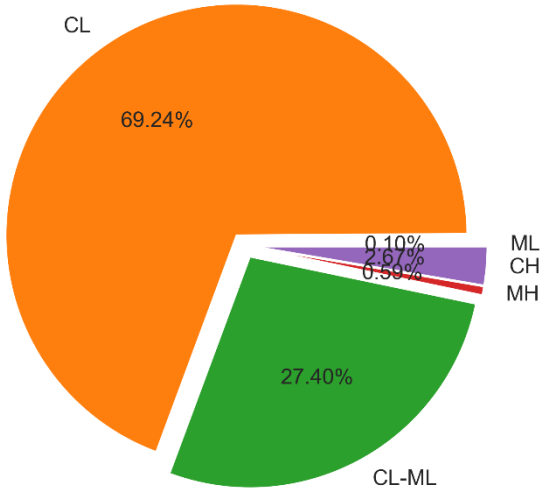


Figure 4.4 Pie plot for the different soil's groups of fine-grained soil.

Table 4.2 Variation in geotechnical parameters based on soil group.

Soil group	Observed variations in geotechnical parameters								
	Gravel	Sand	FC	LL	PL	PI	MDD	OMC	CBR
CL	0-27	2-44	53-96	26-49	12-34	7-22	1.56-1.959	9.5-24.8	6-13
ML	1	11	88	29	27	2	1.909	10.98	8
CH	0-11	5-24	76-93	54-85	23-50	22-39	1.455-1.934	12.5-29.5	1-8
MH	0-6	13-27	72-87	50-57	33-37	16-21	1.5-1.66	19.1-27.6	3-10
CL-ML	0-23	4-49	51-96	24-33	19-27	5-7	1.715-1952	9.6-13.96	6-13

The parameters which exhibit a good strength of association with the output were incorporated as input parameters for the model development. The most significant input parameters such as S, FC, PL, PI, MDD and OMC were selected using the correlation matrix given in Figure 4.3. Using those input parameters, various tentative combinations were prepared, which are tabulated in Table 4.3, to adopt the most reliable model.

Table 4.3 Tentative combinations of input parameters for developing the CBR prediction model

Model No.	Input parameter combinations	Output
1	S, FC, PL, PI, MDD, OMC	CBR
2	S, FC, PL, PI, MDD	CBR
3	S, FC, PL, PI, OMC	CBR
4	S, FC, PL, PI	CBR
5	S, FC, MDD, OMC	CBR

4.2.4 Division of dataset

Initially, the whole dataset was divided into training (TR) and testing (TS) sets. About 80% of all dataset was considered to train the model and the remaining 20% was kept for testing the model. The TS dataset was not used during the model development, consequently, the accuracy of the developed model could be verified. Three data divisional approaches i.e., statistical approach, K-fold and FCM (as discussed earlier in 2.4.2), were adopted to identify which 80% of the complete dataset is to be used to train the model and the remaining for testing the model. The results obtained for each of the data divisional approach are as follow:

4.2.4.1 Statistical approach

A simple code written in a python programming language was used to distinguish the training and testing dataset. After numerous trial and error approach, the results obtained for the descriptive statistics value of the training and testing dataset are presented in Table 4.4 and Table 4.5, respectively.

Table 4.4 Descriptive statistics value for the selected input and output parameters of the TR dataset obtained through a statistical approach

	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Sand (%)	2.25	48.85	46.60	13.95	12.56	12.79	6.53	42.67
FC (%)	50.65	96.28	45.63	83.21	85.49	84.10	7.57	57.27
PL (%)	11.81	50.00	38.20	21.55	21.11	21.30	2.82	7.94
PI (%)	1.93	35.00	33.07	8.19	7.65	7.75	3.48	12.12
MDD (g/cc)	1.455	1.959	0.504	1.868	1.885	1.900	0.068	0.005
OMC (%)	9.50	29.30	19.80	11.91	11.48	10.70	2.31	5.35
CBR (%)	1.00	13.20	12.20	9.018	9.10	10.00	1.15	1.32

Table 4.5 Descriptive statistics value for the selected input and output parameters of TS dataset obtained through a statistical approach

	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Sand (%)	4.30	47.70	43.40	14.17	12.85	11.00	7.06	49.77
FC (%)	51.55	94.88	43.33	83.05	85.42	65.57	8.14	66.23
PL (%)	19.05	42.00	22.95	21.83	21.07	20.50	3.15	9.93
PI (%)	4.85	39.00	34.16	8.49	7.60	8.10	4.55	20.69
MDD (g/cc)	1.490	1.956	0.466	1.861	1.881	1.915	0.089	0.008
OMC (%)	9.85	29.50	19.65	12.18	11.35	12.30	3.23	10.41
CBR (%)	3.00	12.80	9.80	9.03	9.28	10.00	1.23	1.50

4.2.4.2 K-fold cross-validation approach

Using a total of five folds, the K-fold splitting was done through the procedural method given in section 2.4.2.2. The K-Fold cross-validation (CV) code was written in python programming language. The descriptive statistics value obtained for the TR and TS datasets is shown in Table 4.6 and Table 4.7, respectively.

Table 4.6 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 4.7 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

4.2.4.3 Fuzzy C-means clustering approach

The procedural method given in section 2.4.2.3 is used for splitting the dataset according to the FCM approach. The code for the FCM approach was written in a python programming language. Initially, two number of cluster was taken and the silhouette value was estimated. The number of cluster was increased gradually and silhouette values for each of the corresponding clusters were estimated. The silhouette score obtained corresponding to the number of clusters has been depicted in Figure 4.5. It can clearly be observed from Figure 4.5 that the maximum silhouette score was obtained when two number of clusters were used 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 section 4.2.4.2), and labelled as $Train_1$ and $Test_1$, respectively.

Similarly, C_2 TR and TS dataset was designated as $Train_2$ and $Test_2$, respectively. The final TR dataset was obtained by concatenating the $Train_1$ and $Train_2$ datasets. Similarly, the TS dataset was achieved through the concatenation of the $Test_1$ and $Test_2$ datasets. The descriptive statistics value of the final TR and TS dataset is tabulated in Table 4.8 and Table 4.9.

Figure 4.5 Silhouette value obtained for each of the cluster

Table 4.8 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 4.9 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

4.3 DEVELOPMENT OF CBR PREDICTION MODEL

4.3.1 Multi-expression programming (MEP) model development

In the present investigation, Multi expression programming X (MEPX) software generated by M Oltean (2004) was adopted for the analysis and some of the hyper-parameters were adjusted as per the existing problem. The population size is defined as the number of the program in the population. Three different levels were set for population size. Chromosome length is a set of parameters that provides the proposed solution to a problem. The algorithm success increases with an increase in chromosome length, though, after the peak, it gets decreases as the excess chromosomes are not consumed by the program, which requires the extra memory space to store the value, consequently resulting in a less fitting value (Mihai Oltean and D Dumitrescu, 2002; M Oltean and D Dumitrescu, 2002). In the present investigation, two levels of chromosome length were taken for analysis. The number of generations is defined as the number of runs, which is continue until there is no significant improvement in the model's performance. Three levels of the run were set for the number of generations. In general, the computational time of the program increases with an increase in population size, chromosome length and a number of generations. The probability of mutation and crossover determines the probability of an offspring subjected to crossover and mutation operators, and also plays a key role in developing a reliable model. Three sets for each, crossover and mutation, were adopted for the program run. The tentative values for population size, chromosome length, number of generations, mutation probability and crossover probability are listed in Table 4.10. Basic mathematical functions were used to achieve the optimal solution. The values for other parameters were selected based on the previously attempted study (Alavi et al., 2013; Alavi et al., 2010; F.E. Jalal et al., 2021; Taskiran, 2010; Tenpe and Patel, 2018, 2020) and by analyzing the performance of

several runs. A total of 486 (3×3×2×3×3×3) different combinations of code parameters were run. The model's performance was measured in terms of performance measurement indices, as discussed in section 2.4.1.2.

Table 4.10 Hyper-parameters setting for MEP algorithm

MEP algorithm hyper-parameters	Hyper-parameters setting
Number of subpopulations	10, 20, 30
Population size	500, 1000, 2000
Chromosome length	25, 50
Number of generation	5000, 10000, 15000
Mutation probability	0.01, 0.1, 0.9
Crossover probability	0.1, 0.5, 0.9
Crossover type	Uniform
Functional set	+, -, ×, /, sqrt
Terminal set	Problem input

4.3.2 eXtreme gradient boosting (XGBoost) model development

XGBoost prediction model for CBR value was developed by writing the algorithm's code in a python programming language. Using the algorithm code, all five tentative combinations (as shown in Table 4.3) were examined for selecting the most appropriate XGBoost model. Various hyper-parameters such as booster type, Colsample by level, Colsample by node, Colsample by the tree, learning rate, maximum depth, n estimator and subsample were adjusted by analyzing the useful information available for this kind of problem in the literature. These hyper-parameters were varied as per their ranges defined in Table 4.11. The grid search cross-validation approach was adopted to better adjust these hyper-parameters. This grid search approach provides the best combinations of the algorithm's hyper-parameters within their pre-defined ranges. Grid search facilitates not only the desired hyper-parameters values but also the values of their desired outcomes.

