A Study of Relationship among Correlation Coefficient, Performance, and Overfitting using Regression Analysis

Jitendra Khatti, Dr. Kamaldeep Singh Grover

Abstract— The correlation coefficient is the method for presenting the relationship between features and labels. The present research work aims to map a relationship among correlation coefficient, performance, and overfitting of the regression models. To carry out the present study, the consistency limits and compaction parameters of soil are used. The datasets are collected from the published articles. Furthermore, the datasets are divided into training, testing, and validation dataset. The training datasets are subdivided from 50% to 100% at 10% intervals. The simple linear regression, simple polynomial regression, and multiple linear regression analyses are performed to study the correlation coefficient and performance relationship. The results show that the correlation coefficient influences the performance of regression models. The moderately (0.61≥CC≤0.80) correlated datasets predict consistency limits of soil with better performance. The excellent prediction of OMC and MDD can be achieved if very strongly (0.81≥CC≤1.0) correlated consistency limits features and strongly (0.61≥CC≤0.80) correlated sand and fine content features exist in the dataset. It is concluded that the overfitting of regression models is decreased because of very strongly correlated datasets. The sensitivity analysis shows that consistency limits are highly influenced by fine content, and compaction parameters are highly influenced by fine content, LL, PL, and PI.

Index Terms— Correlation coefficient, Polynomial Regression, Consistency Limits, Compaction Parameters, Multiple Linear Regression Anaylsis, Soil Properties,

1 Introduction

In India, majorly five types of soil are found, i.e., alluvial deposits, black cotton soil, lateritic soils, desert soils, and marine deposits [3]. Every soil has consistency limits, compaction parameters, and strength parameters. The liquid limit, plasticity index, plastic limits are the consistency limits or Atterberg's limits. The liquid limit is determined experimentally using the Casagrande and cone penetration methods. The plastic limit of soil is determined by preparing 3mm thread at suitable water content. The optimum moisture content (OMC) and maximum dry density (MDD) are compaction parameters of soil. The standard and modified proctor test is performed to determine the optimum moisture content and maximum dry density of soil. The most popular method for determining OMC and MDD is the standard proctor test or light compaction test. The consistency limits and compaction parameters play an important role in every civil engineering project. The soil is classified into cohesive and cohesionless soil. Cohesive soils have high consistency limits than cohesionless soil.

The laboratory procedures for determining consistency limits and compaction parameters of soil are time-consuming. Therefore, numerous researchers evolved different methodologies and applied them to compute consistency limits and compaction parameters. The multi expression programming predicts the compaction parameters with training performances of 0.916 (for OMC) & 0.872 (for MDD) and testing performance of 0.923 (for OMC) & 0.858 (for MDD) if optimum moisture content has COD of 0.263, 0.324, 0.48, 0.164, 0.482, & 0.027 with G, S, FC, LL, PL, & E, respectively, and maximum dry density has COD of 0.304, 0.265, 0.447, 0.243, 0.541, and 0.031 with G, S, FC, LL, PL, and E, respectively. The fine content and plastic limit affect the OMC of soil [15]. The support vector machine predicts the compaction parameters of soil if the liquid limit is strongly related to MDD than OMC [6]. The OMC increases with LL, but MDD decreases. The PL has linearity with OMC and MDD but unlike the liquid limit. Therefore, the LL is essential for predicting OMC and MDD [4]. The plasticity index of soil increases with clay content. Therefore, the plasticity index significantly impacts the OMC and MDD of soil [8].

The maximum dry density and optimum moisture content have a correlation coefficient of 0.70-0.90 and 0.74-0.92 with liquid limit and plastic limit, respectively. Therefore, the GMDH-ANN model predicts the OMC and MDD with a testing performance of 0.96 and 0.93, respectively. Furthermore, the cosine amplitude method shows that the LL and plastic limit influence the OMC and MDD [2]. The input parameters coefficient of uniformity and D$_{30}$ of soil has COD of 0.74 and 0.81 for OMC and MDD, respectively, in multiple regression analysis [11]. The artificial neural network predicts OMC and MDD of soil with COD of 0.8691 and 0.8531, respectively [12]. The S, FC, D$_{50}$, and C$_u$ have a correlation coefficient of 0.447, 0.431, 0.332, 0.774, and 0.42 with maximum dry density. Similarly, S, Fc, D$_{50}$, and C$_u$ have a correlation coefficient of 0.309, 0.284, 0.205, 0.455 with optimum moisture content. The sand content is susceptible to OMC and MDD of soil. Therefore, the MARS approach performs better than the SVM and ANN in predicting OMC and MDD of soil [9]. The input parameters LL, PI, and E, predict the OMC and MDD of soil with ±9.5%

- *Mr. Jitendra Khatti is pursuing PhD in geotechnical engineering at Department of Civil Engineering, Rajasthan Technical University, Kota-324010, Rajasthan, Ph: +91 94138-26197.*
  E-mail: jitendrakhatti197@gmail.com
- *Dr. Kamaldeep Singh Grover is a Professor at Department of Civil Engineering, Rajasthan Technical University, Kota-324010, Rajasthan.*
  E-mail: ksgrover@rtu.ac.in

and ±2.5% accuracy at a 95% confidence interval [7]. The input parameters sand, silt, and clay content predict the LL, PL, and PI of soil using ANN and ANFIS. The ANN model predicts LL, PL, PI with COD of 0.928, 0.974, and 0.976. Thus, ANFIS predicts LL, PL, and PI better than artificial neural networks [10]. The prediction of soil properties using artificial neural networks depends on datasets' quantity and quality [1]. The multiple regression equation of gravel, S, FC, LL, and PI has 0.987 and 0.999 correlation coefficients (R) with OMC and MDD. Therefore, the GEP predicts OMC and MDD better than the MLR approach [13]. The literature shows that the relationship among correlation coefficient, performance and overfitting has not been mapped using regression analysis and for regression models. Thus, the novelty of the present study is

(i) Determine the correlation coefficient for 50%, 60%, 70%, 80%, 90%, and 100% of 190 training dataset.
(ii) Determine the effect of the different training datasets on has been studied in predicting the consistency limits and compaction parameters of soil.
(ii) The effect of the correlation coefficient has been studied for the performance of regression models.
(iii) The effect of the correlation coefficient has been studied for the overfitting of regression models.

2 Data Collection and Analysis

The data source, division of training datasets, descriptive statistics, and Pearson's product-moment correlation coefficient are discussed in this section.

2.1 Data Source

In this research work, 332 datasets were collected from different sources. The datasets are divided into training, testing, and validation of the models. The detail of the data source is given in Table 1.

Table 1 – Data Sources

Data Type	Description	Numbers
Training	Benson C. H. et al. (1994)	67
	Benson C. H. et al. (1995)	13
	Najjar Y. M. et al. (1996)	47
	Nagaraj H. B. et al. (2014)	44
	O. Gunaydin (2008)	73
	NG. K. S. (2015)	09
Testing	O. Gunaydin (2008)	53
Validation	Ibrahim A. O. (2013), Vukićević M. (2013), Khalid F. (2015), Senol A. (2002), Wathiq Al-Jaban (2019), Tuncer B. (2006), Kamal M. H. I. I. (2014). Erdem O. T. (2005), Tayel El-Hasan (2014), IHRB Project (2005), Kawther Y. (2018a), Kawther Y (2018b), Nahal S. (2018a, 2018b), Xiaobin Z. (2020), K. V. Manjunath (2012), Rana A (2020), Fattah M. Y. (2013), Wathiq Al-Jabban (2017), Hawkar H. I. (2019), Rizgar A. (2020), Guoqid X. (2021)	26

2.2 Preprocessing of Dataset

Preprocessing is a method of removing missing data and outliers from the datasets and transforming raw data into an understandable form. The min-max normalization function has transformed the datasets.

2.3 Descriptive Statistics

In the present research work, the minimum, maximum, mean (average), standard deviation (St. Dev), and confidence interval (CL) at 95% is determined for each feature of the dataset. The descriptive statistics of training datasets of consistency limits with OMC & MDD are shown in Table 2.

Table 2 – Descriptive statistics of training datasets

Features	Min	Max	Mean	StDev	CL	Min	Max	Mean	StDev	CL	Min	Max	Mean	StDev	CL
-	50% Training Datasets					60% Training Datasets					70% Training Datasets				
S (%)	3.02	70.28	30.71	18.14	3.70	3.02	70.28	28.59	17.37	3.22	3.02	70.28	28.54	16.92	2.90
FC (%)	25.65	96.98	67.15	19.41	3.95	25.65	96.98	69.32	18.54	3.44	25.65	96.98	69.15	18.01	3.09
S:FC	0.03	2.36	0.61	0.56	0.11	0.03	2.36	0.55	0.54	0.10	0.03	2.36	0.53	0.51	0.09
LL (%)	21.58	62.21	34.85	9.87	2.01	21.97	65.13	36.81	10.56	1.96	21.34	65.13	35.48	10.67	1.83
PL (%)	5.27	25.91	14.00	4.76	0.97	5.23	29.46	14.72	5.29	0.98	4.63	29.46	14.03	5.34	0.92
PI (%)	13.74	38.72	20.85	5.53	1.13	14.94	38.72	22.09	5.86	1.09	13.74	38.72	21.45	5.87	1.01
OMC (%)	9.30	30.40	15.52	4.32	0.88	9.00	30.40	16.10	4.60	0.85	9.00	30.40	15.89	4.56	0.78
MDD (g/cc)	1.44	2.01	1.77	0.12	0.02	1.44	2.01	1.75	0.13	0.02	1.44	2.01	1.75	0.13	0.02
-	80% Training Datasets					90% Training Datasets					100% Training Datasets				
S (%)	3.02	70.28	31.15	17.60	2.82	3.02	70.28	29.53	18.14	2.74	3.02	70.28	29.29	17.29	2.47
FC (%)	25.65	96.98	66.72	18.46	2.96	25.65	96.98	68.54	19.04	2.87	25.65	96.98	68.78	18.08	2.59
S:FC	0.03	2.36	0.61	0.54	0.09	0.03	2.36	0.57	0.53	0.08	0.03	2.36	0.55	0.51	0.07
LL (%)	21.34	65.13	34.51	10.35	1.66	21.34	65.13	35.68	10.63	1.60	21.34	65.13	35.41	10.40	1.49
PL (%)	4.63	29.46	13.70	5.34	0.86	4.63	29.46	14.17	5.41	0.82	4.63	29.46	14.03	5.28	0.75
PI (%)	13.74	38.72	20.81	5.47	0.88	13.74	38.72	21.51	5.73	0.87	13.74	38.72	21.38	5.64	0.81
OMC (%)	9.00	30.40	15.25	4.22	0.68	9.00	30.40	15.81	4.45	0.67	9.00	30.40	15.74	4.33	0.62
MDD (g/cc)	1.45	2.01	1.77	0.12	0.02	1.44	2.01	1.76	0.13	0.02	1.44	2.01	1.76	0.12	0.02

2.4 Division of Training Dataset

The training datasets are divided from 50% to 100% at 10% intervals to study the relationship among correlation coefficient, performance and overfitting of regression models, as shown in Table 3.

