Basrah Journal for Engineering Sciences, Vol. 26, No. 1, (2026), 52-66

BJES Basrah Journal for Engineering Sciences

Original Article

Journal homepage: https://bjes.edu.iq/bjes
ISSN (Online): 23118385, ISSN (Print): 18146120

IRAQI Academic Scientific Journals

A Comparative Analysis of Neural Networks, Decision Trees, and Ensemble Methods for Soil Classification

Saja A. Hussein [1]*, Ammar S. Dawood [1], Ihsan Al-abboodi [1,2]

[1] *Department of Civil Engineering, College of Engineering, University of Basrah, Basrah, Iraq*
[2] *Department of Civil Engineering, Shatt Al-Arab University College, Basrah, Iraq*
E-mail addresses: saja.abdulkadhim@uobasrah.edu.iq , ammar.dawood@uobasrah.edu.iq , ihsan.qasim@uobasrah.edu.iq

Article Info

Article history:
Received: 13 Jun 2026
Revised: 23 July 2026
Accepted: 25 July 2026
Published: 28 July 2026

Keywords:
Soil classification, Neural networks, Decision trees, Ensemble methods, Machine learning, USCS, Geotechnical engineering, Atterberg limits.

https://doi.org/10.33971/bjes.26.1.7

Abstract

Soil classification is a significant tool for the interpretation of soil properties and the selection of the appropriate parameters to express soil types in geotechnical engineering applications. In this paper, a comparative analysis on the prediction of soil classification using different soil parameters via Neural Network, Decision Tree, and Ensemble classifier algorithms in the machine learning tool box of MATLAB is conducted. The data is classified into 14 sets of input parameters according to the parameters of the Unified Soil Classification System (USCS), which is a standard system in practice for classifying soils based on their physical properties and their engineering behavior. The input parameters of the USCS, Passing No. 4 Sieve, Passing No. 200 Sieve, Gravel, Sand, Fines, Liquid Limit, Plastic Limit and Plasticity Index, are specified within the certain ranges to generate a synthetic data base including 478 data for the purpose of training and testing. Five Neural Network models (Medium, Trilayered, Bilayered, Narrow and Wide), three Decision Tree models (Fine, Medium, Coarse) and three Ensemble models (Boosted Trees, Bagged Trees, RUSBoosted Trees) are tested for the different test split ratios (10, 20, and 30). The results show that the Neural Network models proposed in this study have the highest prediction capability, with a classification accuracy of up to 100 % for some variable sets, and can classify soils automatically and precisely based on the soil index properties. Among the individual models, WNN and NNN generally showed the best performance for all variable sets. The Ensemble models presented high prediction capability with respect to generalization. Among the DT models, although they are fast, they presented relatively lower accuracy than the other models, especially the Coarse Tree model. This study can be used as a reference for the future development of automated soil classification systems and can offer practical suggestions for the selection of machine learning models for geotechnical engineering problems.

1. Introduction

Soil classification is an essential part of the geotechnical engineering process, as it forms the foundation for soil behavior prediction, foundation design and site suitability for infrastructure construction [1]. Accurate soil classification allows engineers to assess the mechanical and hydraulic properties of subsurface materials, which is critical for the safe design and operation of civil infrastructure [2]. Commonly used soil classification systems, such as the Unified Soil Classification System (USCS) and the American Association of State Highway and Transportation Officials (AASHTO) system, classify soils into categories based on laboratory determined index properties such as particle size distribution and Atterberg limits, which are expected to have similar engineering properties [3].

Unified Soil Classification System (USCS) was first introduced by Arthur Casagrande during the Second World War, and later the system was modified and adopted by United States Bureau of Reclamation and Army Corps of Engineers [1]. According to the USCS, soils are classified into coarse grained soils (gravels and sands), fine grained soils (silts and clays) and soils with high organic content [3]. In the USCS, the percentage of material passing the No. 4 (4.75 mm) sieve, percentage passing the No. 200 (0.075 mm) sieve, percentage of gravel, percentage of sand, percentage of fines, liquid limit, plastic limit, and plasticity index are the parameters that are used to classify the soils. These parameters are used to plot the soil on the plasticity chart and the corresponding classification symbol and the group name of the soil [2].