Table 4.11 Hyper-parameters setting for XGBoost algorithm code

XGBoost Code hyper-parameters	Hyper-parameters setting
Base score	0.1, 0.5, 0.9
Booster	gblinear, gbtree
Column sample by level	1
Column sample by node	1
Column sample by tree	1
Gamma	0
Learning rate	0.01 - 0.20
Maximum delta step	0
Maximum depth	1, 3, 5, 10
Minimum child weight	1
N estimators	10, 20, 40, 60, 80, 100
Number of parallel tree	1
Regularization alpha	0
Regularization lambda	1
Scale position weight	1
Subsample	0.1, 0.3, 0.5, 0.7, 0.9

4.4 RESULTS AND DISCUSSION

4.4.1 Performance of MEP model

4.4.1.1 Statistical details of the MEP model

Table 4.12 depicts the performance of each of the trained MEP models in terms of several statistical performance indices. Here, MEP_S, MEP_K, and MEP_F indicate the MEP model developed through statistical, K-fold and FCM approaches, respectively. It can easily be observed from Table 4.12 that the R^2 value of MEP_S models ranges from 0.56 to 0.61, which means that models are able to explain the minimum variability of 56% and a maximum of up to 61% in the CBR value through adopted input parameters. The Pearson's correlation coefficient value ranges from 0.75 to 0.79. The MAE and RMSE values lie between 0.516 to 0.537 and 0.711 and 0.756, respectively, for all the trained models. The a20-index varies from 95.2% to 96.7%, which means that models are able to predict this much percentage of the dataset with in ±20% variations. In the case

of the K-fold CV approach, all the trained models are able to explain 54% to 60% variability in the CBR value. The correlation coefficient ranges from 0.74 to 0.77. The MAE and RMSE value was founded in 0.516 to 0.543 and 0.719 to 0.768, respectively. The models trained through the FCM approach can explain a minimum of 59% and a maximum of up to 62% variability in the CBR value. The correlation coefficient varies from 0.77 to 0.79. The MAE and RMSE exist in the range of 0.522 to 0.540 and 0.726 to 0.754, respectively. It is perceived from the above observations that the model trained through the statistical approach outperforms very well, which is followed by FCM and K-fold approaches.

Table 4.12 Statistical performance of various MEP models for TR dataset

Data division approach	Model no.	Statistical performance indices							
		R^2	R	MAE	RMSE	VAF	IOA	IOS	a20-index
MEP_S	1	0.6154	0.7853	0.5275	0.7111	61.6752	0.8702	0.0785	0.9666
	2	0.6006	0.7769	0.5264	0.7247	60.3456	0.8624	0.0798	0.9664
	3	0.6035	0.7785	0.5157	0.7221	60.5902	0.8671	0.0796	0.9631
	4	0.5864	0.7692	0.5363	0.7375	58.8793	0.8656	0.0813	0.9629
	5	0.5654	0.7537	0.5365	0.7560	56.7480	0.8426	0.0833	0.9518
MEP_K	1	0.5954	0.7734	0.5162	0.7191	59.8097	0.8630	0.0791	0.9592
	2	0.5647	0.7536	0.5293	0.7459	56.6914	0.8519	0.0821	0.9543
	3	0.5627	0.7542	0.5202	0.7477	56.6758	0.8535	0.0821	0.9592
	4	0.5625	0.7530	0.5390	0.7478	56.5982	0.8512	0.0822	0.9604
	5	0.5387	0.7382	0.5429	0.7679	54.4144	0.8381	0.0843	0.9617
MEP_F	1	0.6104	0.7818	0.5337	0.7368	61.1036	0.8691	0.0815	0.9654
	2	0.6216	0.7900	0.5220	0.7262	62.4040	0.8737	0.0801	0.9630
	3	0.6056	0.7788	0.5397	0.7414	60.5998	0.8677	0.0820	0.9568
	4	0.5962	0.7758	0.5256	0.7502	59.9915	0.8687	0.0826	0.9617
	5	0.5918	0.7715	0.5397	0.7542	59.5030	0.8605	0.0831	0.9605

The performance of developed models on the TS dataset is presented in Table 4.13. The R^2 value of the MEP_S, MEP_K and MEP_F model varies from 0.61 to 0.69, 0.68 to 0.72 and 0.52 to 0.60, respectively. The MAE value varies from 0.499 to 0.532, 0.504 to 0.530 and 0.496 to 0.546 for MEP_S, MEP_K and MEP_F, respectively. Therefore, it can be clearly understood from the comparative analysis of Table 4.12 and Table 4.13 that

the R^2 obtained in the TS dataset is much higher than the TR dataset. Moreover, the maximum enhancement in R^2 of the TS dataset was detected for the K-fold approach, followed by statistical and FCM approaches.

Table 4.13 Statistical performance of various MEP models for TS dataset

Data division approach	Model no.	Statistical performance indices							
		R^2	R	MAE	RMSE	VAF	IOA	IOS	a20-index
MEP_S	1	0.6868	0.8294	0.4993	0.6844	68.6849	0.8962	0.0759	0.9604
	2	0.6875	0.8316	0.5056	0.6837	68.7458	0.8933	0.0757	0.9554
	3	0.6048	0.7858	0.5281	0.7688	60.4937	0.8806	0.0853	0.9505
	4	0.6140	0.7890	0.5234	0.7598	61.4118	0.8814	0.0843	0.9406
	5	0.6101	0.7831	0.5317	0.7636	61.0420	0.8577	0.0848	0.9455
MEP_K	1	0.7188	0.8497	0.5038	0.6791	72.1108	0.9107	0.0751	0.9653
	2	0.7009	0.8391	0.5095	0.7005	70.1265	0.9010	0.0778	0.9703
	3	0.6755	0.8353	0.5302	0.7296	68.0688	0.8821	0.0804	0.9653
	4	0.7187	0.8510	0.5009	0.6793	72.2982	0.9099	0.0750	0.9554
	5	0.6787	0.8311	0.5101	0.7259	68.3004	0.8870	0.0801	0.9703
MEP_F	1	0.5172	0.7333	0.5456	0.7545	51.7157	0.8475	0.0833	0.9552
	2	0.6038	0.7808	0.4963	0.6835	60.5521	0.8742	0.0751	0.9602
	3	0.5632	0.7538	0.5291	0.7176	56.5925	0.8542	0.0788	0.9602
	4	0.5440	0.7497	0.5260	0.7332	55.8690	0.8487	0.0798	0.9552
	5	0.5366	0.7355	0.5209	0.7391	53.9591	0.8371	0.0811	0.9701

From the overall analysis of Table 4.12 and Table 4.13, one can easily understood that the maximum enhancement in accuracy was detected for the K-fold approach followed by statistical and FCM approaches.

4.4.1.2 Selection of MEP model

4.4.1.2.1 Ranking analysis (RA)

To evaluate the performance of the models, numerous statistical performance indices were adopted and the conclusions can easily be drawn by comparing their values. However, the situation becomes more mysterious when the value of different performance indices describes their own best models. For instance, in Table 4.13, model no. 2 of MEP_S depicts a higher R^2 value than model no. 1, but at the same time, the MAE value obtained for model no. 2 is high as well the a20-index is low. Therefore, it becomes

more complicated to judge the best model and which performance indices should be given more preference. In that particular situation, the ranking analysis (RA) can be useful to identify the best model as it provides the overall consideration of all the performance indices. According to RA, used by many researchers in the past (Asteris et al., 2021; Kardani, Bardhan, Kim, et al., 2021; Kardani, Bardhan, Samui, et al., 2021; H. Zhang et al., 2020), 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. For instance, consider R^2 performance indices for MEP_S models in Table 4.13. Among a total of 5 models of MEP_S, model no. 2 presents the highest value of R^2, therefore, the score for that is five and the score for the second highest model i.e., model no. 1, is four and so on. Like R^2, the score for the models of other performance indices is obtained in this way, as shown in Table 4.14. The total score for the TR and TS set of each of the models is calculated and the final score for a model is designed as the summation of the score of the TR and TS dataset.

Table 4.14, Table 4.15 and Table 4.16 describe the RA results for the statistical, K-fold and FCM data division approaches, respectively. It is observed from Table 4.14, Table 4.15 and Table 4.16 that model no.1, 1 and 2, respectively, demonstrates higher score for their respective data division approaches; therefore, top rank was assigned to them. These models are labelled as $MEP_{S\text{-}1}$, $MEP_{K\text{-}1}$ and $MEP_{F\text{-}2}$ models.

Table 4.14 Rank analysis for selecting the MEP model from a statistical approach

Performance measurement indices	MEP_S									
	1		2		3		4		5	
	TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
R^2	5	4	3	5	4	1	2	3	1	2
R	5	4	3	5	4	2	2	3	1	1

Table 4.14 (Cont.)