TABLE 3 – DIVISION OF TRAINING DATASET

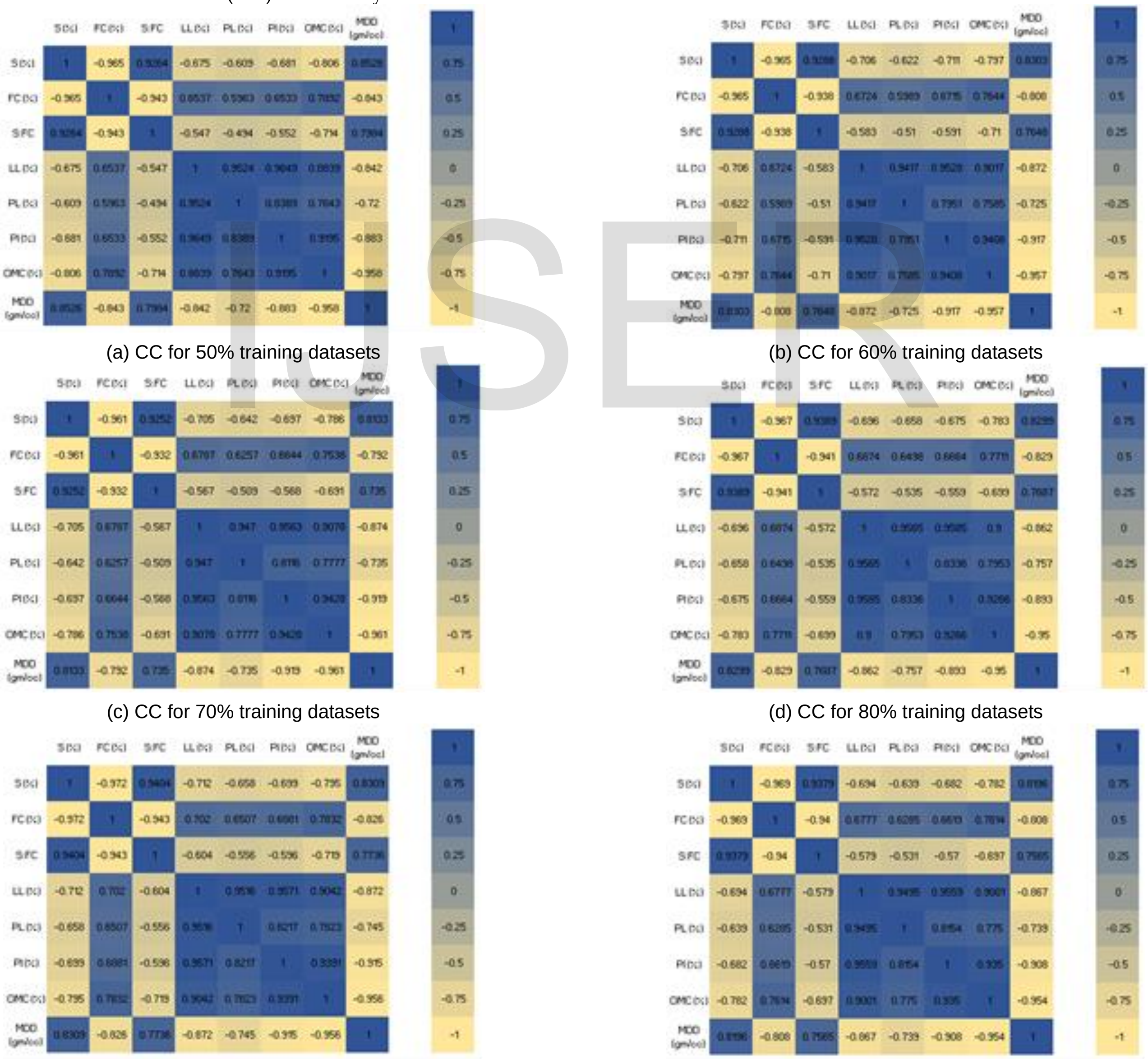
Percentage	Training Data
50%	95
60%	114
70%	133
80%	152
90%	171
100%	190

2.5 Pearson's Correlation Coefficient

The correlation coefficient (CC) is the way to determine the strength of the linear relationship between independent and dependent variables. The Linear or curvilinear correlation, scatter diagram method, Pearson's product-moment correlation coefficient, spearman's rand correlation coefficient are the methods for determining correlation coefficient or relationship. The relationship of the pair of datasets according to the range of correlation coefficients is given in Table 4 [5].

TABLE 4 – RELATIONSHIP LEVEL

Correlation Coefficient	Relationship Level
±0.81 - ±1.00	Very Strong
±0.61 - ±0.80	Strong
±0.41 - ±0.60	Moderate
±0.21 - ±0.40	Weak
±0.00 - ±0.20	No relationship



(a) CC for 50% training datasets

(b) CC for 60% training datasets

(c) CC for 70% training datasets

(d) CC for 80% training datasets

(e) CC for 90% training datasets

(f) CC for 100% training datasets

Fig. 1. Correlation coefficient for different percentages of training datasets

Fig. 2. depicts the correlation coefficient for LL, PL, PI, OMC, and MDD. The consistency limits (LL, PI, PL) get affected by sand and fine content. Therefore, sand and fine contents are selected as input parameters to predict LL, PI, and PL. Similarly, S, FC, LL, PL, and PI are selected as input parameters to predict OMC and MDD of soil. Fig. 2 shows that the LL, PL, and PI strongly correlate with sand (S) and fine content (FC) for 50-100% datasets. Fig. 2 also shows that OMC has a strong to very strong relationship with input parameters. Similarly, Fig. 2 also depicts that MDD strongly relationship with input parameters. It is also observed that the correlation coefficient or level of relationship increases by increasing datasets.

3 Methodologies

The simple linear and polynomial regression analysis is performed to determine the effect of the correlation coefficient on the performance and overfitting of the regression model. In addition, the predictive multiple linear regression models are also constructed to study the effect of multi-input parameters on the performance and overfitting of regression models. The developed simple linear and polynomial regression equations for consistency limits are given in Table 5.

Table 5 – Equations derived from simple linear and polynomial regression analysis for LL, PL, PI

Data	Input	Simple Linear Equations	R²	Eq.	Simple Polynomial Equations	R²	Eq.
50%	S	LL′ = -0.3674*S + 46.131	0.4561	(2)	LL′ = 0.0089*S2 - 0.9473*S + 52.708	0.5334	(38)
	FC	LL′ = 0.3324*FC + 12.529	0.4273	(3)	LL′ = 0.0079*FC2 - 0.6829*FC + 42.216	0.5198	(39)
	S	PL′ = -0.1597*S + 18.904	0.3706	(4)	PL′ = 0.0036*S2 - 0.3949*S + 21.572	0.4252	(40)
	FC	PL′ = 0.1463*FC + 4.1771	0.3556	(5)	PL′= 0.0036*FC2 - 0.3171*FC + 17.728	0.4384	(41)
	S	PI′ = -0.2077*S + 27.227	0.4643	(6)	PI′ = 0.0053*S2 - 0.5523*S + 31.137	0.5512	(42)
	FC	PI′ = 0.1861*FC + 8.3519	0.4268	(7)	PI′ = 0.0043*FC2 - 0.3657*FC + 24.488	0.5139	(43)
60%	S	LL′ = -0.4293*S + 49.084	0.4982	(8)	LL′ = 0.0083*S2 - 0.9815*S + 55.636	0.5509	(44)
	FC	LL′ = 0.3832*FC + 10.252	0.4521	(9)	LL′ = 0.007*FC2 - 0.5126*FC + 36.371	0.5082	(45)
	S	PL′ = -0.1893*S + 20.13	0.3869	(10)	PL′ = 0.0036*S2 - 0.4273*S + 22.954	0.4260	(46)
	FC	PL′ = 0.1708*FC + 2.8801	0.3587	(11)	PL′ = 0.0034*FC2 - 0.2675*FC + 15.658	0.4123	(47)
	S	PI′ = -0.24*S + 28.954	0.5056	(12)	PI′ = 0.0047*S2 - 0.5542*S + 32.683	0.5609	(48)
	FC	PI′ = 0.2124*FC + 7.3718	0.4509	(13)	PI′ = 0.0036*FC2 - 0.2452*FC + 20.712	0.4985	(49)
70%	S	LL′ = -0.4446*S + 48.168	0.4967	(14)	LL′ = 0.0102*S2 - 1.1032*S + 55.777	0.5699	(50)
	FC	LL′ = 0.4021*FC + 7.6788	0.4606	(15)	LL′ = 0.0086*FC2 - 0.7205*FC + 41.252	0.5366	(51)
	S	PL′ = -0.2028*S + 19.82	0.4125	(16)	PL′ = 0.0047*S2 - 0.5078*S + 23.343	0.4751	(52)
	FC	PL′ = 0.1856*FC + 1.1992	0.3915	(17)	PL′ = 0.0044*FC2 - 0.3895*FC + 18.4	0.4712	(53)
	S	PI′ = -0.2418*S + 28.349	0.4856	(18)	PI′ = 0.0055*S2 - 0.5954*S + 32.434	0.5554	(54)
	FC	PI′ = 0.2165*FC + 6.4796	0.4415	(19)	PI′ = 0.0042*FC2 - 0.3309*FC + 22.852	0.5013	(55)
80%	S	LL′ = -0.4092*S + 47.255	0.4844	(20)	LL′ = 0.0107*S2 - 1.1459*S + 56.472	0.5721	(56)
	FC	LL′ = 0.3854*FC + 8.7933	0.4725	(21)	LL′ = 0.0096*FC2 - 0.8389*FC + 44.491	0.5658	(57)
	S	PL′ = -0.1996*S + 19.911	0.4326	(22)	PL′ = 0.0051*S2 - 0.5471*S + 24.26	0.5059	(58)
	FC	PL′ = 0.188*FC + 1.1518	0.4222	(23)	PL′ = 0.0049*FC2 - 0.4346*FC + 19.305	0.5127	(59)
	S	PI′ = -0.2096*S + 27.343	0.4555	(24)	PI′ = 0.0057*S2 - 0.5988*S + 32.212	0.5432	(60)
	FC	PI′ = 0.1974*FC + 7.6415	0.4441	(25)	PI′ = 0.0047*FC2 - 0.4043*FC + 25.186	0.5248	(61)
90%	S	LL′ = -0.417*S + 47.995	0.5065	(26)	LL′ = 0.0085*S2 - 0.9801*S + 54.403	0.5576	(62)
	FC	LL′ = 0.3919*FC + 8.8183	0.4927	(27)	LL′ = 0.0078*FC2 - 0.6262*FC + 39.066	0.5497	(63)
	S	PL′ = -0.1962*S + 19.964	0.4332	(28)	PL′ = 0.0037*S2 - 0.4407*S + 22.746	0.4704	(64)
	FC	PL′ = 0.1848*FC + 1.504	0.4234	(29)	PL′ = 0.0038*FC2 - 0.3069*FC + 16.111	0.4748	(65)
	S	PI′ = -0.2209*S + 28.031	0.4888	(30)	PI′ = 0.0048*S2 - 0.5394*S + 31.657	0.5452	(66)
	FC	PI′ = 0.2071*FC + 7.3143	0.4735	(31)	PI′ = 0.004*FC2 - 0.3193*FC + 22.955	0.5260	(67)
100%	S	LL′ = -0.4175*S + 47.636	0.4817	(32)	LL′ = 0.0092*S2 - 1.0244*S + 54.806	0.5440	(68)
	FC	LL′ = 0.3898*FC + 8.5949	0.4593	(33)	LL′ = 0.0081*FC2 - 0.6662*FC + 40.176	0.5234	(69)
	S	PL′ = -0.1949*S + 19.742	0.4080	(34)	PL′ = 0.0039*S2 - 0.4559*S + 22.825	0.4528	(70)
	FC	PL′ = 0.1834*FC + 1.4182	0.3950	(35)	PL′ = 0.0039*FC2 - 0.3268*FC + 16.676	0.4532	(71)
	S	PI′ = -0.2226*S + 27.894	0.4656	(36)	PI′ = 0.0052*S2 - 0.5685*S + 31.981	0.5345	(72)
	FC	PI′ = 0.2064*FC + 7.1767	0.4380	(37)	PI′ = 0.0042*FC2 - 0.3394*FC + 23.499	0.4963	(73)