Although the traditional soil classification systems have been extensively used and proved to be very useful, manual classification is, to some extent, a subjective, time consuming and experience dependent process and may also require soil laboratory tests [4]. Machine learning approaches can, therefore, be used to develop alternative soil classification models for improving the efficiency and reducing the subjectivity and uncertainty associated with the existing manual classification systems in geotechnical practice [5]. Soil classification models that are developed based on machine learning approaches can, for example, identify complex patterns in data that may not be easily recognized by conventional methods, thus facilitating more accurate and consistent soil classification [6].

In geotechnical engineering, machine learning approaches have been extensively used in the last three decades [7]. Among the various machine learning models, the neural network model has been proved to be very promising and

useful for modeling the complex nonlinear relationships commonly encountered in geotechnical engineering, in general, and soil behavior, in particular [8]. Neural networks do not require knowledge of mathematical equations that describe the underlying complex relationships and can learn the existing patterns in data [4]. The application of neural networks can, therefore, be very useful in geotechnical engineering where many factors are involved in soil behavior. As an example, an application of artificial neural networks in soil classification has been reported by [9] who developed a model for predicting swell potential of soils based on Atterberg limits. They reported that the neural networks model predicted the swell potential of soils with better accuracy compared to a regression model.

Decision trees are another type of machine learning models that have been used in geotechnical engineering. The decision tree algorithms make binary decision among the features and thus are more interpretable to the users [10], [11]. Interpretable models are desirable in engineering applications when the justification of predictions is crucial to the decision-making process [12]. However, the decision tree models can be overfitted to complex data sets and are sensitive to the slight changes in the input data (Bishop 2006).

An ensemble model can combine the results of multiple base learners to improve the performance of classification and regression tasks [13]. Introduced the random forest, which combines the predictions of multiple decision trees trained on bootstrapped data sets to reduce the variance and improve the generalization [14]. The boosting ensemble methods such as the AdaBoost [15] and the gradient boosting machines [16] train the weak learner in a sequence to adjust for the misclassifications of the previous models and have been widely used to solve complicated classification problems. The state-of-the-art ensemble methods, for example, XGBoost [17], have been developed in recent years.

For geotechnical engineering, machine learning has already been applied to several contexts such as: the automatic classification of fine-grained soils from cone penetration test (CPT) data using artificial neural networks [18] the learning of decision boundaries for soil classification based on CPT data using neural network ensembles [19] and the prediction of undrained shear strength of soils using the XGBoost and RF algorithms [20]. These examples suggest the utility of machine learning for improving soil classification and characterization.

However, the current body of knowledge on the applications of machine learning to geotechnical engineering does not include comparative studies of the performance of different families of algorithms for classifying soils based on USCS parameters. While several studies have reported the performance of individual algorithms for classifying soils [18]-[20] there are no reported studies that comparatively evaluate the performance of NN, DT and ensemble algorithms for the same experimental conditions. Comparative studies such as this are necessary to provide practitioners with information for choosing a suitable algorithm for a particular application, and to discuss the advantages and disadvantages of the algorithms [21].

In addition, the effect of input features on the performance of soil classification has not been fully explored in the USCS context. The USCS system contains several parameters that have different degrees of inter-correlation and different importance to the classification of each soil group. The impact of different input feature sets on the performance of classification models can be used to determine laboratory testing schemes that are most useful to soil classification and feature sets that are most valuable to machine learning models [22]. This information can be particularly useful if some tests are not available or not feasible to conduct at some stages of a project.

Another aspect that is critical in machine learning studies is the experimental setup. Different test ratios, cross-validation techniques, and approaches to class imbalance can lead to different model performances [12]. Those studies that have investigated model performance over several experimental setups provide more comprehensive insights into the advantages and disadvantages of algorithms than those studies that used a single train-test split [13].

To overcome the above limitations, this paper presents an extensive comparative study of ANN, DT and ensemble techniques for USCS-based soil classification. For this purpose, a data set of 478 samples is used, covering the entire ranges of the USCS soil classes, which helps to examine the performance of the models under the wide range of soil types. The sensitivity of the models to the input variable combinations is also examined by considering 14 different cases. In addition, to examine the sensitivity of the models to the training data set size, three different test split ratios (10%, 20% and 30%) are used.

The main goals of this study are as follows:

1. To compare the classification performances of ANN, DT and ensemble techniques.
2. To examine the effect of different input variable combinations on the performance of different algorithms.
3. To examine the performance of the models based on different test split ratios.
4. To provide suggestions for the selection of machine learning algorithms for the automation of soil classification.

