International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

Development of a Machine Learning Model for IS Soil Classification using Index Properties of Soil

Swathi K[1], Gowthami G Pai[2], Shyam Kumar M H[3], Shankargiri K S[4]

[123]*B.E. Student, Department of Civil Engineering, VCET, Puttur, India*
[4]*Professor, Department of Civil Engineering, VCET, Puttur, India*

Abstract - *Soil classification according to the Indian Standard IS 1498:1970 is a fundamental step in geotechnical investigations, but the conventional procedure depends on laboratory testing followed by manual chart-based interpretation, which is time consuming and prone to human error. This paper presents a machine learning model for automated IS soil classification using soil index properties, namely gravel, sand and fines percentages, liquid limit and plastic limit, together with derived parameters such as plasticity index, coefficient of uniformity and coefficient of curvature. A balanced synthetic dataset of more than 5000 samples covering 15 inorganic IS soil groups was generated using SMOTE and divided into 70 percent training and 30 percent testing data. A Random Forest classifier was trained and achieved an overall accuracy of 91.73 percent on the unseen test data. Boundary tests covering the 50 percent fines limit, the 50/50 sand gravel split, dual symbol soils such as GW-GC and SW-SC, the A-line, and liquid limit values of 35 and 50 were also performed. The model was integrated into a Streamlit based web application that returns the IS group symbol instantly from the entered laboratory values. Six physical soil samples were tested in the laboratory using wet sieve analysis and Atterberg limit tests and classified manually, and the model predictions agreed with the manual classification for all samples. The tool reduces classification time from 15 to 20 minutes per sample to under one second.*

Key Words: Soil classification, IS 1498:1970, Random Forest, Machine learning, Atterberg limits, Index properties, Streamlit, Geotechnical engineering

1. INTRODUCTION

Knowing how a soil will behave under load is the starting point of any foundation, pavement or earth structure design, and classification is the first step toward that knowledge. Indian practice follows IS 1498:1970 [9], which assigns a soil to a group using grain size distribution together with the liquid limit (LL), plastic limit (PL) and plasticity index (PI).

In routine practice the engineer runs the laboratory tests, computes the parameters and then reads the group off the plasticity chart by hand. This is slow and depends on the skill of the person doing it, and on projects with many boreholes the repeated manual work invites inconsistent or erroneous results.

Machine learning (ML) offers a way to automate this step. Models trained on labelled soil data learn the link between measured properties and soil class and can return a prediction in a fraction of a second [2], [3], [10]-[15]. Most published studies, however, target systems such as the Unified Soil Classification System (USCS), and very little has been reported for the Indian Standard system.

This paper therefore develops a Random Forest model for IS soil classification using index properties. The objectives are: (i) to collect soil samples and perform manual laboratory classification as a baseline for testing; (ii) to develop a predictive ML model that classifies soils according to IS 1498:1970; (iii) to evaluate the model using statistical metrics and boundary tests; and (iv) to deploy a user-friendly web application that gives instantaneous classification from raw laboratory inputs.

2. LITERATURE REVIEW

Automation of soil classification has been explored on several fronts. Ahmad et al. [1] built a Python desktop tool that applies USCS rules directly to test data, removing the need for manual look-up and lowering the chance of transcription mistakes. Liu et al. [2] trained a Random Forest on a range of soil parameters and used its feature importance to see which inputs drive the class; the ensemble coped well with nonlinear relations and reached high accuracy. Aydin et al. [3] compared several learning algorithms on field datasets containing gaps and uncertainty, and concluded that dependable classes can be obtained with fewer laboratory tests.

The choice of classification scheme also matters: Abrar and Hasan [4] found that one fine-grained soil can fall into different groups under USCS and AASHTO, with possible consequences for foundation design. Jameel and Chowdhary [5] merged hybrid learners with spatial data and reported accuracy above 95%, with the benefit most visible where soil records are sparse. Tran et al. [6] applied K-Nearest Neighbours and Support Vector Machines to more than 5800 cohesive soil records from Vietnam to speed up routine classification.

Work on the index parameters themselves includes Rosas et al. [7], who estimated the liquid and plastic limits by regression models so that fine-grained soils can be located on the Casagrande chart with less testing, and Zhu et al. [8], who found that pooling data from different sources gives better shear strength predictions than any single source. Neural network and cone penetration test-based classifiers have likewise been studied [10]-[15]. Taken together, these studies confirm that index properties are sufficient inputs for learned classifiers, yet the Indian Standard system has received little attention, which is the gap this paper addresses.