Table 5 shows that the coefficient of determination (R^2) increases by increasing the percentage of datasets for simple linear regression analysis. Similarly, the simple polynomial regression analysis shows that the coefficient of determination (R^2) increased by increasing the percentage of datasets in the case of LL, PI, and PL. The comparison of R^2 of simple linear and polynomial regression analysis shows that the nonlinear analysis gives better results than linear analysis for consistency limits. Similarly, the developed simple linear and polynomial regression equations to predict OMC are given in Table 6.

TABLE 6. EQUATIONS DERIVED FROM SIMPLE LINEAR AND POLYNOMIAL REGRESSION ANALYSIS FOR OMC

Data	Input	Simple Linear Equations	R^2	Eq.	Simple Polynomial Equations	R^2	Eq.
50%	S	OMC′ = -0.1917*S + 21.405	0.6488	(74)	OMC′ = 0.0028*S2 - 0.3726*S + 23.457	0.6882	(104)
	FC	OMC′ = 0.1755*FC + 3.7323	0.6228	(75)	OMC′ = 0.002*FC2 - 0.0853*FC + 11.36	0.6547	(105)
	LL	OMC′ = 0.3866*LL + 2.0459	0.7812	(76)	OMC′ = 0.0063*LL2 - 0.1069*LL + 10.974	0.8057	(106)
	PL	OMC′ = 0.6931*PL + 5.816	0.5842	(77)	OMC′ = 0.0376*PL2 - 0.4228*PL + 13.222	0.6397	(107)
	PI	OMC′ = 0.7179*PI + 0.5504	0.8455	(78)	OMC′ = 0.0087*PI2 + 0.2992*PI + 5.2561	0.8506	(108)
60%	S	OMC′ = -0.2108*S + 22.123	0.6348	(79)	OMC′ = 0.0024*S2 - 0.3678*S + 23.985	0.6573	(109)
	FC	OMC′ = 0.1895*FC + 2.9592	0.5843	(80)	OMC′ = 0.0015*FC2 - 0.0034*FC + 8.5847	0.5980	(110)
	LL	OMC′ = 0.3923*LL + 1.6553	0.8130	(81)	OMC′ = 0.0051*LL2 - 0.0089*LL + 9.0002	0.8293	(111)
	PL	OMC′ = 0.6595*PL + 6.3892	0.5753	(82)	OMC′ = 0.0216*PL2 + 0.0085*PL + 10.683	0.6003	(112)
	PI	OMC′ = 0.7375*PI - 0.1982	0.8851	(83)	OMC′ = 0.0053*PI2 + 0.4779*PI + 2.7848	0.8866	(113)
70%	S	OMC′ = -0.2119*S + 21.938	0.6171	(84)	OMC′ = 0.0028*S2 - 0.3899*S + 23.995	0.6464	(114)
	FC	OMC′ = 0.1909*FC + 2.688	0.5682	(85)	OMC′ = 0.0018*FC2 - 0.0463*FC + 9.7846	0.5868	(115)
	LL	OMC′ = 0.3881*LL + 2.1197	0.8241	(86)	OMC′ = 0.0047*LL2 + 0.0136*LL + 8.901	0.8381	(116)
	PL	OMC′ = 0.6642*PL + 6.5716	0.6048	(87)	OMC′ = 0.0219*PL2 + 0.0067*PL + 10.86	0.6306	(117)
	PI	OMC′ = 0.7329*PI + 0.1704	0.8888	(88)	OMC′ = 0.0075*PI2 + 0.3665*PI + 4.3363	0.8920	(118)
80%	S	OMC′ = -0.1876*S + 21.098	0.6131	(89)	OMC′ = 0.0029*S2 - 0.388*S + 23.605	0.6521	(119)
	FC	OMC′ = 0.1762*FC + 3.495	0.5946	(90)	OMC′ = 0.0021*FC2 - 0.0856*FC + 11.13	0.6203	(120)
	LL	OMC′ = 0.3668*LL + 2.5939	0.8099	(91)	OMC′ = 0.005*LL2 - 0.0216*LL + 9.5632	0.8290	(121)
	PL	OMC′ = 0.6282*PL + 6.6496	0.6325	(92)	OMC′ = 0.0259*PL2 - 0.1434*PL + 11.626	0.6765	(122)
	PI	OMC′ = 0.7149*PI + 0.374	0.8586	(93)	OMC′ = 0.0083*PI2 + 0.3145*PI + 4.8645	0.8628	(123)
90%	S	OMC′ = -0.1948*S + 21.561	0.6316	(94)	OMC′ = 0.0025*S2 - 0.3593*S + 23.433	0.6566	(124)
	FC	OMC′ = 0.1829*FC + 3.2715	0.6134	(95)	OMC′ = 0.0018*FC2 - 0.0451*FC + 10.045	0.6297	(125)
	LL	OMC′ = 0.3782*LL + 2.3132	0.8175	(96)	OMC′ = 0.004*LL2 + 0.0614*LL + 8.03	0.8289	(126)
	PL	OMC′ = 0.6433*PL + 6.6922	0.6120	(97)	OMC′ = 0.0205*PL2 + 0.0297*PL + 10.679	0.6364	(127)
	PI	OMC′ = 0.7287*PI + 0.1339	0.8819	(98)	OMC′ = 0.0058*PI2 + 0.4481*PI + 3.3076	0.8838	(128)
100%	S	OMC′ = -0.1959*S + 21.478	0.6171	(99)	OMC′ = 0.0028*S2 - 0.3815*S + 23.671	0.6449	(129)
	FC	OMC′ = 0.1825*FC + 3.1906	0.5682	(100)	OMC′ = 0.0019*FC2 - 0.0664*FC + 10.631	0.6002	(130)
	LL	OMC′ = 0.375*LL + 2.462	0.8241	(101)	OMC′ = 0.0045*LL2 + 0.0202*LL + 8.857	0.8247	(131)
	PL	OMC′ = 0.6364*PL + 6.8091	0.6048	(102)	OMC′ = 0.0211*PL2 + 0.0088*PL + 10.88	0.6265	(132)
	PI	OMC′ = 0.7183*PI + 0.3851	0.8888	(103)	OMC′ = 0.0074*PI2 + 0.3607*PI + 4.4361	0.8774	(133)

From Table 6, the following points are observed, (i) the sand content, liquid limit, and plasticity index has a strong to the very strong relationship with optimum moisture content for simple linear regression analysis, (ii) The fine content and plastic limit have moderate to strong relationship with optimum moisture content for simple linear regression analysis, (iii) the liquid limit and plasticity index has very strong relationship with optimum moisture content for simple polynomial regression analysis, (iv) the sand content, fine content, and plastic limit have a strong relationship with OMC for simple polynomial regression analysis. The comparison of coefficient of determination also shows that the nonlinear regression analysis gives better training performance than linear regression analysis. Thus, the developed simple linear and polynomial regression equations for predicting MDD are given in Table 7.

From Table 7, the following points are observed, (i) the sand, fine content, liquid limit, and plasticity index are strongly correlated with a maximum dry density of soil, (ii) the plastic limit is moderately correlated with a maximum dry density of soil.

The multiple linear regression analysis is also performed to predict the LL, PL, PI, OMC, and MDD of soil and determine the effect of correlation coefficient on performance and overfitting of multiple linear regression model. The following equations are derived from the multiple linear regression analysis at different percentages of training datasets.