MAE	3	5	4	4	5	2	2	3	1	1
RMSE	5	4	3	5	4	1	2	3	1	2
VAF	5	4	3	5	4	1	2	3	1	2
IOA	5	5	2	4	4	2	3	3	1	1
IOS	5	4	3	5	4	1	2	3	1	2
a20-index	5	5	4	4	3	3	2	1	1	2
Total score	38	35	25	37	32	13	17	22	8	13
Final score	73		62		45		39		21	
Rank	1		2		3		4		5	

Table 4.15 Rank analysis for selecting the MEP model from the K-fold approach

Performance measurement indices	MEP_K									
	1		2		3		4		5	
	TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
R^2	5	5	4	3	3	1	2	4	1	2
R	5	4	3	3	4	2	2	5	1	1
MAE	5	4	3	3	4	1	2	5	1	2
RMSE	5	5	4	3	3	1	2	4	1	2
VAF	5	4	4	3	3	1	2	5	1	2
IOA	5	5	3	3	4	1	2	4	1	2
IOS	5	4	4	3	3	2	2	5	1	1
a20-index	3	3	1	5	2	2	4	1	5	4
Total score	38	34	24	26	26	11	18	33	12	16
Final score	72		50		37		51		28	
Rank	1		3		4		2		5	

Table 4.16 Rank analysis for selecting the MEP model from the FCM approach

Performance measurement indices	MEP_F									
	1		2		3		4		5	
	TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
R^2	4	1	5	5	3	4	2	3	1	2
R	4	1	5	5	3	4	2	3	1	2
MAE	3	1	5	5	2	2	4	3	1	4
RMSE	4	1	5	5	3	4	2	3	1	2
VAF	4	1	5	5	3	4	2	3	1	2
IOA	4	2	5	5	2	4	3	3	1	1
IOS	4	1	5	5	3	4	2	3	1	2
a20-index	5	1	4	4	1	3	3	2	2	5
Total score	32	9	39	39	20	29	20	23	9	20

Table 4.16 (Cont.)

Final score	41	78	49	43	29
Rank	4	1	2	3	5

4.4.1.2.2 Overfitting analysis

Overfitting is a common problem faced by the developed prediction models. Overfitting occurs when a model outperforms very well on the relevant dataset (training dataset) only and is irrelevant to any other datasets. Thus, the overfitting of a model reduces its generalizability outside the original datasets. In a regression problem, overfitting can produce misleading to the performance of any model. Therefore, to avoid the selection of an over-fitting model, it is essential to test for the overfitting analysis. In general, the overfitting ratio (OR) is estimated using equation (4.1) corresponding to each of the performance measurement parameters. According to this formula, a model that has the highest value of OR for the trend class parameters (R^2, R, VAF, IOA, and a20-index) and a lower value for the error class parameters (MAE, RMSE and IOS) is considered to be less prone to the overfitting.

$$OR = \frac{Performance\ parameter\ of\ testing\ dataset}{Performance\ parameters\ of\ training\ dataset} \quad (4.1)$$

For selecting the final MEP algorithm model, further, RA was performed on the results achieved for the overfitting analysis in MEP_{S-1}, MEP_{K-1} and MEP_{F-2} models. Table 4.17 tabulates the OR values along with the RA value obtained corresponding to each of the statistical performance measurement parameters in MEP models. It is observed from Table 4.17 that the MEP_{K-1} model comprises the highest score, whereas MEP_{S-1} and MEP_{F-2} model shares same score. Based on the above findings, MEP_{K-1} is ranked as 1 and MEP_{S-1} and MEP_{F-2} models are ranked as 2.

Table 4.17 Overfitting ratio values along with the rank analysis in MEP models

Performance measurement indices	MEP_{S-1}		MEP_{K-1}		MEP_{F-2}	
	OR		OR		OR	
R^2	1.1160	2	1.2073	3	0.9714	1
R	1.0562	2	1.0987	3	0.9883	1
MAE	0.9465	3	0.9760	1	0.9508	2
RMSE	0.9625	1	0.9444	2	0.9412	3
VAF	1.1137	2	1.2057	3	0.9703	1
IOA	1.0299	2	1.0553	3	1.0006	1
IOS	0.9669	1	0.9494	2	0.9376	3
a20-index	0.9936	1	1.0064	3	0.9971	2
Total score	14		20		14	
Rank	2		1		2	

The best MEP model (MEP_{K-1}) comprises of S, FC, PI and MDD as input parameters and CBR as an output parameter. Table 4.18 presents the best achieved MEP algorithm hyper-parameters for predicting the CBR value of fine-grained soil.

Table 4.18 Selected combination of MEP algorithm hyper-parameters

Code parameters	Parameters setting
Number of subpopulations	20
Population size	1000
Chromosome length	50
Number of generation	10000
Mutation probability	0.01
Crossover probability	0.9
Crossover type	Uniform
Functional set	+, -, ×, /, sqrt
Terminal set	Problem input

Through these final code parameters, the final equation for predicting the CBR value is A9, which is obtained as discussed below:

$$A_1 = \left(S - 2MDD - \frac{FC}{S}\right)$$

$$A_2 = \frac{FC(2MDD + 1)}{S}$$

$$A_3 = \frac{A_1\left(\sqrt{FC/S}\right)}{(PI + A_1^2)}$$

$$A_4 = \frac{S\left(\sqrt{S \times FC}\right)}{2FC \times PI \times MDD}$$

$$A_5 = \frac{PI - 4MDD}{S}$$

$$A_6 = (S \times FC)^{\frac{1}{4}} + 2MDD - A_5$$

$$A_7 = S - 2MDD + \frac{\sqrt{S \times FC}}{PI} - A_3(S \times FC)^{\frac{1}{4}}$$

$$A_8 = A_4A_5 + A_6 + \frac{A_4}{4MDD \times A_2 \times A_3} + A_3A_5$$

$$A_9 = A_3 + A_8 + A_5\left(A_4 - \frac{1}{A_2A_3}\right) - \frac{A_7}{A_2 - PI + 2MDD + S} + \frac{A_1A_4A_5}{4MDD \times A_1 - PI - A_1^2} \quad (4.2)$$

4.4.2 Performance of XGBoost model

4.4.2.1 Statistical details of the XGBoost model

The performance of models trained through the XGBoost algorithm is presented in Table 4.19. Here, XGB_S, XGB_K, and XGB_F label the XGBoost model developed through statistical, K-fold and FCM approach, respectively. It is observed from Table

4.19 that the R^2 value of XGB_S models lies between 0.79 to 0.82, which means that models are able to explain the minimum variability of 79% and a maximum of up to 82% in the CBR value through adopted input parameters. The Pearson's correlation coefficient value ranges from 0.89 to 0.91. The MAE value lies between 0.378 to 0.406 and RMSE varies from 0.482 to 0.527. The a20-index varies from 99% to 99.5%, which means that models can predict this much percentage of the dataset with in ±20% variations. In the case of the K-fold cross-validation approach, the respective input parameters are able to explain 78% to 81% variability in the CBR value. The correlation coefficient ranges from 0.88 to 0.90. The MAE and RMSE value was founded in the range of 0.386 to 0.412 and 0.495 to 0.535, respectively. The a20-index varies from 99.1% to 99.5%. The models trained through the FCM approach can explain a minimum of 80% and a maximum of up to 84% variability in the CBR value. The R value varies from 0.89 to 0.92. The MAE and RMSE exist in the range of 0.370 to 0.405 and 0.478 to 0.532, respectively. It is understood from the above findings that the model trained through the FCM approach outperforms very well and is followed by statistical and K-fold approaches. Additionally, the performance measurement indices followed a specific trend in many of the models of all three data divisional approaches, which states that there is no uncertainty in the results of the model.

Table 4.19 Statistical performance of various XGBoost models for TR dataset

Data division approach	Model no.	Statistical performance indices							
		R^2	R	MAE	RMSE	VAF	IOA	IOS	a20-index
XGB_S	1	0.8234	0.9089	0.3779	0.4818	82.3444	0.9469	0.0534	0.9913
	2	0.8135	0.9034	0.3888	0.4952	81.3510	0.9435	0.0549	0.9951
	3	0.8162	0.9052	0.3880	0.4916	81.6264	0.9442	0.0545	0.9951
	4	0.8014	0.8967	0.4016	0.5110	80.1423	0.9392	0.0567	0.9901
	5	0.7885	0.8896	0.4062	0.5273	78.8538	0.9346	0.0585	0.9901
XGB_K	1	0.8086	0.9002	0.3855	0.4946	80.8656	0.9423	0.0548	0.9938
	2	0.8049	0.8988	0.3884	0.4994	80.4886	0.9403	0.0553	0.9938
	3	0.7954	0.8940	0.3980	0.5114	79.5420	0.9364	0.0566	0.9951
	4	0.7882	0.8898	0.4079	0.5204	78.8163	0.9339	0.0576	0.9926

Table 4.19 (Cont.)

	5	0.7757	0.8825	0.4123	0.5354	77.5743	0.9296	0.0593	0.9913
XGB_F	1	0.8359	0.9155	0.3701	0.4782	83.5940	0.9514	0.0531	0.9926
	2	0.8207	0.9075	0.3882	0.4998	82.0747	0.9460	0.0555	0.9938
	3	0.8325	0.9141	0.3740	0.4832	82.2494	0.9498	0.0536	0.9951
	4	0.8101	0.9017	0.4024	0.5145	81.0087	0.9421	0.0571	0.9914
	5	0.7968	0.8944	0.4053	0.5322	79.6784	0.9373	0.0591	0.9864

The performance of developed XGBoost models on the TS dataset is presented in Table 4.20. The R^2 value of XGB_S, XGB_K and XGB_F model varies from 0.60 to 0.65, 0.71 to 0.76 and 0.65 to 0.70, respectively. The MAE value varies from 0.515 to 0.540, 0.467 to 0.487 and 0.465 to 0.491 for XGB_S, XGB_K and XGB_F, respectively. The values obtained for a20-index reveals that XGB_S models are able to predict 94.1% to 96.5% observations within ±20% variations whereas for XGB_K and XGB_F models ranges from 97% to 98.5% and 96.5% to 98.5%, respectively. Therefore, it is understood from the above findings that the accuracy of the TS dataset is lower than the TR dataset and the maximum deficiency was noticed for the statistical approach. Moreover, the comparative analysis of Table 4.19 and Table 4.20 illustrates the presence of overfitting for the statistical and FCM approaches as the value R^2 obtained for the TS dataset is comparatively lower than the TR dataset.