3. METHODOLOGY

The proposed workflow, shown in Fig -1, comprises data collection, feature engineering, ML model development, validation and deployment as a web application. Soil data are collected and classified manually as per IS 1498:1970, pre-processed by calculating the required parameters, and used to train a Random Forest model, whose performance is evaluated using accuracy, confusion matrix and F1-score.



Fig -1: Proposed project methodology workflow

3.1 Soil Classification as per IS 1498:1970

IS 1498:1970 places soils in one of three divisions: coarse-grained, fine-grained and highly organic. The deciding quantity is the share of material passing the 75 micron IS sieve. Where more than half is retained, the soil is coarse-grained; it is called gravel (G) or sand (S) depending on the portion held on the 4.75 mm sieve, and a second letter records gradation (well-graded W, poorly-graded P) or the character of the fines (silty M, clayey C). Where more than half passes, the soil is fine-grained, and its position relative to the A-line separates silt (M) from clay (C), while the liquid limit gives the compressibility class (low L, intermediate I, high H). The broad classification and the plasticity chart are shown in Fig -2 and Fig -3.

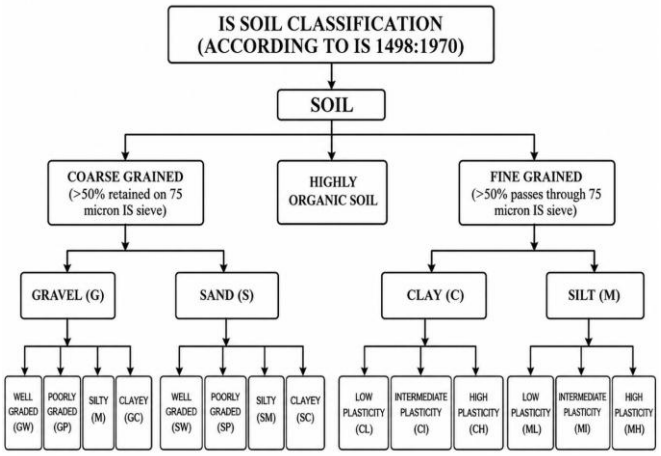

Fig -2: Broad classification of soils based on IS 1498:1970

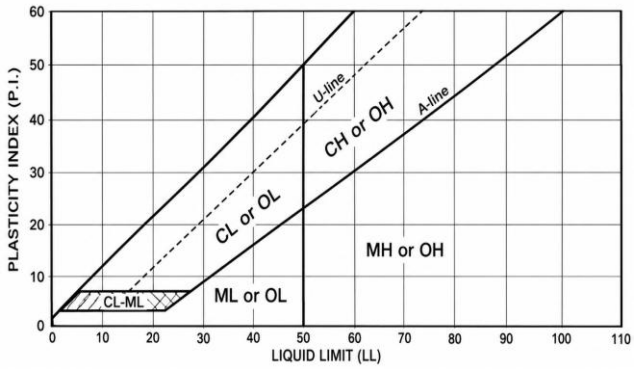

Fig -3: IS 1498 plasticity chart for classification of fine-grained soils

3.2 Synthetic Training Data

A comprehensive dataset was synthetically generated to represent the variety of soils encountered in geotechnical practice. Data generation techniques, including SMOTE (Synthetic Minority Over-sampling Technique), were used to obtain a balanced dataset that satisfies the criteria of IS 1498:1970, so that the 15 targeted IS soil groups are evenly represented without the bias found in historical datasets. Organic (O) soils and peat (Pt) were deliberately excluded so that the model is optimized for standard inorganic soils.

3.3 Field Samples and Laboratory Determination

To validate the model against real inputs, soil samples were collected from different locations. Standard IS laboratory tests were then carried out to obtain the input parameters: grain size distribution (percentages of gravel, sand and fines) and the Atterberg limits (liquid limit and plastic limit). These values were also used for manual classification to IS 1498:1970, which serves as the benchmark for the ML predictions.