50% Training Dataset

$LL = -0.3513*S + 0.0156*FC + 44.59$	$R^2 = 0.4562$	(194)
$PL = -0.1266*S + 0.0321*FC + 15.73$	$R^2 = 0.3717$	(195)
$PI = -0.2247*S - 0.0165*FC + 28.85$	$R^2 = 0.4645$	(196)
$OMC = -0.0282*S + 0.0507*FC + 0.5884*LL - 0.6572*PL + 0*PI + 1.67$	$R^2 = 0.9102$	(197)
$MDD = 0.0009*S - 0.0022*FC - 0.0148*LL + 0.0183*PL + 0*PI + 2.14$	$R^2 = 0.9097$	(198)

60% Training Dataset

$LL = -0.503*S - 0.0716*FC + 56.15$	$R^2 = 0.4993$	(199)
$PL = -0.1946*S - 0.0051*FC + 20.63$	$R^2 = 0.3869$	(200)
$PI = -0.3084*S - 0.0664*FC + 35.51$	$R^2 = 0.5086$	(201)
$OMC = -0.0432*S + 0.0244*FC + 0.6035*LL - 0.6157*PL + 0*PI + 2.48$	$R^2 = 0.9188$	(202)
$MDD = 0.0009*S - 0.0017*FC - 0.0156*LL + 0.0174*PL + 0*PI + 2.15$	$R^2 = 0.9113$	(203)

TABLE 7. EQUATIONS DERIVED FROM SIMPLE LINEAR AND POLYNOMIAL REGRESSION ANALYSIS FOR MDD

Data	Input	Simple Linear Equations	R^2	Eq.	Simple Polynomial Equations	R^2	Eq.
50%	S	MDD′ = 0.0057*S + 1.5976	0.7269	(134)	MDD′ = -3E-05*S2 + 0.0078*S + 1.5738	0.7337	(164)
	FC	MDD′ = -0.0052*FC + 2.122	0.7100	(135)	MDD′ = -2E-05*FC2 - 0.0023*FC + 2.036	0.7153	(165)
	LL	MDD′ = -0.0103*LL + 2.129	0.7088	(136)	MDD′ = -0.0001*LL2 - 2E-05*LL + 1.944	0.7224	(166)
	PL	MDD′ = -0.0182*PL + 2.026	0.5190	(137)	MDD′ = -0.001*PL2 + 0.0106*PL + 1.8351	0.5665	(167)
	PI	MDD′ = -0.0192*PI + 2.1717	0.7788	(138)	MDD′ = -8E-05*PI2 - 0.0155*PI + 2.1304	0.7793	(168)
60%	S	MDD′ = 0.0061*S + 1.5717	0.6893	(139)	MDD′ = -4E-05*S2 + 0.009*S + 1.5364	0.6999	(169)
	FC	MDD′ = -0.0055*FC + 2.128	0.6527	(140)	MDD′ = -2E-05*FC2 - 0.0029*FC + 2.051	0.6562	(170)
	LL	MDD′ = -0.0105*LL + 2.131	0.7597	(141)	MDD′ = -8E-05*LL2 - 0.0039*LL + 2.010	0.7655	(171)
	PL	MDD′ = -0.0174*PL + 2.001	0.5252	(142)	MDD′ = -0.0006*PL2 + 0.0001*PL + 1.886	0.5490	(172)
	PI	MDD′ = -0.0198*PI + 2.1836	0.8412	(143)	MDD′ = 0.0002*PI2 - 0.0275*PI + 2.2718	0.8430	(173)
70%	S	MDD′ = 0.0061*S + 1.5802	0.6615	(144)	MDD′ = -6E-05*S2 + 0.0101*S + 1.5343	0.6803	(174)
	FC	MDD′ = -0.0056*FC + 2.139	0.6271	(145)	MDD′ = -3E-05*FC2 - 0.0013*FC + 2.010	0.6351	(175)
	LL	MDD′ = -0.0104*LL + 2.122	0.7630	(146)	MDD′ = -9E-05*LL2 - 0.0034*LL + 1.997	0.7692	(176)
	PL	MDD′ = -0.0174*PL + 1.999	0.5400	(147)	MDD′ = -0.0006*PL2 - 0.0002*PL + 1.887	0.5628	(177)
	PI	MDD′ = -0.0199*PI + 2.18	0.8452	(148)	MDD′ = 1E-05*PI2 - 0.0203*PI + 2.1856	0.8452	(178)
80%	S	MDD′ = 0.0056*S + 1.5966	0.6887	(149)	MDD′ = -5E-05*S2 + 0.0093*S + 1.5499	0.7057	(179)
	FC	MDD′ = -0.0053*FC + 2.128	0.6871	(150)	MDD′ = -3E-05*FC2 - 0.0012*FC + 2.006	0.6954	(180)
	LL	MDD′ = -0.0099*LL + 2.114	0.7437	(151)	MDD′ = -8E-05*LL2 - 0.0037*LL + 2.002	0.7495	(181)
	PL	MDD′ = -0.0169*PL + 2.003	0.5733	(152)	MDD′ = -0.0007*PL2 + 0.0033*PL + 1.873	0.6109	(182)
	PI	MDD′ = -0.0194*PI + 2.176	0.7969	(153)	MDD′ = 8E-05*PI2 - 0.0234*PI + 2.2208	0.7975	(183)
90%	S	MDD′ = 0.0057*S + 1.5883	0.6903	(154)	MDD′ = -5E-05*S2 + 0.0089*S + 1.5528	0.7016	(184)
	FC	MDD′ = -0.0054*FC + 2.13	0.6816	(155)	MDD′ = -3E-05*FC2 - 0.002*FC + 2.0273	0.6863	(185)
	LL	MDD′ = -0.0103*LL + 2.124	0.7606	(156)	MDD′ = -8E-05*LL2 - 0.0043*LL + 2.016	0.7657	(186)
	PL	MDD′ = -0.0173*PL + 2.002	0.5555	(157)	MDD′ = -0.0006*PL2 + 6E-05*PL + 1.889	0.5801	(187)
	PI	MDD′ = -0.02*PI + 2.1877	0.8363	(158)	MDD′ = 6E-05*PI2 - 0.0231*PI + 2.2224	0.8366	(188)
100%	S	MDD′ = 0.0058*S + 1.589	0.6717	(159)	MDD′ = -5E-05*S2 + 0.0091*S + 1.549	0.6859	(189)
	FC	MDD′ = -0.0054*FC + 2.131	0.6528	(160)	MDD′ = -3E-05*FC2 - 0.0018*FC + 2.022	0.6584	(190)
	LL	MDD′ = -0.0101*LL + 2.117	0.7523	(161)	MDD′ = -9E-05*LL2 - 0.0034*LL + 1.995	0.7589	(191)
	PL	MDD′ = -0.017*PL + 1.997	0.5457	(162)	MDD′ = -0.0006*PL2 + 0.0002*PL + 1.885	0.5703	(192)
	PI	MDD′ = -0.0196*PI + 2.1763	0.8252	(163)	MDD′ = 1E-05*PI2 - 0.0203*PI + 2.184	0.8253	(193)

70% Training Dataset

$$LL = -0.4325*S + 0.0118*FC + 47.00 \quad R^2 = 0.4967 \quad (204)$$

$$PL = -0.1686*S + 0.0334*FC + 16.53 \quad R^2 = 0.4134 \quad (205)$$

$$PI = -0.2639*S - 0.0216*FC + 30.47 \quad R^2 = 0.4859 \quad (206)$$

$$OMC = -0.0574*S + 0.0106*FC - 0.0186*LL + 0*PL + 0.6284*PI + 3.97 \quad R^2 = 0.9212 \quad (207)$$

$$MDD = 0.0012*S - 0.0014*FC + 0.0028*LL + 0*PL - 0.0195*PI + 2.13 \quad R^2 = 0.9104 \quad (208)$$

80% Training Dataset

$$LL = -0.2831*S + 0.1244*FC + 35.02 \quad R^2 = 0.4876 \quad (209)$$

$$PL = -0.1373*S + 0.0614*FC + 13.87 \quad R^2 = 0.4355 \quad (210)$$

$$PI = -0.1458*S + 0.063*FC + 21.15 \quad R^2 = 0.4585 \quad (211)$$

$$OMC = -0.0561*S + 0.014*FC + 0.5724*LL - 0.5858*PL + 0*PI + 4.33 \quad R^2 = 0.9045 \quad (212)$$

$$MDD = 0.0011*S - 0.0018*FC - 0.0146*LL + 0.0168*PL + 0*PI + 2.13 \quad R^2 = 0.8996 \quad (213)$$

90% Training Dataset

$$LL = -0.3131*S + 0.1019*FC + 37.94 \quad R^2 = 0.5083 \quad (214)$$

$$PL = -0.139*S + 0.056*FC + 14.43 \quad R^2 = 0.4353 \quad (215)$$

$$PI = -0.1741*S + 0.0458*FC + 23.50 \quad R^2 = 0.4901 \quad (216)$$

$$OMC = -0.0489*S + 0.0188*FC - 0.0327*LL + 0*PL + 0.6356*PI + 3.46 \quad R^2 = 0.9201 \quad (217)$$

$$MDD = 0.0012*S - 0.0016*FC + 0.0029*LL + 0*PL - 0.019*PI + 2.14 \quad R^2 = 0.9156 \quad (218)$$

100% Training Dataset

$$LL = -0.3674*S + 0.0495*FC + 42.76 \quad R^2 = 0.4822 \quad (219)$$

$$PL = -0.1485*S + 0.0458*FC + 15.23 \quad R^2 = 0.4095 \quad (220)$$

$$PI = -0.2188*S + 0.0037*FC + 27.53 \quad R^2 = 0.4656 \quad (221)$$

$$OMC = -0.0544*S + 0.0141*FC + 0.5943*LL - 0.6201*PL + 0*PI + 4.02 \quad R^2 = 0.9134 \quad (222)$$

$$MDD = 0.0012*S - 0.0016*FC - 0.0157*LL + 0.0182*PL + 0*PI + 2.13 \quad R^2 = 0.9066 \quad (223)$$

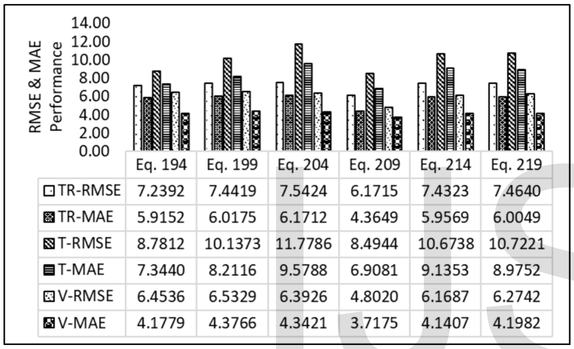
The ratio of sand to fine content is eliminated in the present work because the sand: fine content ratio has multicollinearity. The sand and fine content have a strong relationship, but the sand: fine content ratio has a moderate relationship. The sand: fine content ratio may reduce the performance of regression models.