Table 4.20 Statistical performance of various XGBoost models for TS dataset

Data division approach	Model no.	Statistical performance indices							
		R^2	R	MAE	RMSE	VAF	IOA	IOS	a20-index
XGB_S	1	0.6018	0.7862	0.5351	0.7717	60.5000	0.8791	0.0861	0.9406
	2	0.6464	0.8061	0.5146	0.7271	64.7607	0.8884	0.0809	0.9653
	3	0.6457	0.8105	0.5185	0.7279	65.0491	0.8920	0.0814	0.9554
	4	0.6224	0.7940	0.5397	0.7515	62.5313	0.8817	0.0838	0.9505
	5	0.6473	0.8083	0.5226	0.7263	65.1844	0.8876	0.0812	0.9505
XGB_K	1	0.7606	0.8793	0.4705	0.6267	76.0937	0.9198	0.0696	0.9802
	2	0.7620	0.8782	0.4761	0.6248	76.2060	0.9214	0.0697	0.9851
	3	0.7621	0.8815	0.4670	0.6247	76.2632	0.9201	0.0694	0.9703
	4	0.7555	0.8753	0.4908	0.6333	75.6146	0.9185	0.0703	0.9802
	5	0.7115	0.8481	0.4870	0.6879	71.1585	0.9010	0.0765	0.9802
XGB_F	1	0.6846	0.8291	0.4743	0.6097	68.4901	0.9043	0.0672	0.9851

Table 4.20 (Cont.)

2	0.6965	0.8367	0.4650	0.5982	69.6625	0.9098	0.0660	0.9801
3	0.6604	0.8143	0.4861	0.6327	66.0621	0.8951	0.0698	0.9751
4	0.6472	0.8072	0.4913	0.6449	64.7452	0.8914	0.0711	0.9652
5	0.6659	0.8162	0.4776	0.6276	66.6027	0.8935	0.0692	0.9751

From the overall analysis of Table 4.19 and Table 4.20, one can easily understood that the maximum enhancement in accuracy was detected for the K-fold approach followed by FCM and statistical approaches.

4.4.2.2 Selection of XGBoost model

4.4.2.2.1 Ranking analysis

The selection of the best-fitted model in the XGBoost algorithm for all three data divisional approaches is also performed through the rank analysis technique, as discussed previously in section 4.4.1.2. Table 4.21, Table 4.22 and Table 4.23 describe the rank analysis for the statistical, K-fold and FCM model, respectively, of the XGBoost algorithm. It is observed from Table 4.21, Table 4.22 and Table 4.23 that model no. 3, 1 and 1, respectively, exhibit the higher score in their respective data division approaches therefore highest rank was assigned to them. According to the model number, these selected models are labelled as $XGB_{S\text{-}3}$, $XGB_{K\text{-}1}$ and $XGB_{F\text{-}1}$.

Table 4.21 Rank analysis for selecting the XGBoost model from the statistical approach

Performance measurement indices	XGB_S									
	1		2		3		4		5	
	TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
R^2	5	1	3	4	4	3	2	2	1	5
R	5	1	3	3	4	5	2	2	1	4
MAE	5	2	3	5	4	4	2	1	1	3
RMSE	5	1	3	4	4	3	2	2	1	5
VAF	5	1	3	3	4	4	2	2	1	5
IOA	5	1	3	4	4	5	2	2	1	3
IOS	5	1	3	5	4	3	2	2	1	4
a20-index	3	1	5	5	4	4	2	2	1	3
Total score	38	9	26	33	32	31	16	15	8	32

Table 4.21 (Cont.)

Final score	47	59	63	31	40
Rank	3	2	1	5	4

Table 4.22 Rank analysis for selecting the XGBoost model from the K-fold approach

Performance measurement indices	XGB_K									
	1		2		3		4		5	
	TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
R^2	5	3	4	4	3	5	2	2	1	1
R	5	4	4	3	3	5	2	2	1	1
MAE	5	4	4	3	3	5	2	1	1	2
RMSE	5	3	4	4	3	5	2	2	1	1
VAF	5	3	4	4	3	5	2	2	1	1
IOA	5	3	4	5	3	4	2	2	1	1
IOS	5	4	4	3	3	5	2	2	1	1
a20-index	4	4	3	5	5	1	2	3	1	2
Total score	39	28	31	31	26	35	16	16	8	10
Final score	67		62		61		32		18	
Rank	1		2		3		4		5	

Table 4.23 Rank analysis for selecting the XGBoost model from the FCM approach

Performance measurement indices	XGB_F									
	1		2		3		4		5	
	TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
R^2	5	4	3	5	4	2	2	1	1	3
R	5	4	3	5	4	2	2	1	1	3
MAE	5	4	3	5	4	2	2	1	1	3
RMSE	5	4	3	5	4	2	2	1	1	3
VAF	5	4	3	5	4	2	2	1	1	3
IOA	5	4	3	5	4	3	2	1	1	2
IOS	5	4	3	5	4	2	2	1	1	3
a20-index	3	5	4	4	5	3	2	1	1	2
Total score	38	33	25	39	33	18	16	8	8	22
Final score	71		64		51		24		30	
Rank	1		2		3		5		4	

4.4.2.2.2 Overfitting analysis

Using the above equation (4.1), the overfitting ratio was estimated for each of the performance measurement parameters of XGB models. For selecting the final XGBoost algorithm model, further, RA was performed on the results achieved for the overfitting analysis in XGB_{S-3}, XGB_{K-1} and XGB_{F-1} models. Table 4.24 tabulates the OR values along with the RA value obtained corresponding to each of the statistical performance measurement parameters in XGB models. It is observed from Table 4.24 that the XGB_{K-1} model comprises the highest score, followed by XGB_{F-1} and XGB_{S-3} models. Therefore, based on the above findings, XGB_{K-1}, XGB_{F-1} and XGB_{S-3} models are ranked as 1, 2 and 3, respectively.

Table 4.24 Overfitting ratio values along with the rank analysis in XGBoost models

Performance measurement indices	XGB_{S-3}		XGB_{K-1}		XGB_{F-1}	
	OR		OR		OR	
R^2	0.7911	1	0.9406	3	0.8190	2
R	0.8954	1	0.9768	3	0.9056	2
MAE	1.3363	1	1.2205	3	1.2815	2
RMSE	1.4807	1	1.2671	3	1.2750	2
VAF	0.7969	1	0.9410	3	0.8193	2
IOA	0.9447	1	0.9761	3	0.9505	2
IOS	1.4938	1	1.2701	2	1.2655	3
a20-index	0.9601	1	0.9863	2	0.9924	3
Total score	8		22		18	
Rank	3		1		2	

The final selected XGBoost model is comprised of S, FC, PL, PI, MDD and OMC as input parameters and CBR as an output parameter. Table 4.25 presents the best-achieved hyper-parameters of the XGBoost algorithm for predicting the CBR value of fine-grained soil.

Table 4.25 Selected combination of XGBoost algorithm hyper-parameters

Code parameters	Parameters setting
Base score	0.5
Booster type	gbtree
Col sample by level	1
Col sample by node	1
Col sample by tree	1
Gamma	0
Learning rate	0.161
Maximum delta step	0
Maximum depth	3
Minimum child weight	1
N estimators	80
N jobs	-1
Number of parallel trees	1
Regularization alpha	0
Regularization lambda	1
Subsample	0.5

4.4.3 Comparative performance of selected MEP and XGBoost model

4.4.3.1 Visual interpretation of the results

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

4.4.3.1.1 Trend and error plot for the selected models

Based on the ranking and overfitting analysis results, the MEP_{K-1} and XGB_{K-1} model was selected for MEP and XGBoost algorithms, respectively. Figure 4.6 and Figure 4.7 depicts the actual versus predicted value for the TR dataset of MEP_{K-1} and XGB_{K-1} models, respectively. The center line represents the line of equality or 1:1 line whereas the upper and lower line denotes the upper and lower bound which was taken as ±20%. It is observed from the scatter plot results that the maximum number of observations follows a specific trend and are very closely distributed to line of equality. The closeness

of data points toward the line of equality is maximum for the XGB_{K-1} model, and an acceptable inclination is obtained for the MEP_{K-1} model. This means that the soaked CBR value predicted through XGBoost algorithm are much closer to the actual value. Similarly, the results obtained for the TS dataset of MEP_{K-1} and XGB_{K-1} model is presented in Figure 4.8 and Figure 4.9, respectively. The overall trend analysis of the TR and TS dataset reveals that the XGBoost algorithm is much superior in predicting the CBR value more closely to the actual value as compared to the MEP algorithm.