3.4 Pre-processing and Feature Engineering

Data cleaning handled missing values and normalized numerical inputs. Additional features were then derived from the raw inputs: the plasticity index from the Atterberg limits, and the coefficients of uniformity and curvature from the grain size distribution, as given by

$$I_P = LL - PL \quad (1)$$

$$C_u = D_{60} / D_{10} \quad (2)$$

$$C_c = (D_{30})^2 / (D_{10} \times D_{60}) \quad (3)$$

where D_{10}, D_{30} and D_{60} are the particle diameters corresponding to 10%, 30% and 60% finer. Fig -4 summarizes how these fundamental properties map to the IS 1498 group symbols.

No.	% Gravel	% Sand	% Fines (<75μ)	Liquid Limit (wL)	Plastic Limit (wP)	Cu	Cc	Color/Odor	IS Group
1	55	30	15	45	25	-	-	Reddish	GC
2	2	95	3	-	-	8.5	2.1	Yellow	SW
3	5	80	15	25	20	-	-	Grey	SM
4	0	10	90	28	18	-	-	Brown	CL
5	1	5	94	42	22	-	-	Yellow	CI
6	0	2	98	65	30	-	-	Red	CH
7	0	4	96	32	26	-	-	Light Grey	ML
8	2	8	90	45	35	-	-	Grey	MI
9	0	5	95	55	40	-	-	Dark Grey	MH
10	4	90	6	30	18	7.2	2.5	Brown	SW-SC
11	80	15	5	-	-	2.1	0.8	Brown	GP
12	0	3	97	18	15	-	-	Black/Odor	OL
13	5	5	90	25	20	-	-	Grey	CL-ML
14	0	0	100	48	38	-	-	Dark/Odor	OI
15	72	26	2	-	-	5.8	1.5	Light Brown	GW
16	65	32	3	-	-	3.2	0.8	Yellowish	GP
17	58	22	20	32	24	-	-	Brown	GM
18	62	18	20	45	22	-	-	Reddish	GC
19	4	94	2	-	-	7.1	2.2	Tan	SW
20	8	89	3	-	-	2.5	0.7	Light Grey	SP

Fig -4: Soil properties and corresponding IS 1498 group symbols

3.5 Random Forest Model

The prepared dataset of more than 5000 samples was split into 70% training and 30% testing data. The training data was used to learn the relationship between index properties and IS soil groups, while the testing data was kept separate to evaluate the model on unseen samples and to check that it learns general patterns rather than memorizing the training data. A Random Forest classifier, which combines many decision trees to obtain the final class, was trained and saved for testing and integration into the user interface.

3.6 Web Application

A graphical user interface was developed in Python using the Streamlit framework, chosen for its efficiency in deploying data science models as interactive web applications. Users enter gravel, sand and fines percentages and the Atterberg limits without any programming. The Streamlit backend passes these values to the trained model, which instantly predicts and displays the IS soil group symbol, removing the need for manual chart plotting.

4. EXPERIMENTS AND TESTING

4.1 Laboratory Tests

Wet sieve analysis was used to find the grain size distribution. A known quantity of soil was washed through the 75-micron sieve, and the retained material was dried and sieved through a series of standard sieves. The mass retained on each sieve was recorded and the gravel, sand and fines percentages were calculated (Fig -5). The liquid limit and plastic limit were determined by the standard Atterberg limit tests (Fig -6 and Fig -7).



Fig -5: Wet sieve analysis



Fig -6: Plastic limit test



Fig -7: Liquid limit test

4.2 Machine Learning Tests

The Random Forest model was trained on the prepared dataset, and the terminal output confirming training and model accuracy is shown in Fig -8. The model was then tested on the separate 30% testing dataset and achieved an overall accuracy of 91.73%.

```
PS C:\Users\swathi\OneDrive\Desktop\Soil Project> python train_model.py
Loading dataset...
Splitting data into training and testing sets...
Training the Random Forest model (this will take a few seconds)...
------------------------------
Model Training Complete!
Model Accuracy: 91.73%
------------------------------
Model saved successfully as 'soil_classification_model.pkl'!
PS C:\Users\swathi\OneDrive\Desktop\Soil Project>
```