4 Results and Discussions

A total of 96 equations are derived for each simple linear and polynomial regression analysis to predict consistency limits and compaction parameters. The proposed simple regression equations have a good coefficient of determination [14]. In addition, a total of 30 multi-linear equations are derived for predicting consistency limits and compaction parameters for different percentages of training datasets. The results of consistency limits and compaction parameters are discussed below.

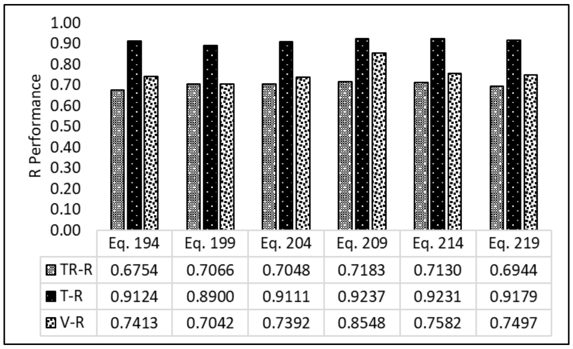
4.1 Prediction of Liquid Limit

Equations 194, 199, 204, 209, 214, and 219 are used to predict the liquid limit of soil. The proposed multiple regression models of the liquid limit are tested and validated by 53 and 26 datasets, respectively. Equations 194, 199, 204, 209, 214, and 219 are derived at 50%, 60%, 70%, 80%, 90% and 100% training datasets, respectively. The training, testing, and validation performances of multi-linear regression (MRA) models are compared, as shown in Figs. 2 and 3.



	Eq. 194	Eq. 199	Eq. 204	Eq. 209	Eq. 214	Eq. 219
□ TR-RMSE	7.2392	7.4419	7.5424	6.1715	7.4323	7.4640
▨ TR-MAE	5.9152	6.0175	6.1712	4.3649	5.9569	6.0049
▨ T-RMSE	8.7812	10.1373	11.7786	8.4944	10.6738	10.7221
▤ T-MAE	7.3440	8.2116	9.5788	6.9081	9.1353	8.9752
▨ V-RMSE	6.4536	6.5329	6.3926	4.8020	6.1687	6.2742
▨ V-MAE	4.1779	4.3766	4.3421	3.7175	4.1407	4.1982

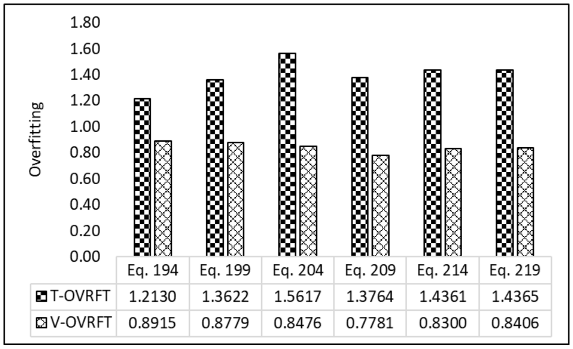
Fig. 2. Training, testing, validation performances (RMSE & MAE) of MRA models for LL



	Eq. 194	Eq. 199	Eq. 204	Eq. 209	Eq. 214	Eq. 219
▨ TR-R	0.6754	0.7066	0.7048	0.7183	0.7130	0.6944
■ T-R	0.9124	0.8900	0.9111	0.9237	0.9231	0.9179
▨ V-R	0.7413	0.7042	0.7392	0.8548	0.7582	0.7497

Fig 3. Training, testing, validation performances (R) of MRA models for LL

Figs. 2 and 3 show that equation 209 predicted the liquid limit of soil with the testing and validation performance of 0.9226 and 0.8548, respectively. In addition, equation 209 predicted the liquid limit of soil with the testing and validation prediction error (RMSE) of 4.3649 and 6.7794, respectively. Therefore, equation 209 based MRA model is identified as the best architecture MRA model for LL. The training performance increases by increasing the percentage of input parameters [10]. On the other hand, equation 209 is developed using 80% training datasets having a strong correlation with the liquid limit. Therefore, it may be stated that the better performance of the MRA model may be achieved by 80% training dataset or having a strong correlation with pair of the dataset. The overfitting of Eqs. 194, 199, 204, 209, 214, and 219 are also calculated, as shown in Fig. 4.



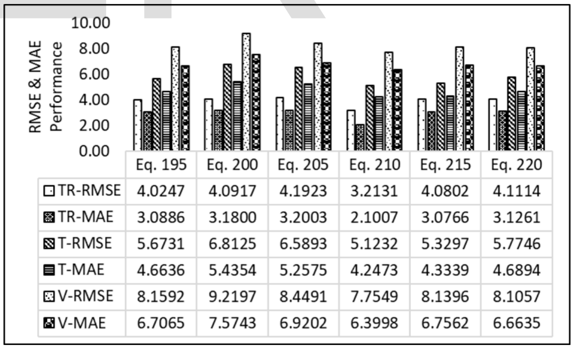
	Eq. 194	Eq. 199	Eq. 204	Eq. 209	Eq. 214	Eq. 219
▩ T-OVRFT	1.2130	1.3622	1.5617	1.3764	1.4361	1.4365
▨ V-OVRFT	0.8915	0.8779	0.8476	0.7781	0.8300	0.8406

Fig. 4. Overfitting of MRA models in predicting liquid limit

Fig. 4 shows the overfitting of MRA models of liquid limit during testing and validation. The overfitting comparison shows that the equation 209 based MRA model predicted liquid limit in testing and validation with overfitting of 1.0985 and 0.7781, respectively.

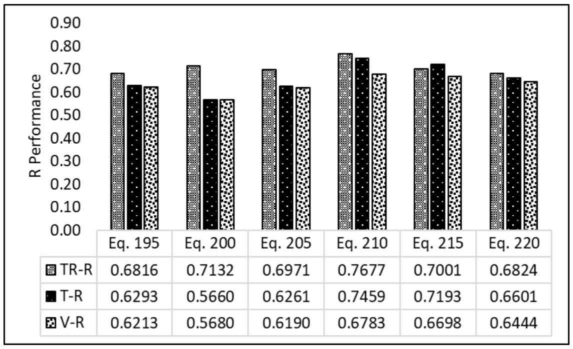
4.2 Prediction of Plastic Limit

Equations 195, 200, 205, 210, 215, and 220 are used to predict the plastic limit of soil. The proposed multiple regression models of the plastic limit are tested and validated by 53 and 26 datasets, respectively. Equations 194, 199, 204, 209, 214, and 219 are derived at 50%, 60%, 70%, 80%, 90% and 100% training datasets, respectively. The training, testing, and validation performance of MRA models are compared, as shown in Figs. 5 and 6.



	Eq. 195	Eq. 200	Eq. 205	Eq. 210	Eq. 215	Eq. 220
□ TR-RMSE	4.0247	4.0917	4.1923	3.2131	4.0802	4.1114
▨ TR-MAE	3.0886	3.1800	3.2003	2.1007	3.0766	3.1261
▨ T-RMSE	5.6731	6.8125	6.5893	5.1232	5.3297	5.7746
▤ T-MAE	4.6636	5.4354	5.2575	4.2473	4.3339	4.6894
▨ V-RMSE	8.1592	9.2197	8.4491	7.7549	8.1396	8.1057
▨ V-MAE	6.7065	7.5743	6.9202	6.3998	6.7562	6.6635

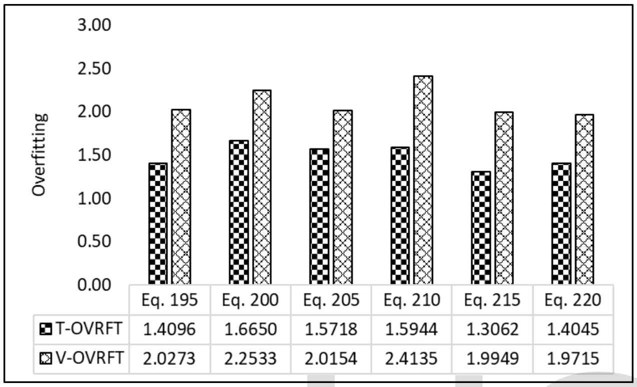
Fig. 5. Training, testing, validation performances (RMSE & MAE) of MRA models for PL



	Eq. 195	Eq. 200	Eq. 205	Eq. 210	Eq. 215	Eq. 220
▨ TR-R	0.6816	0.7132	0.6971	0.7677	0.7001	0.6824
■ T-R	0.6293	0.5660	0.6261	0.7459	0.7193	0.6601
▨ V-R	0.6213	0.5680	0.6190	0.6783	0.6698	0.6444

Fig. 6. Training, testing, validation performances (R) of MRA models for PL

Figs. 5 and 6 show that equation 210 predicted the plastic limit of soil with the testing and validation performance of 0.7461 and 0.7529, respectively. In addition, equation 210 predicted the plastic limit of soil with the testing and validation prediction error (RMSE) of 7.7976 and 5.2416, respectively. Therefore, equation 210 based MRA model is identified as the best architecture MRA model for PL. On the other hand, equation 210 is developed using 80% training datasets having a strong correlation with the plastic limit. Therefore, it may be stated that the better performance of the MRA model may be achieved by 80% training dataset or having a strong correlation with pair of the dataset. The overfitting of Eqs. 195, 200, 205, 210, 215, and 220 are also calculated, as shown in Fig. 7.



	Eq. 195	Eq. 200	Eq. 205	Eq. 210	Eq. 215	Eq. 220
▣ T-OVRFT	1.4096	1.6650	1.5718	1.5944	1.3062	1.4045
▨ V-OVRFT	2.0273	2.2533	2.0154	2.4135	1.9949	1.9715

Fig. 7. Overfitting of MRA models in predicting plastic limit

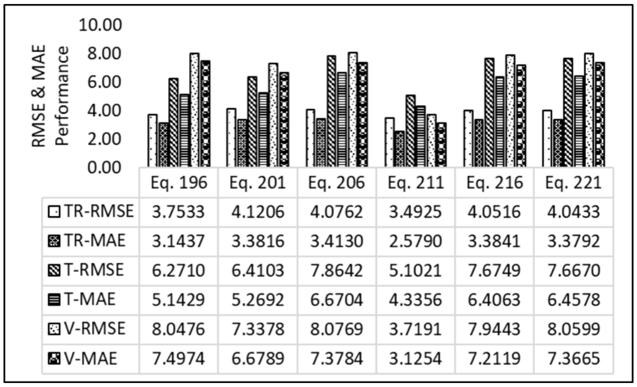
Fig. 7 shows the overfitting of MRA models of plastic limit during testing and validation. The overfitting comparison shows that the equation 210 based MRA model predicted a plastic limit in testing and validation with overfitting of 2.2327 and 1.5008, respectively.