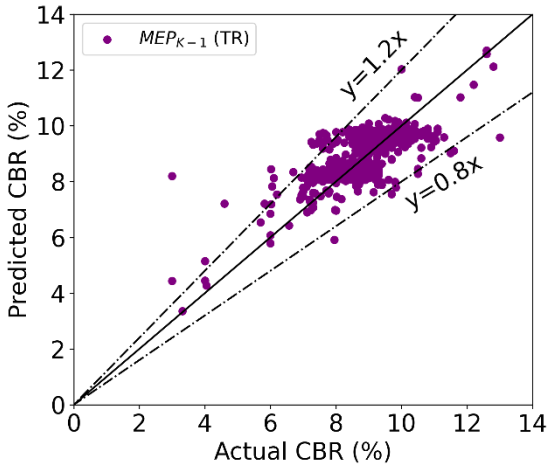


Figure 4.6 Actual versus predicted CBR value for the TR datasets of the MEP_{K-1} model

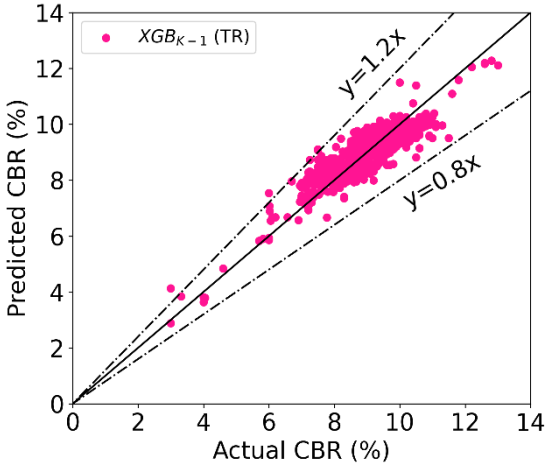


Figure 4.7 Actual versus predicted CBR value for the TR datasets of the XGB_{K-1} model

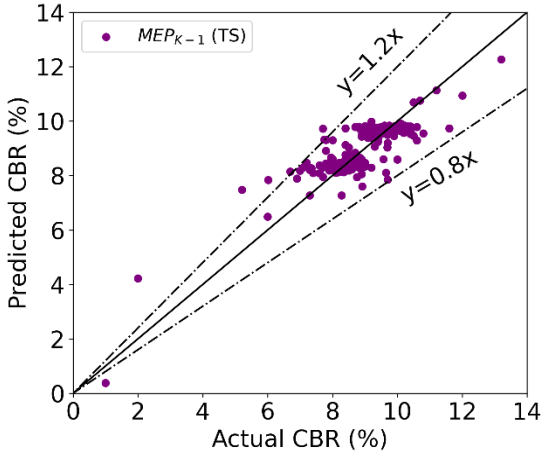


Figure 4.8 Actual versus predicted CBR value for the TS datasets of the MEP_{K-1} model

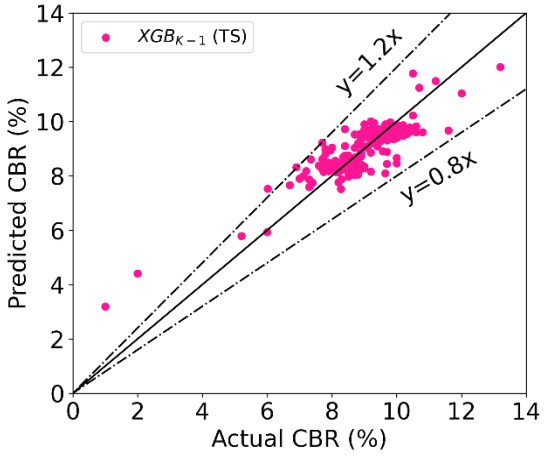


Figure 4.9 Actual versus predicted CBR value for the TS datasets of the XGB_{K-1} model

The error distribution plot for the TR dataset of MEP_{K-1} and XGB_{K-1} models is shown in Figure 4.10 and Figure 4.11, respectively. The center horizontal line 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 Figure 4.10 and Figure 4.11 that some datasets are below the zero error line (displays negative error) and some are above the zero error line (displays positive error). This random pattern of the error indicates that the selected MEP_{K-1} and XGB_{K-1} model demonstrates a decent fit to the dataset. However, the existence of a dataset within ±20% variation seems to be maximum for the XGB_{K-1} model (refer to Figure 4.11) as compared to the MEP_{K-1} model (refer to Figure 4.10). This can also be confirmed from Figure 4.12 and Figure 4.13 shows the error frequency plot for the TR dataset of MEP_{K-1} and XGB_{K-1} model, respectively. As seen from Figure 4.12 (for the MEP_{K-1} model), almost 83% and 96% of observations can be predicted within ±10% and ±20% variations, respectively, whereas for the XGB_{K-1} model (see Figure 4.13) was found to be 92% and 99% observations within ±10% and ±20% variations, respectively. A similar trend was observed for the TS dataset, as shown in Figure 4.14 to Figure 4.17. Conclusively, it is perceived that the soaked CBR value of fine-grained plastic soil predicted through the XGBoost algorithm is much perfection as compared to the CBR predicted through the MEP algorithm. Therefore, one can understand that the predictive ability is prominently influenced by the type of ML algorithm used for developing the prediction model.