Fig -8: Model training results

Because classification can change when a property lies very close to a limit, a targeted test script was used for boundary conditions: a 50/50 sand-gravel split, exactly 50% fines, dual-symbol cases GW-GC (at exactly 10% fines) and SW-SC (at exactly 12% fines), plasticity index just below the A-line, and liquid limits of 35 and 50. The model handled the tested boundary cases, including prediction of MI when the plasticity index was just below the A-line (Fig -9).

```
PS C:\Users\swathi\OneDrive\Desktop\Soil Project> python test_model.py
Waking up the trained model...
Analyzing soil parameters...

--------------------------------------------------
PREDICTED IS CLASSIFICATIONS:
--------------------------------------------------
Test 1 | GW     | Pure Well-Graded Gravel (<5% Fines, Cu>4, 1<Cc<3)
Test 2 | CL     | Borderline Coarse/Fine (Exactly 50% Fines, LL=34)
Test 3 | GW-GC  | Dual Symbol (Exactly 10% Fines)
Test 4 | MI     | A-Line Borderline (PI is just below the A-Line)
Test 5 | CI     | High Compressibility Boundary (LL exactly 50)
Test 6 | SW     | Sand vs. Gravel Boundary (Exactly 50/50 split)
Test 7 | SP     | Poorly Graded Sand (Fails Cu > 6 and Cc limits)
Test 8 | SP-SM  | Silty Sand (Fines > 12%, plots below A-Line)
Test 9 | SW-SC  | Dual Symbol Boundary (Exactly 12% Fines)
Test 10 | CI    | Low Compressibility Boundary (LL exactly 35)
--------------------------------------------------
PS C:\Users\swathi\OneDrive\Desktop\Soil Project>
```

Fig -9: Model boundary testing results

4.3 Interface Testing

The model was integrated into the interface and tested by entering soil parameters and checking the returned IS group symbol (Fig -10). The interface gave the classification directly from the entered values, which makes quick classification easier than manual chart plotting.

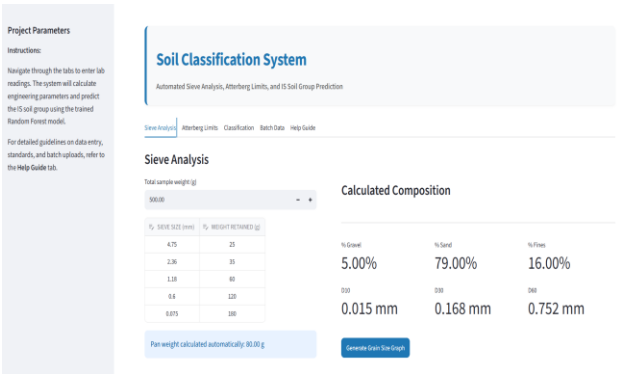

Fig -10: User interface of the developed system

5. RESULTS AND DISCUSSION

5.1 Laboratory Results

Six soil samples were tested. The percentage passing each sieve is given in Table -1, and the resulting gravel, sand and fines fractions, together with the Atterberg limits and plasticity index, are given in Table -2. Sample 5 was non-plastic (NP).

Table -1: Sieve analysis, percentage passing

S.No.	4.75 mm	2.36 mm	1.18 mm	0.6 mm	0.075 mm
1	85	76	69	63	48
2	90	70	50	30	20
3	100	100	90	80	70
4	40	30	20	10	2
5	95	80	40	10	2
6	100	99	97	93	85

Table -2: Grain size fractions and Atterberg limits

S.No.	Gravel (%)	Sand (%)	Fines (%)	LL (%)	PL (%)	PI (%)
1	15	37	48	44.68	26.53	18.15
2	10	70	20	29.93	25.00	4.93
3	0	30	70	64.84	25.00	39.84
4	60	38	2	22.20	18.00	4.20
5	5	93	2	NP	NP	NP
6	0	15	85	41.93	20.00	21.93

5.2 Manual versus Random Forest Classification

The same index properties were classified manually as per IS 1498:1970 and passed through the web application. The two sets of results are compared in Table -3.

Table -3: Manual and Random Forest classification

S.No.	Manual (IS 1498)	Random Forest	Description
1	SC	SC	Clayey sand
2	SM	SM	Silty sand
3	CH	CH	High plasticity clay
4	GW/GP	GW	Well graded gravel
5	SP	SP	Poorly graded sand
6	CI	CI	Intermediate plasticity clay

The Random Forest model predicted the IS soil group in agreement with the manual classification for all six samples. For sample 4 the manual result is the dual designation GW/GP, and the model resolved it to a single definite class, GW. It should be noted that the 91.73% accuracy refers to the synthetic test data, whereas the six physical samples are a small validation set; a larger set of regional field samples is needed to confirm performance on real soils.