4.3 Prediction of Plasticity Index

Equations 196, 201, 206, 211, 216, and 221 are used to predict the plasticity index of soil and derived at 50%, 60%, 70%, 80%, 90% and 100% training datasets, respectively. The proposed multiple regression models of the plasticity index are tested and validated by 53 and 26 datasets, respectively. The training, testing, and validation performances of MRA models are compared, as shown in Figs. 8 and 9.

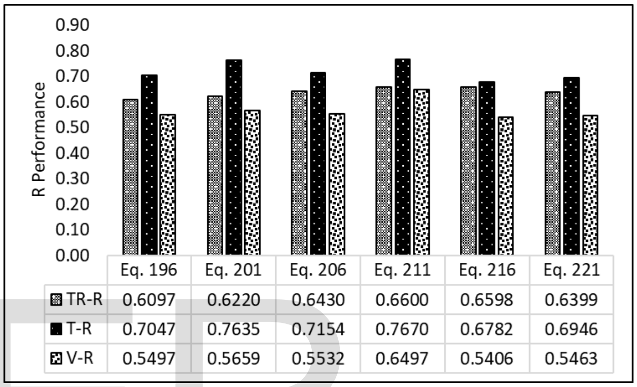
Figs. 8 and 9 show that equation 201 predicted the plasticity index of soil with the testing and validation performance of 0.8326 and 0.5927, respectively. In addition, equation 201 predicted the plasticity index of soil with the testing and validation prediction error (RMSE) of 3.8771 and 3.4213, respectively. Therefore, equation 201 based MRA model is identified as the best architecture MRA model for PI. On the other hand, equation 201 is developed using 60% training datasets having a strong correlation with the plasticity index. Therefore, it may be stated that the better performance of the MRA model may be achieved by 60% training dataset or having a strong correlation with pair of the dataset. The overfitting of Eqs. 196, 201, 206, 211, 216, and 221 are also calculated, as shown in Fig. 10.

Fig. 10 shows the overfitting of MRA models of plasticity index during testing and validation. The overfitting comparison shows that the equation 201 based MRA model predicted a plasticity index in testing and validation with overfitting of 0.9476 and 0.8362, respectively.



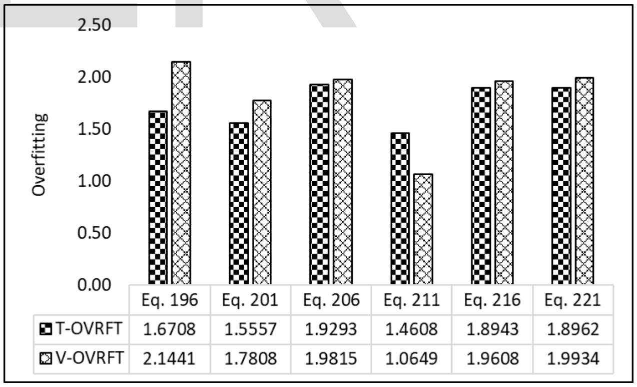
	Eq. 196	Eq. 201	Eq. 206	Eq. 211	Eq. 216	Eq. 221
□ TR-RMSE	3.7533	4.1206	4.0762	3.4925	4.0516	4.0433
▩ TR-MAE	3.1437	3.3816	3.4130	2.5790	3.3841	3.3792
▨ T-RMSE	6.2710	6.4103	7.8642	5.1021	7.6749	7.6670
▤ T-MAE	5.1429	5.2692	6.6704	4.3356	6.4063	6.4578
▧ V-RMSE	8.0476	7.3378	8.0769	3.7191	7.9443	8.0599
▣ V-MAE	7.4974	6.6789	7.3784	3.1254	7.2119	7.3665

Fig. 8. Training, testing, validation performances (RMSE & MAE) of MRA models for PI



	Eq. 196	Eq. 201	Eq. 206	Eq. 211	Eq. 216	Eq. 221
▩ TR-R	0.6097	0.6220	0.6430	0.6600	0.6598	0.6399
■ T-R	0.7047	0.7635	0.7154	0.7670	0.6782	0.6946
▧ V-R	0.5497	0.5659	0.5532	0.6497	0.5406	0.5463

Fig. 9 Training, testing, validation performances (R) of MRA models for PI



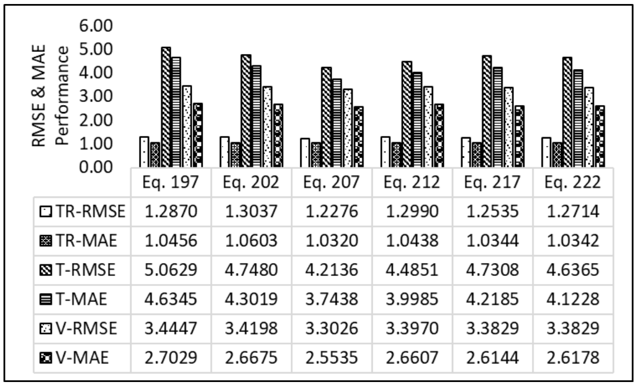
	Eq. 196	Eq. 201	Eq. 206	Eq. 211	Eq. 216	Eq. 221
▣ T-OVRFT	1.6708	1.5557	1.9293	1.4608	1.8943	1.8962
▨ V-OVRFT	2.1441	1.7808	1.9815	1.0649	1.9608	1.9934

Fig. 10. Overfitting of MRA models in predicting plasticity index

4.4 Prediction of Optimum Moisture Content

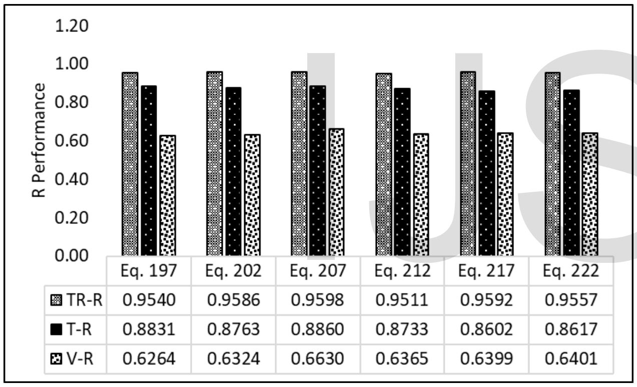
Equations 197, 202, 207, 212, 217, and 222 are used to predict the optimum moisture content of soil and derived at 50%, 60%, 70%, 80%, 90% and 100% training datasets, respectively. The proposed multiple regression models of the OMC are tested and validated by 53 and 26 datasets, respectively. The training, testing, and validation performances of MRA models are compared, as shown in Figs. 11 and 12.

Figs. 11 and 12 show that equation 207 predicted the OMC of soil with the testing and validation performance of 0.8824 and 0.6630, respectively. In addition, equation 207 predicted the OMC of soil with the testing and validation prediction error (RMSE) of 4.2336 and 3.3026, respectively. Therefore, the equation 207 based MRA model is identified as the best architecture MRA model for OMC. On the other hand, equation 207 is developed using 70% training datasets having a strong to very strong correlation with the OMC. Therefore, it may be stated that the excellent performance of the MRA model may be achieved by 70% training dataset or having a strong-very strong correlation with pair of the dataset. The overfitting of Eqs. 197, 202, 207, 212, 217, and 222 are also calculated, as shown in Fig. 13.



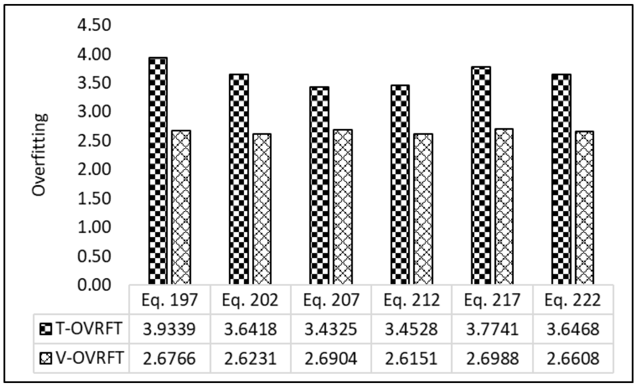
	Eq. 197	Eq. 202	Eq. 207	Eq. 212	Eq. 217	Eq. 222
□ TR-RMSE	1.2870	1.3037	1.2276	1.2990	1.2535	1.2714
▩ TR-MAE	1.0456	1.0603	1.0320	1.0438	1.0344	1.0342
▧ T-RMSE	5.0629	4.7480	4.2136	4.4851	4.7308	4.6365
▤ T-MAE	4.6345	4.3019	3.7438	3.9985	4.2185	4.1228
▨ V-RMSE	3.4447	3.4198	3.3026	3.3970	3.3829	3.3829
▦ V-MAE	2.7029	2.6675	2.5535	2.6607	2.6144	2.6178

Fig. 11. Training, testing, validation performances (RMSE & MAE) of MRA models for OMC



	Eq. 197	Eq. 202	Eq. 207	Eq. 212	Eq. 217	Eq. 222
▩ TR-R	0.9540	0.9586	0.9598	0.9511	0.9592	0.9557
■ T-R	0.8831	0.8763	0.8860	0.8733	0.8602	0.8617
▨ V-R	0.6264	0.6324	0.6630	0.6365	0.6399	0.6401

Fig. 12. Training, testing, validation performances (R) of MRA models for OMC



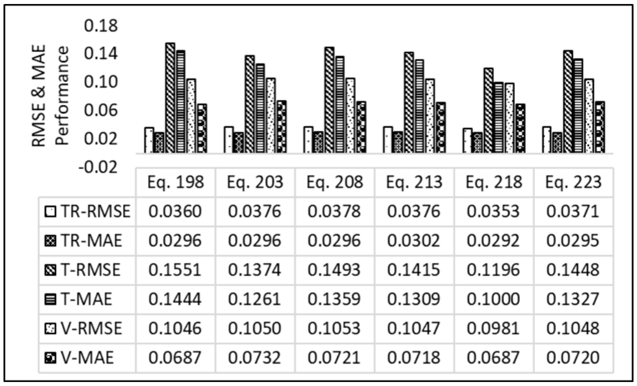
	Eq. 197	Eq. 202	Eq. 207	Eq. 212	Eq. 217	Eq. 222
▦ T-OVRFT	3.9339	3.6418	3.4325	3.4528	3.7741	3.6468
▨ V-OVRFT	2.6766	2.6231	2.6904	2.6151	2.6988	2.6608

Fig. 13. Overfitting of MRA models in predicting OMC

Fig. 13 shows the overfitting of MRA models of OMC during testing and validation. The overfitting comparison shows that the equation 207 based MRA model predicted OMC in testing and validation with overfitting of 3.3186 and 2.5888.