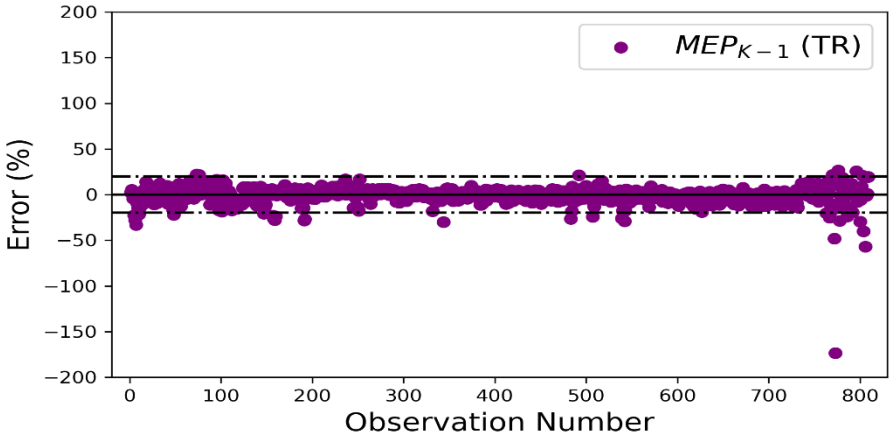


Figure 4.10 Error distribution plot for the TR datasets of the MEP_{K-1} model

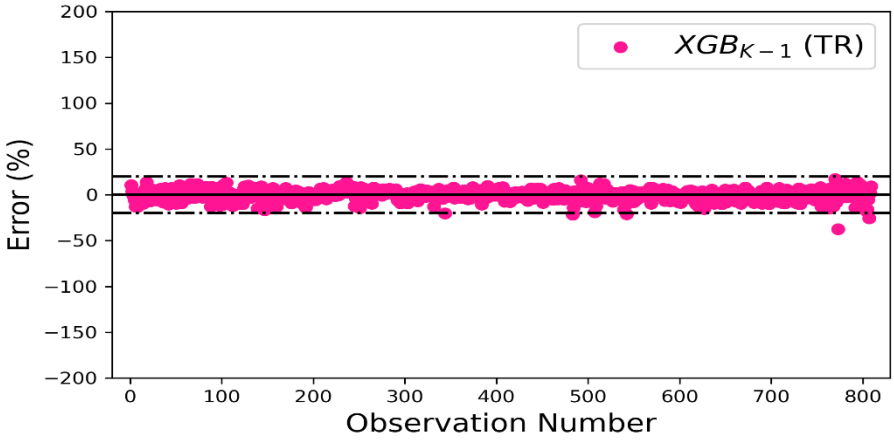


Figure 4.11 Error distribution plot for the TR datasets of the XGB_{K-1} model

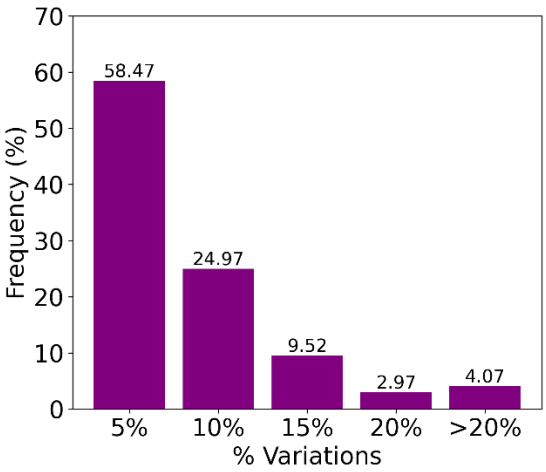


Figure 4.12 Error histogram plot for the TR datasets of MEP_{K-1} model

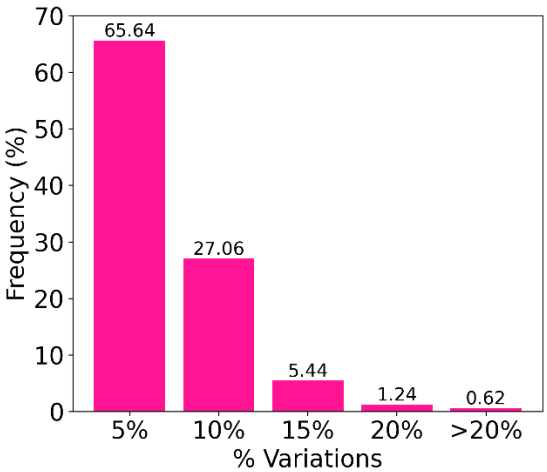


Figure 4.13 Error histogram plot for the TR datasets of the XGB_{K-1} model

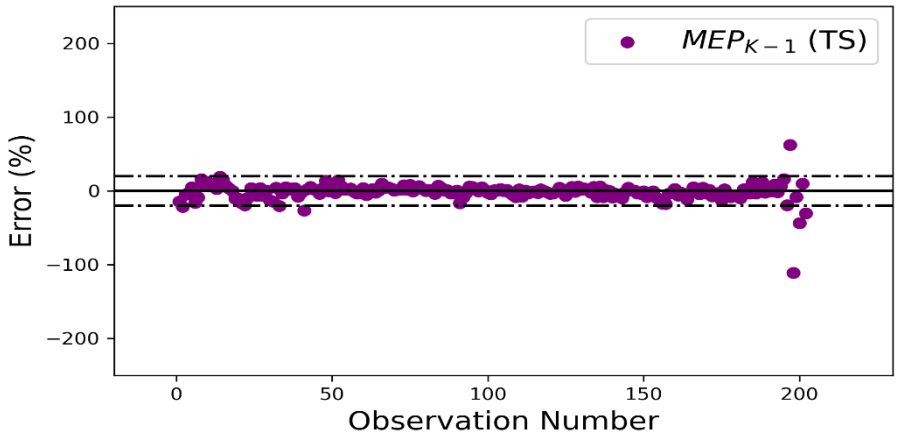


Figure 4.14 Error distribution plot for the TS datasets of MEP_{K-1} model

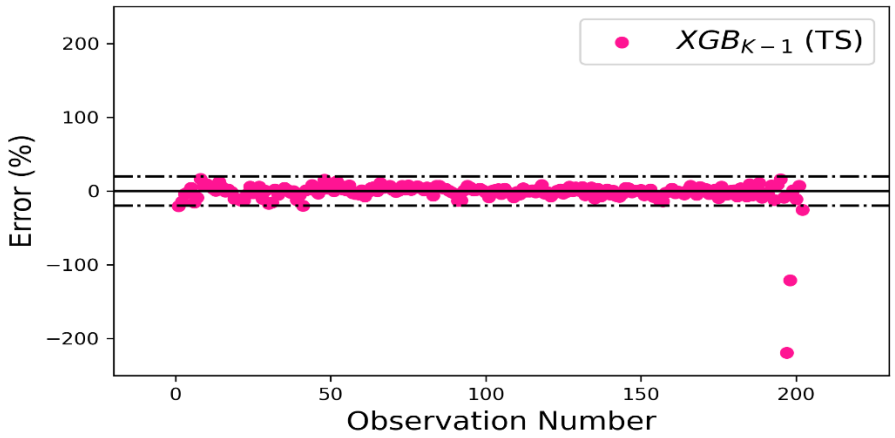


Figure 4.15 Error distribution plot for the TS datasets of the XGB_{K-1} model

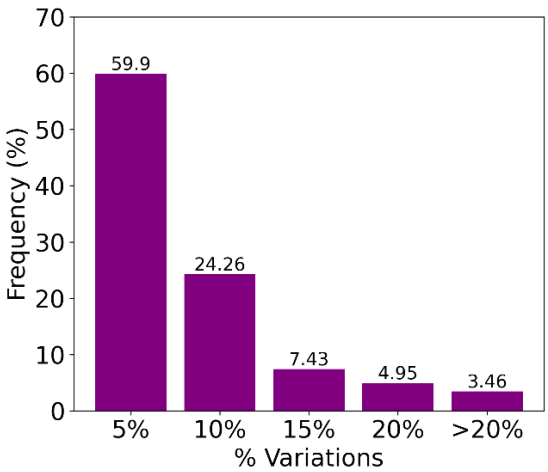


Figure 4.16 Error histogram plot for the TS datasets of MEP_{K-1} model

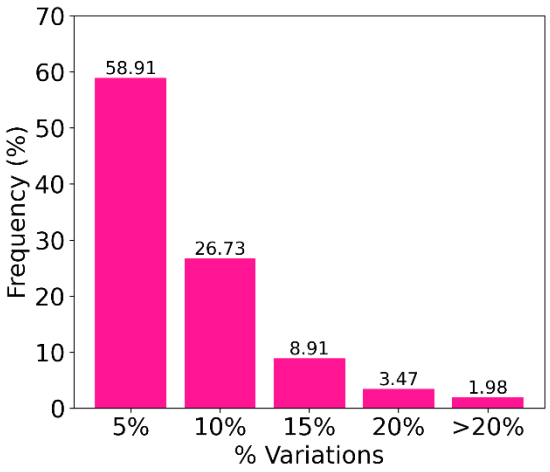


Figure 4.17 Error histogram plot for the TS datasets of the XGB_{K-1} model

4.4.3.1.2 Regression error characteristics (REC) curve

In the 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 (Asteris et al., 2021; Kardani, Bardhan, Kim, et al., 2021; Kardani, Bardhan, Samui, et al., 2021). 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.

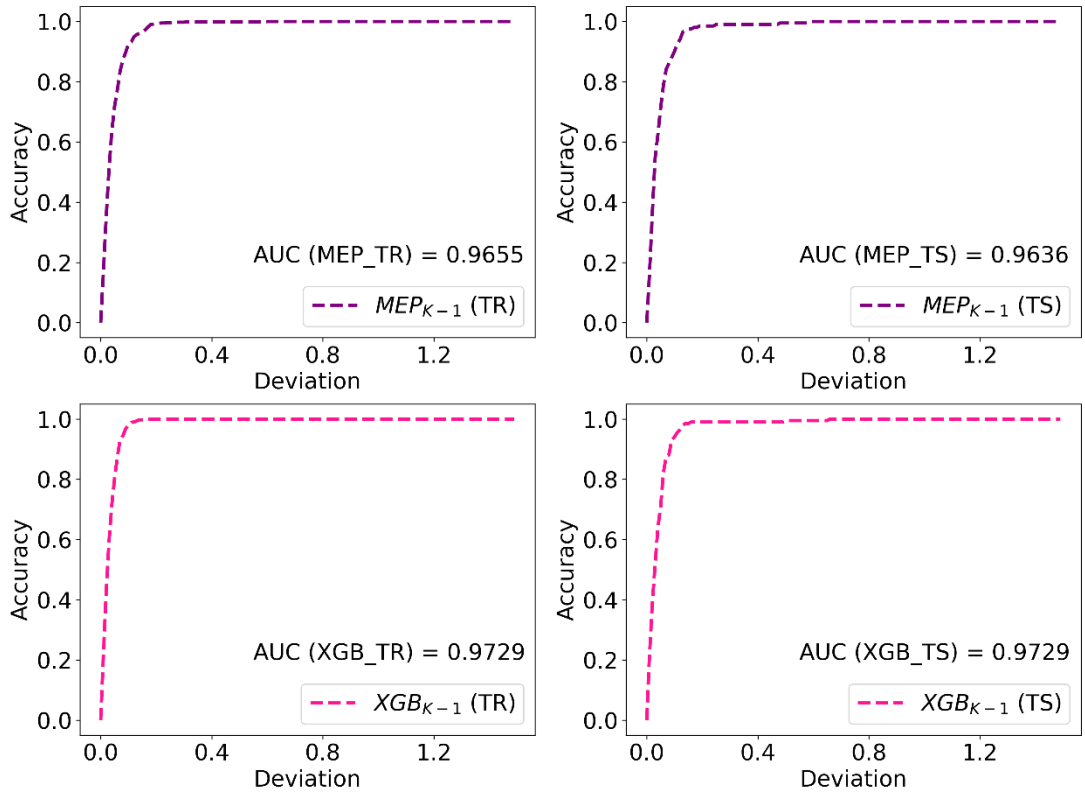


Figure 4.18 REC curve for MEP_{K-1} TR dataset, MEP_{K-1} TS dataset, XGB_{K-1} TR dataset and XGB_{K-1} TS dataset

Figure 4.18 (a) and (c) presents the REC curve obtained corresponding to the training dataset of MEP_{K-1} and XGB_{K-1} models, respectively, and the testing dataset of both models is shown in Figure 4.18 (b) and (d). As seen from these figures, the AUC value obtained for MEP_{K-1} and XGB_{K-1} models is higher than 0.9, which means that both models outperform very well and are stated to be reliable in predicting the soaked CBR value of fine-grained plastic soils. However, it is clearly observed from the comparative analysis of both models in the training and testing set that the REC curve for the XGB_{K-1} model exists more closely to the upper left corner as compared to the MEP_{K-1} model and the AUC value achieved for XGB_{K-1} model is higher than that of MEP_{K-1} model. Hence, the model developed through the XGBoost algorithm is considered to be the better predictive model.

4.4.3.1.3 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 (see Table 2.3), using equations (4.3) and (4.4).

$$A_e = |(1 - |m_e|)| \times 100 \tag{4.3}$$

$$A_t = \frac{|m_t|}{i_t} \times 100 \tag{4.4}$$

Where, A_e and A_t denotes 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, RMSE and IOS, belong to the class of error, whereas R^2, R, VAF, IOA and a20-index belong to the trend. i_t represents the ideal value of the respective error and trend parameters.

Table 4.26 tabulates the accuracy of the selected MEP and XGBoost model. As seen from Table 4.26, the accuracy achieved for the XGBoost algorithm is much higher than that of the MEP algorithm from all aspects of the statistical performance measurement parameters. Furthermore, the R^2, R, MAE, RMSE, VAF, IOA, IOS and a20-index value get improved by 27%, 13%, 23%, 65%, 27%, 7%, 2% and 3% when XGBoost algorithm is adopted over the MEP algorithm.