6. CONCLUSIONS

A Random Forest model for classification of soils as per IS 1498:1970 from index properties was developed, tested and deployed as a web application. The main conclusions are:

1. The model achieved an overall accuracy of 91.73% on unseen test data and handled the tested boundary and dual-symbol cases.

2. Its predictions agreed with manual IS 1498:1970 classification for all six physical samples tested in the laboratory.

3. Automating the calculations and plasticity chart plotting through the Streamlit application reduced classification time from 15-20 minutes of manual work per sample to under one second.

4. Replacing visual chart interpolation with strict decision boundaries removes subjectivity in plasticity evaluation, and ambiguous dual designations such as GW/GP are resolved into a single defined class.

Future work includes collecting diverse regional soil samples to enrich the dataset, comparing alternative ML models with the Random Forest, adding compaction (OMC/MDD) and strength properties, developing a mobile-optimized interface for on-site use, estimating soil type from images using deep learning, and extending the framework to USCS and AASHTO classification.

ACKNOWLEDGEMENT

The authors thank the Department of Civil Engineering, VCET, Puttur, for providing the laboratory facilities and guidance for this work.

REFERENCES

[1] Z. Ahmad, B. A. Khan, I. Muhammad, I. Ullah, and U. Malik, "A Python based application for automated soil classification using the Unified Soil Classification System (USCS)," Int. J. Mobile Applications and Technologies, vol. 1, no. 2, 2025.

[2] C.-Y. Liu, C.-Y. Ku, T.-Y. Wu, and Y.-C. Ku, "An advanced soil classification method employing the Random Forest technique in machine learning," Applied Sciences, vol. 14, no. 16, p. 7202, Aug. 2024.

[3] Y. Aydin, U. Isikdag, G. Bekdas, S. M. Nigdeli, and Z. W. Geem, "Use of machine learning techniques in soil classification," Sustainability, 2022.

[4] M. M. Abrar and M. R. Hasan, "Classification of fine-grained soils using two soil classification systems: a case study," Geotechnical Research, Sep. 2025.

[5] M. Jameel and Y. Chowdhary, "Soil quality prediction and classification using machine learning algorithms," J. Research in Engineering and Computer Sciences, vol. 3, no. 3, pp. 41-47, Jun. 2025.

[6] D. T. Tran, D. X. Tran, and V. H. Truong, "Machine learning techniques for cohesive soil classification in construction in Vietnam," HCMCOUJS - Engineering and Technology, vol. 15, no. 2, pp. 16-35, 2025.

[7] D. A. Rosas, D. Burgos, J. W. Branch, and A. Corbi, "Automatic determination of the Atterberg limits with machine learning," DYNA, Sep. 2022.

[8] L. Zhu et al., "Prediction of soil shear strength parameters using combined data and different machine learning models," Applied Sciences, vol. 12, no. 10, p. 5100, May 2022.

[9] IS 1498:1970, "Classification and identification of soils for general engineering purposes," Bureau of Indian Standards.

[10] B. Bhattacharya and D. P. Solomatine, "Machine learning in soil classification," Neural Networks, 2006.

[11] C. Reale, B. C. O'Kelly, and M. Pirotton, "Automatic classification of fine-grained soils using artificial neural networks," Int. J. Geotechnical Engineering, 2018.

[12] A. Ghaderi, A. Abbaszadeh Shahri, and S. Larsson, "An artificial neural network-based model for soil type classification using CPT data," Bulletin of Engineering Geology and the Environment, 2019.

[13] I. Esfandiarpour-Boroujeni, F. Shahbazi, H. Siadat, and H. Shirani, "Evaluation of machine learning algorithms for soil classification," Catena, 2020.

[14] S. Rauter and B. Tschuchnigg, "Machine learning approaches for automated soil classification using cone penetration test data," Computers and Geotechnics, 2021.

[15] A. T. Chala and R. Ray, "Machine learning approaches for soil classification using cone penetration test data," Applied Sciences, vol. 13, no. 9, p. 5758, 2023.