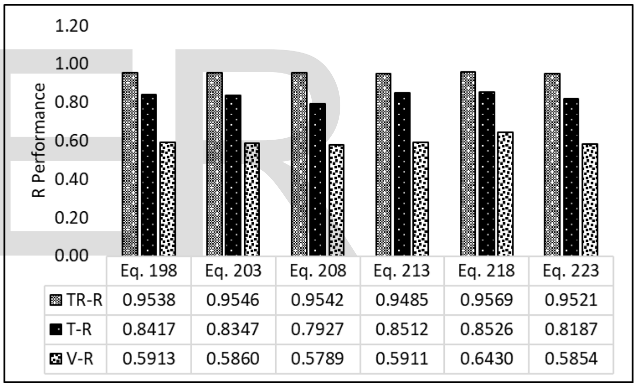
4.5 Prediction of Maximum Dry Density

Equations 198, 203, 208, 213, 218, and 223 are used to predict the maximum dry density of soil. The proposed multiple regression models of the MDD are tested and validated by 53 and 26 datasets, respectively. Equations 198, 203, 208, 213, 218, and 223 are derived at 50%, 60%, 70%, 80%, 90% and 100% training datasets, respectively. The training, testing, and validation performances of multi-linear regression (MRA) models are compared, as shown in Figs. 14 and 15.



	Eq. 198	Eq. 203	Eq. 208	Eq. 213	Eq. 218	Eq. 223
□ TR-RMSE	0.0360	0.0376	0.0378	0.0376	0.0353	0.0371
▩ TR-MAE	0.0296	0.0296	0.0296	0.0302	0.0292	0.0295
▧ T-RMSE	0.1551	0.1374	0.1493	0.1415	0.1196	0.1448
▤ T-MAE	0.1444	0.1261	0.1359	0.1309	0.1000	0.1327
▨ V-RMSE	0.1046	0.1050	0.1053	0.1047	0.0981	0.1048
▦ V-MAE	0.0687	0.0732	0.0721	0.0718	0.0687	0.0720

Fig. 14. Training, testing, validation performances (RMSE & MAE) of MRA models for MDD



	Eq. 198	Eq. 203	Eq. 208	Eq. 213	Eq. 218	Eq. 223
▩ TR-R	0.9538	0.9546	0.9542	0.9485	0.9569	0.9521
■ T-R	0.8417	0.8347	0.7927	0.8512	0.8526	0.8187
▨ V-R	0.5913	0.5860	0.5789	0.5911	0.6430	0.5854

Fig. 15 Training, testing, validation performances (R) of MRA models for MDD

Figs. 14 and 15 show that equation 218 predicted the MDD of soil with the testing and validation performance of 0.8799 and 0.6430, respectively. In addition, equation 218 predicted the MDD of soil with the testing and validation prediction error (RMSE) of 0.1346 and 0.0981, respectively. Therefore, the equation 218 based MRA model is identified as the best architecture MRA model for MDD. On the other hand, equation 218 is developed using 90% training datasets having a strong to very strong correlation with MDD. Therefore, it may be stated that the better performance of the MRA model may be achieved by 90% training dataset or having a strong-very strong correlation with pair of the dataset. The overfitting of Eqs. 198, 203, 208, 213, 218, and 223 are also calculated, as shown in Fig. 16.

Fig. 16 shows the overfitting of MRA models of MDD during testing and validation. The overfitting comparison shows that the equation 218 based MRA model predicted MDD in testing and validation with overfitting of 3.7085 and 2.7023, respectively.

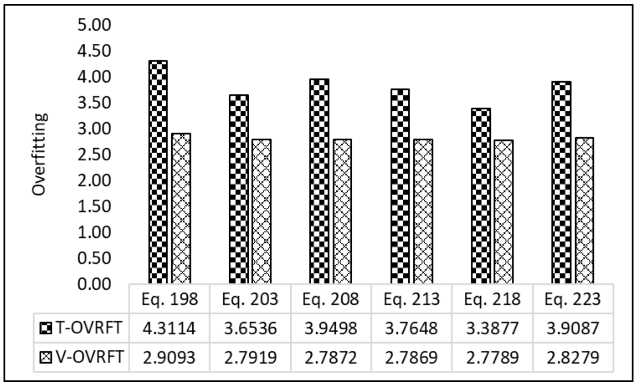


Fig. 16. Overfitting of MRA models in predicting MDD

5 Soil Classification using Predicted Results

The particle size distribution and consistency limits procedures are used to classify soil. The particle size distribution classifies coarse-grained soils, and consistency limits classify fine-grained soils. The liquid limit and plasticity index are required to classify fine-grained soils. The liquid limit and plasticity index are predicted using Eqs. 209 and 201, respectively, because of Eqs. 209 and 201 based MRA models outperformed other liquid limit and plasticity index models. The liquid limit and plasticity index MRA models are tested and validated by 53 and 26 datasets. Therefore, the test and validation soil datasets are classified, as shown in Figs. 17 and 18.

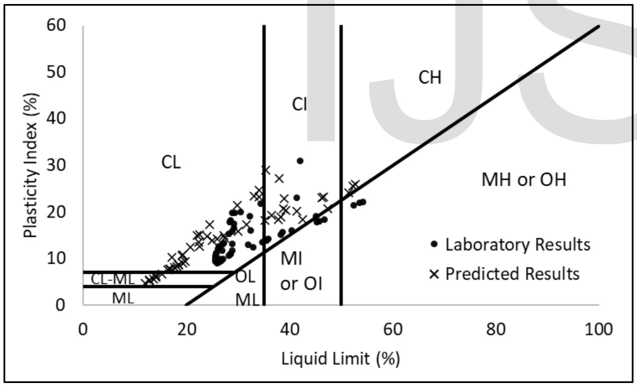


Fig. 17 Classification of test soil dataset using Eqs. 209 and 201 based MRA models

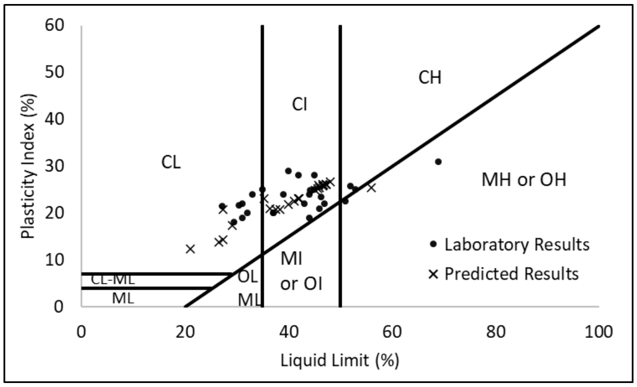


Fig. 18 Classification of validation soil dataset using Eqs. 209 and 201 based MRA models

The testing dataset has more soil specimens than validation soil specimens. Therefore, better results of soil classification are achieved during the testing of MRA models. Finally, the number of soil datasets also affects the prediction of soil classification.

6 Sensitivity Analysis

The cosine amplitude method determines the sensitivity of input parameters for soil LL, PL, PI, OMC, and MDD. The following equation illustrates the cosine amplitude method.

$$S_s = \frac{\sum_{k=1}^{j}(X_{ik} * Y_{ik})}{\sqrt{\sum_{k=1}^{j} X_{ik}^2 \sum_{k=1}^{j} Y_{ik}^2}} \qquad (224)$$

Where Xi and Yi are input and output parameters, respectively, the sensitivity (Ss) ranges between 0 and 1. The value of sensitivity (Ss) closer to one shows the strength between input and output parameters of models. If the input and output parameters have no relation, the Ss value is zero. The liquid limit, plastic limit, plasticity index is predicted using sand and fine content.

Similarly, the OMC and MDD of soil are predicted using sand content, fine content, liquid limit, plastic limit, and plasticity index. The sensitivity analysis is performed for liquid limit, plastic limit, plasticity index, optimum moisture content, and maximum dry density at different percentages of training datasets, as shown in Figs. 19-23.