Table 4.26 Accuracy of selected MEP_{K-1} and XGB_{K-1} models

Statistical performance measurement parameters	Accuracy of MEP_{K-1} model	Accuracy of XGB_{K-1} model	Difference in the accuracy of both the models (%)
R^2	62.55	79.70	27.42
R	79.25	89.46	12.88
MAE	48.63	59.75	22.87
RMSE	28.87	47.64	65.02
VAF	62.81	79.71	26.91
IOA	87.52	93.73	7.06
IOS	92.17	94.20	2.20
a20-index	96.04	99.11	3.20

4.4.4 Validation of present and literature study models

The selected MEP_{K-1} and XGB_{K-1} model developed through the TR dataset demonstrates good results in the TS dataset as well. The MEP_{K-1} and XGB_{K-1} model shows higher accuracy, having R value greater than 0.80. According to Smith (1986), if R value of a model is higher than 0.8, the actual and predicted values are strongly correlated with each other as well as in good agreement. Although the prediction model exhibits higher accuracy in the TR and TS phase, it can't be considered reliable without assessing its generalization capability. In the present study, internal and external validation was conducted through K-Fold CV approach and some new datasets collected from literature studies, respectively.

4.4.4.1 Internal validation of present study models through the K-Fold CV approach

The K-Fold CV approach with five folds was effectively utilized to evaluate the predictive capability of the MEP_{K-1} and XGB_{K-1} models. Figure 4.19 and Figure 4.20 demonstrate the validation results of the MEP_{K-1} and XGB_{K-1} model, respectively. The results indicate the fulfillment of the MEP_{K-1} and XGB_{K-1} models. Additionally, the fact that the R value ranged from 0.9913 to 0.9950 for the XGB_{K-1} model, whereas for the MEP_{K-1} model lies between 0.9496 to 0.9573. Thus, it can be perceived from Figure 4.19 and Figure 4.20 that the XGB_{K-1} model depicts higher accuracy as compared to the MEP_{K-1} model.

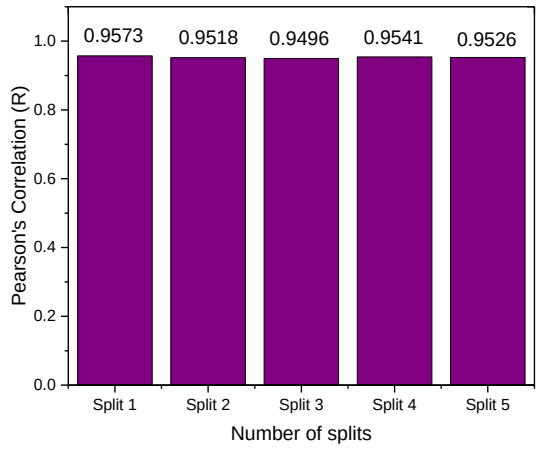


Figure 4.19 Performance measure of MEP_{K-1} model through K-Fold CV approach

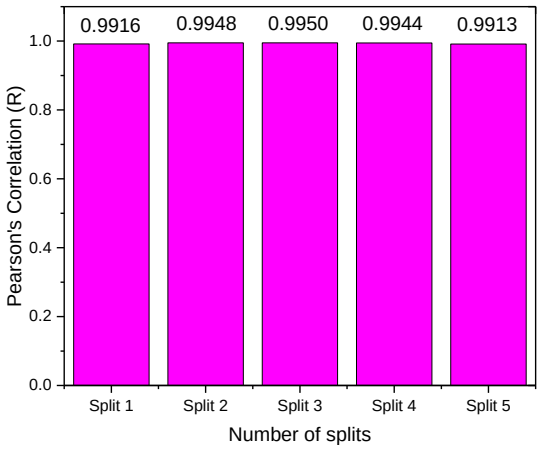


Figure 4.20 Performance measure of XGB_{K-1} model through K-Fold CV approach

4.4.4.2 External validation of present study models through the literature dataset

The external validation of the present study model was performed by collecting the soil dataset from Kin (2006), Yared (2013), Farias et al. (2018) and Gül and Çayir (2020) studies belong to the various country in the world. Table 4.27 presents the origin of the soil dataset used for the investigation. From these studies, the only dataset is occupied, which exists within the range of minimum and maximum value of the

geotechnical parameters of the present study datasets. The descriptive statistic values obtained for the literature datasets are presented in Table 4.28.

Table 4.27 Details of the literature dataset used to validate the present study model

S. No.	Literature study	Soil origin	Datasets from the literature study
1	Kin (2006)	Malaysia	21
2	Yared (2013)	Ethiopia	34
3	Farias et al. (2018)	Peru	20
4	Gül and Çayir (2020)	Turkey	04

Table 4.28 Descriptive statistic details for literature dataset

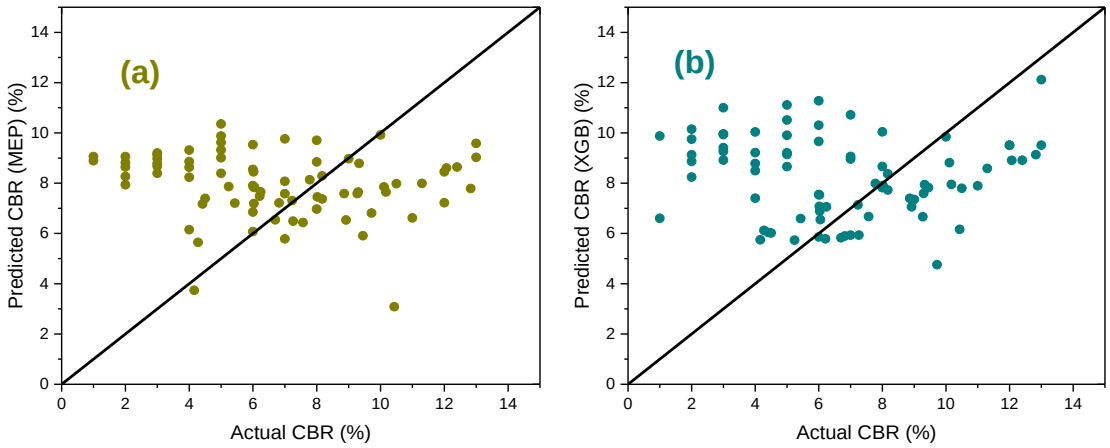
	Min.	Max.	Range	Average	Median	Mode	S.D.	Variance
Gravel (%)	0.00	24.00	24.00	3.49	0.00	0.00	5.75	33.08
Sand (%)	6.00	47.00	41.00	20.78	19.75	20.00	8.74	76.35
FC (%)	51.00	93.30	42.30	75.73	78.85	80.00	10.88	118.41
LL (%)	26.00	71.00	45.00	49.81	50.50	39.00	12.30	151.19
PL (%)	15.00	45.00	30.00	28.57	28.50	28.00	7.52	56.56
PI (%)	5.00	39.00	34.00	21.24	22.00	22.00	7.76	60.21
MDD (g/cc)	1.463	1.950	0.487	1.693	1.645	1.500	0.165	0.027
OMC (%)	9.80	29.30	19.50	19.04	18.90	13.00	5.59	31.26
CBR (%)	1.00	13.00	12.00	6.31	6.00	4.00	3.08	9.49

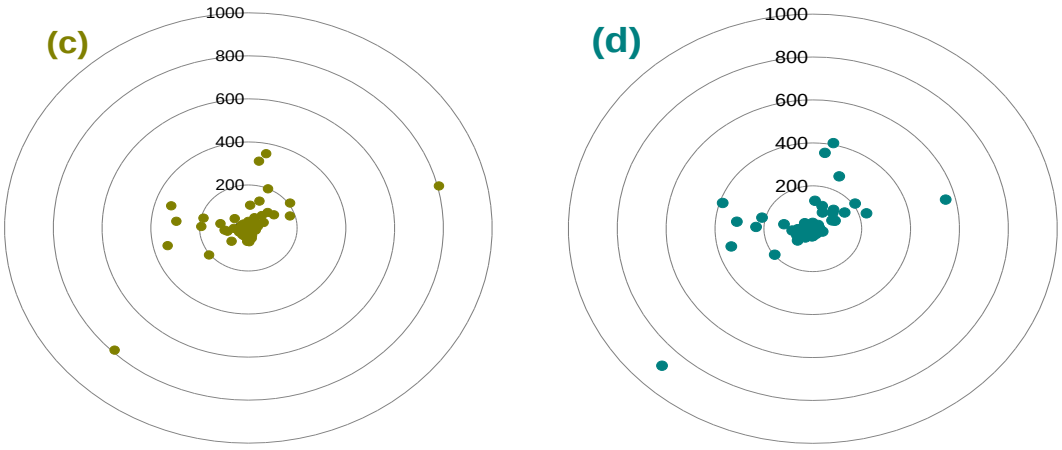
Table 4.29 illustrates the comparative performance of proposed MEP and XGBoost models on the literature dataset in terms of statistical performance indices. It is clearly seen from Table 4.29 that the proposed MEP_{K-1} and XGB_{K-1} models are not efficient in predicting the soaked CBR value of the literature dataset. The R^2 value obtained for both 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. This can also be confirmed from Figure 4.21 (a) and (b) presents the scatter plot for MEP_{K-1} and XGB_{K-1} model, respectively. Results obtained from the error radar plot (c-d) and error histogram plot (e-f) for the MEP_{K-1} and XGB_{K-1} model also reveal that the developed models are almost inefficient in predicting the soaked CBR of literature datasets. Among

79 literature datasets, only 34% observations can be predicted within ±20% variations, which is quite low.