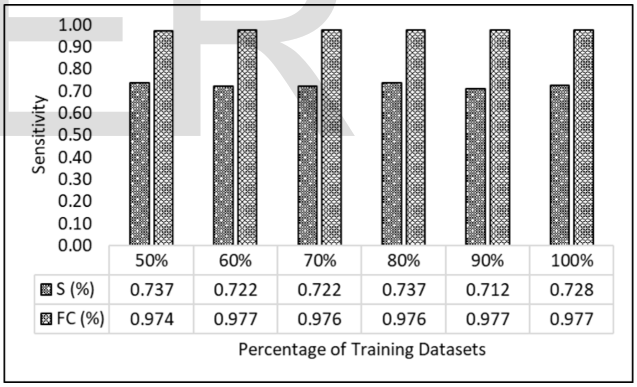


Fig. 19. Sensitivity analysis for liquid limit

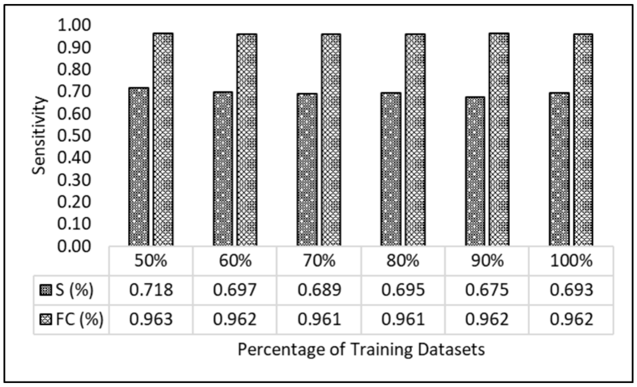


Fig. 20. Sensitivity analysis for plastic limit

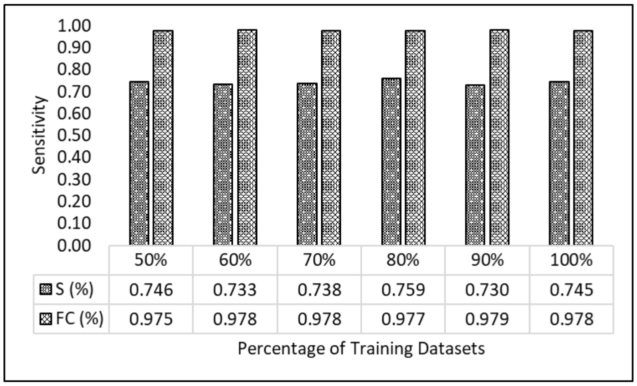


Fig. 21. Sensitivity analysis for plasticity index

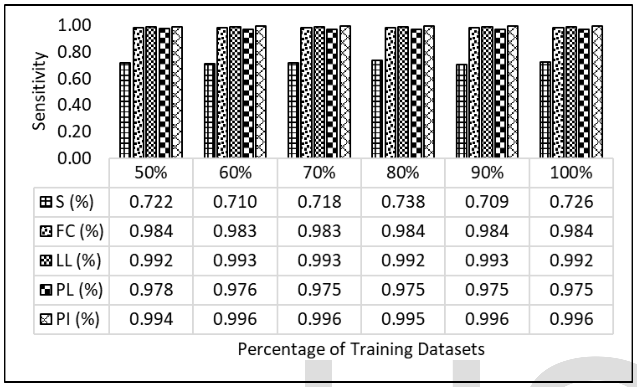


Fig. 22 Sensitivity analysis for optimum moisture cotent

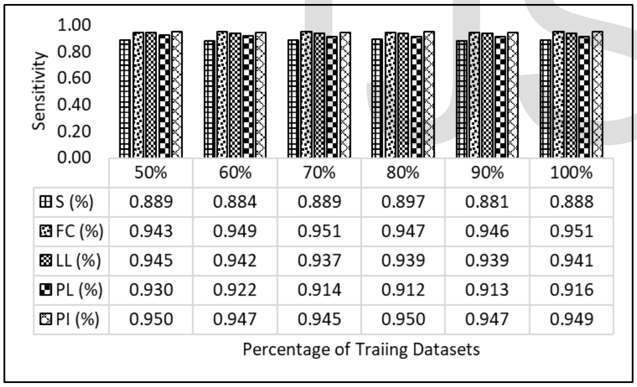


Fig. 23. Sensitivity analysis for maximum dry density

The sensitivity of sand and fine content for liquid limit, plastic limit, and plasticity index is shown in Fig. 14 (a), (b), and (c), respectively. The fine content is highly sensitive for liquid limit, plastic limit, and plasticity index than sand content. In other words, the consistency limits of soil are highly influenced by fine content. The sensitivity of sand, fine content, liquid limit, plastic limit, and plasticity index for OMC and MDD is shown in Fig. 14 (d) and (e), respectively. Fig. 14 (d) shows that the sand content is less influencing in predicting the OMC of soil than other input parameters. Fig. 14 (d) also shows that the OMC is highly influenced by FC, LL, PL, and PI. On the other hand, the maximum dry density is highly influenced by S, FC, LL, PL, and PI, as shown in Fig. 14 (e). It is also observed that the sensitivity of input parameters increases by increasing the percentage of training datasets.

7 Conclusions

The following conclusions are mapped in the present study:

- The coefficient of determination (R2) comparison of simple linear and polynomial regression showed that the nonlinear approach predicts the consistency limits and compaction parameters better than the linear approach.
- The sand and fine content strongly correlate with the liquid limit, plasticity index, plastic limit of soil. The best architecture LL MRA model predicted LL with the testing and validation performance of 0.9226 and 0.8548, respectively. Similarly, the best architecture PL MRA model predicted PL with the testing and validation performance of 0.7461 and 0.7529, respectively. The best architecture PI MRA model also predicted PI with the testing performance of 0.8326. Therefore, it is concluded that a strong relationship or more than 0.65 correlation coefficient of pair of datasets can predict LL, PL, and PI of soil with a performance of 0.70-1.0 and slightest error.
- The plastic limit strongly correlates with the OMC and MDD of soil. The relationship of sand content, fine content, liquid limit, and plasticity index with OMC and MDD improves (strong to very strong) by increasing the percentage of training datasets.
- If the pair of datasets has a strong or very strong correlation, a multiple linear regression model can achieve a prediction performance of more than 0.85.
- The number of validation datasets was less than the test datasets. Therefore, the validation performance was affected while predicting the LL, PL, PI, OMC, and MDD for validation datasets.
- The strong or very strong relationship of pair of datasets decreases the overfitting of the MRA model while predicting consistency limits and compaction parameters.
- The cosine amplitude sensitivity analysis showed that the LL, PL, and PI are highly influenced by fine content than sand content. The compaction parameters are highly influenced by FC, LL, PL, and PI.

Finally, it is concluded that the correlation coefficient affects the performance and overfitting of the regression models. The best prediction of compaction parameters and consistency limits can be achieved by a strongly or very strongly correlated pair of datasets.

8 Future Scope

The present study has the following future scopes.

i. The present study may be used for machine learning approaches to predict soil's compaction parameters, consistency limits, and strength parameters.

ii. The present case may be applied for deep and hybrid learning approaches to predict soil's compaction parameters, consistency limits, and strength parameters.

iii. A comparative study may be carried out for machine, deep, and hybrid learning approaches using the present research work.

ABBREVIATIONS AND NOTATIONS

ANFIS	Adaptive neural fuzzy inference system
ANN	Artificial neural network
CC	Coefficient of correlation
CH	Inorganic clays of high plasticity
CI	Inorganic clays of medium plasticity
CL	Inorganic clays of low plasticity
COD/R^2	Coefficient of determination
Cu	Coefficient of uniformity
D30	Particle size at 30% cumulative finer
D50	Particle size at 50% cumulative finer
FC	Fine content
FD	Frequency distribution
G	Gravel content
GEP	Gene expression programming
GMDH	Group method of data handling
J	Min-max normalization
LL	Liquid limit
LL'	Predicted liquid limit
MAE	Mean absolute error
MARS	Multivariate adaptive regression splines
MDD	Maximum dry density
MDD'	Predicted maximum dry density
MH	Inorganic silts of high compressibility
MI	Inorganic silts of medium compressibility
ML	Inorganic silts of non to low plasticity
MLR	Multi-linear regression
MRA	Multiple regression analysis
OH	Organic clays of medium to high plasticity
OI	Organic silts of medium plasticity
OL	Organic silts of low plasticity
OMC	Optimum moisture content
OMC'	Predicted optimum moisture content
PI	Plasticity index
PI'	Predicted plasticity index
PL	Plastic limit
PL'	Predicted plastic limit
R	Performance
RMSE	Root mean square error
S	Sand content
Ss	Sensitivity
St. Dev.	Standard deviation
SVM	Support vector machine
x	Test value

REFERENCES

[1] Al-Saffar Raghdan Zuhair, Dr. Suhail I. Khattab, Dr. Salim T. Yousif (2013), Prediction of soil's compaction parameter using artificial neural network, Al-Rafidain Engineering, 21(3), pp. 15-27.

[2] Ardakani Alireza, Afshin Kordnaeij (2017), Soil compaction parameters prediction using GMDH-type neural network and genetic algorithm, European Journal of Environmental and Civil Engineering, pp. 449-462, https://doi.org/10.1080/19648189.2017.1304269.

[3] Dr. K. R. Arora (2004), Soil mechanics and foundation engineering, Standard Publisher Distributors, New Delhi, India.

[4] Gunaydin Osman, Abdurrahman Ozbeyaz, Mehmet Soylemez (2018), Regression analysis of soil compaction parameters using support vector machine, Celal Bayar University Journal of Science, 14(4), pp. 443-447.

[5] Hair Jr. J, Wolfnibarger M. C., Ortinau, D. J. & Bush, R. P. (2013), Essentials of Marketing, New York, Mc Graw Hill

[6] Hasnat Arif, Md. Mehedi Hasan, Md Rabiul Islam, Md Abdul Alim (2019), Prediction of compaction parameters of soil using support vector machine, Current Trends in Civil & Structural Engineering, 4(1), pp.1-7, CTCSE.MS.ID.000580.

[7] Khalid Farooq, U. Khalid, H. Mujtaba (2015), Prediction of compaction characteristics of fine-grained soils using consistency limits, Arab Journal for Science and Engineering, pp. 1-10.

[8] Khalid Usama, Rehman Zia Ur (2018), Evaluation of compaction parameters of fine-grained soils using standard and modified efforts, International Journal of Geo-Engineering, 9:15, pp. 1-17, https://doi.org/10.1186/s40703-018-0083-1.

[9] Khuntia Sunil, Mujtaba Hassan, Patra Chittaranjan, Khalid Farooq, Nagaratnam Sivakugan, Braja M. Das (2015), Prediction of compaction parameters of coarse-grained soil using multivariate adaptive regression splines, International Journal of Geotechnical Engineering, 9(1), pp. 79-88, DOI 10.1179/1939787914Y.0000000061.

[10] Mukhlisin Muhammad, Rehman Aini Sharina Binti Abd (2014), Prediction of Atterberg limits via ANN and ANFIS: a comparison, Recent Advances in Environmental Science and Geoscience, pp. 69-74, ISBN: 978-1-61804-224-8.

[11] Rehman Attique ul, Farooq Khalid, Mujtaba Hassan (2017), Prediction of California bearing ratio (CBR) and compaction characteristics of granular soils, Acta Geotechnica Solvenica, pp. 63-72.

[12] Shrivastava A. K., Dr. P. K. Jain (2016), Prediction of compaction parameters using regression and ANN tools, International Journal for Scientific Research and Development, 3(11), pp. 697-702.

[13] Sirvrikaya O., Kayadelen C., Ceren E. (2013), Prediction of the compaction parameters for coarse-grained soils with fines content by MLR and GEP, Acta Geotechnia Slovenica, pp. 29-41.

[14] Smith G. N. (1986), Probability and statistics in civil engineering – An introduction, Collins, London

[15] Wang Han-Lin, Zhen-Yu Yin (2020), High-performance prediction of soil compaction parameters using multi expression programming, Engineering Geology, 276, 105758.