Table 4.29 Comparative performance of proposed ML algorithm models on literature dataset

ML approach	Statistical performance indices							
	R^2	R	MAE	RMSE	VAF	IOA	IOS	a20-index
MEP	-0.4796	-0.1628	3.0408	3.7355	-31.2288	0.3473	0.4698	0.3038
XGBoost	-0.5804	-0.0533	3.0642	3.8606	-32.6223	0.4071	0.4683	0.3418





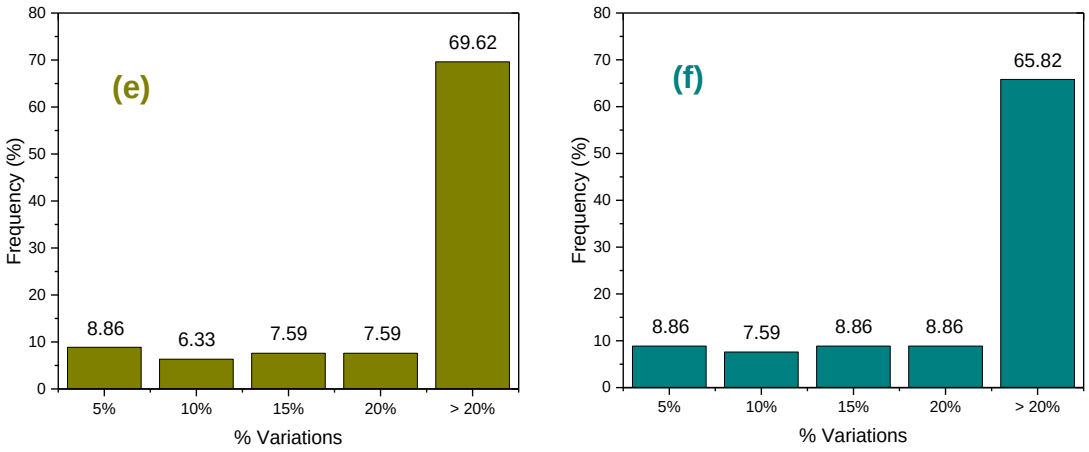


Figure 4.21 Representation of scatter plot (a-b), error radar plot (c-d) and error histogram plot (e-f) for the MEP_{K-1} and XGB_{K-1} models on the literature datasets

4.4.4.3 External validation of literature models on present study datasets

The external validation of literature study models was performed using the present study dataset. Kin (2006), Taskiran (2010), Yildirim and Gunaydin (2011) and Bardhan, Gokceoglu, et al. (2021) models from the various country (as shown in Table 4.30), were selected for the analysis. For this purpose, only those models were selected which are having input parameters similar to the present study geotechnical parameters. For these models, datasets from the present study were selected as per the minimum and maximum values of their input and output parameters.

Table 4.30 Literature models validation on the present study datasets

S. No.	Literature model	Soil origin	Dataset from the present study
1	Kin (2006)	Malaysia	997
2	Taskiran (2010)	Turkey	1011
3	Yildirim and Gunaydin (2011)	Turkey	1011
4	Bardhan, Gokceoglu, et al. (2021)	Indian	610

Table 4.31 exhibits the comparative performance of literature models on the present study dataset in terms of various statistical performance indices. It is clearly observed from Table 4.31 that the literature models are inadequate in predicting the

soaked CBR value of the present study dataset. 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, therefore, leading to a worse fit than the horizontal line.

Table 4.31 Comparative performance of literature models on present study datasets

Statistical performance indices	Literature models				
	Kin (2006)	Taskiran (2010)	Yildirim and Gunaydin (2011)	Bardhan, Gokceoglu, et al. (2021)	Bardhan, Gokceoglu, et al. (2021)
				MARS-L	GP
R^2	-14.5753	-99.5535	-5.4068	-5.0918	-4.7944
R	0.1951	0.5089	0.1114	0.0036	0.1006
MAE	3.8447	9.6759	2.5414	1.5096	1.5594
RMSE	4.3462	11.6557	2.9421	1.7511	1.7078
VAF	-284.1402	-3027.6911	-64.8796	-94.4940	-8.7940
IOA	0.2973	0.1609	0.3506	0.3604	0.3866
IOS	0.8239	0.6235	0.2546	0.2240	0.2211
a20-index	0.1615	0.0079	0.3541	0.6525	0.6705

4.4.4.4 Influence of soil origin on the predictive ability of various models

Section 4.4.4.2 and 4.4.4.3 present the validation process of present study models on the literature dataset and literature model on the present study dataset, respectively. The present study model, developed through Indian origin soil, was attempted to validate on other country soil, and unsatisfactory results were obtained (as shown in Table 4.29). Similarly, when the literature models, developed for their respective country soil, were tried to validate on the Indian soil, they also indicate the insignificant results (see Table 4.31). Although each of the model in Table 4.31 demonstrate worse results, the model developed by Bardhan, Gokceoglu, et al. (2021) indicate slightly good results in comparison to Kin (2006), Taskiran (2010) and Yildirim and Gunaydin (2011) model. The Bardhan, Gokceoglu, et al. (2021) model is able to predict more than 65% observations within ±20% variations, which is quite very high than Kin (2006), Taskiran (2010) and Yildirim and Gunaydin (2011) model. This is because the dataset used for the

analysis and model development belongs to the same country, i.e. India. Nagaraj and Suresh (2018) also shows that soils are likely to be quite variable depending on their geological locations. Therefore, it is established from these investigations that model developed for the soil collected from a particular geological location may or may not be suitable for other geological locations. However, from the author point of view, the more generalized model could be developed by combining the dataset collected from different geological locations.

4.5 GRAPHICAL USER INTERFACE (GUI) FOR PREDICTING THE SOAKED CBR VALUE

A graphical user interface (GUI) is a system of representing the features visually in the form of computer software or tool. It is a simple and easy method to display the components according to user requirements. In this study, a reliable GUI was designed in python language for the XGB_{K-1} model. The designed interface was designated as "*CBR prediction tool for Fine-grained plastic soils*" where CBR referred to California Bearing Ratio (test method for which prediction is being performed). Figure 4.22 depicts the structure of the designed interface. Initially, the values for input parameters i.e., sand content (%), fine content (%), plastic limit (%), plasticity index (%), maximum dry density (g/cc) and optimum moisture content (OMC) are inserted. Ultimately, the predicted CBR value can be achieved directly after clicking the Run button. The developed interface is not only beneficial for the researchers but much user friendly for the site engineers.

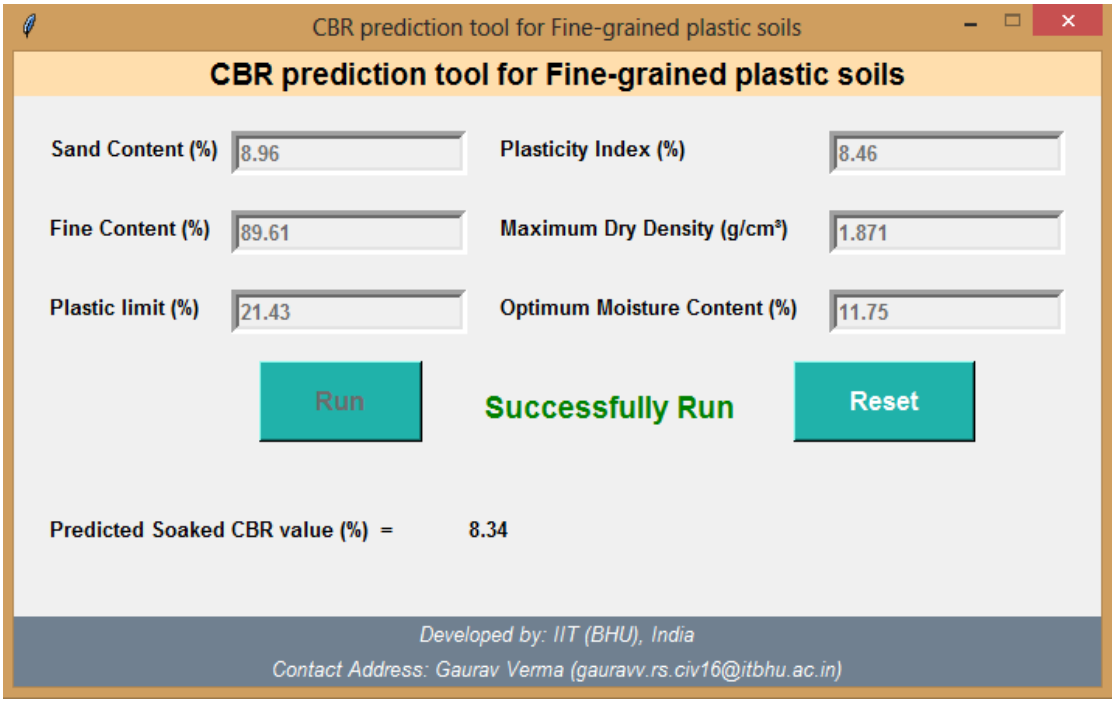


Figure 4.22 Designed GUI for predicting the soaked CBR of fine-grained plastic soils