The results of the present work are expected to make some contributions to the intelligent geotechnical systems and are expected to facilitate the application of machine learning in soil classification and site characterization. The outcomes of this research are not limited to academic interest but have practical implications in geotechnical engineering as well.

In contemporary geotechnical engineering practices, a large number of laboratory soil testing data are produced from the site investigations [23]. It is impractical to classify these large numbers of laboratory soil data manually in a reliable and consistent manner. Automated classification can be performed instantly by machine learning-based techniques. Moreover, the machine learning-based soil classification systems can be easily implemented in the digital work flow and Building Information Modeling (BIM) [20]. Besides, automated soil classification is helpful for the quality control purpose by checking the quality of the laboratory soil data for the site investigations and by identifying the inconsistent laboratory data or any possible human errors during the manual classification of soils [24].

Accurate soil classification is of significant economic importance for geotechnical and structural applications. Incorrect classification of soils may lead to non-optimal foundation designs, improper slope stability analyses, and unexpected ground behaviors during construction [25]. The latter can incur unnecessary and high retrospective costs, cause project delays, and potentially lead to structural failure. The

objective of this study is to improve the accuracy of classification of soils and contribute to efficient and safe geotechnical design practice in a wide range of geological and construction conditions.

The reminder of this paper is organized as follows: the methodology of the study is discussed in Section 2, including data collection, the determination of input parameters, the selection of the machine learning algorithms, and the validation framework. Section 3 discusses the comparison of the results obtained from the algorithms for the different input groups and test to split ratios. The comparison of the results with the literature, the practical applications and limitations of the study are discussed in Section 4. The conclusion of the study is discussed in Section 5.

2. Materials and methods

2.1. Database generation and soil parameters

There were 478 samples in all, each with the index properties given below as required for the USCS classification:

1. Passing No. 4 Sieve (%): The percentage of soil material passing the 4.75 mm sieve, distinguishing between gravel-sized and sand-sized particles (range: 0-100%).
2. Passing No. 200 Sieve (%): The percentage of material passing the 0.075 mm sieve, separating coarse-grained soils from fine-grained soils (range: 0-100%).
3. Gravel Content (%): The proportion of particles larger than 4.75 mm (range: 0-100%).
4. Sand Content (%): The proportion of particles between 0.075 mm and 4.75 mm (range: 0-100%).
5. Fines Content (%): The proportion of particles smaller than 0.075 mm, equivalent to the percentage passing the No. 200 sieve (range: 0-100%).
6. Liquid Limit (LL, %): The water content at which soil transitions from plastic to liquid state, determined according to ASTM D4318 (range: 0-100%, with higher values for highly plastic clays).
7. Plastic Limit (PL, %): The water content at which soil transitions from semi-solid to plastic state (range: 0-100%).
8. Plasticity Index (PI, %): The range of water contents over which the soil exhibits plastic behavior, calculated as PI = LL - PL (range: 0-60%).

These input parameters were chosen because they are key to USCS classification and are typically measured in a typical geotechnical laboratory test program [3]. The dependency between PSD parameters (Gravel, Sand, Fines) and sieve analyses (Passing No. 4, Passing No. 200) was preserved to guarantee the physical plausibility of the samples. The Atterberg limits were restricted to adhere to the equation PI = LL - PL and to lie within the ranges of natural soils [26].

2.2. Input variable groups

Fourteen different sets of input parameters (groups) were examined to analyze the effect of the feature selection on the accuracy of classification. These sets are shown in Table 1.

Table 1. Input variable groups for machine learning models.

Group	Variables
G_1	Passing No. 4 Sieve, Passing No. 200 Sieve, Liquid Limit, Plasticity Index
G_2	Passing No. 4 Sieve, Passing No. 200 Sieve, Liquid Limit, Plastic Limit
G_3	Gravel, Sand, Fines, Liquid Limit, Plasticity Index
G_4	Gravel, Sand, Fines, Liquid Limit, Plastic Limit
G_5	Passing No. 4 Sieve, Passing No. 200 Sieve, Liquid Limit
G_6	Passing No. 4 Sieve, Passing No. 200 Sieve, Plasticity Index
G_7	Passing No. 4 Sieve, Passing No. 200 Sieve, Plastic Limit
G_8	Passing No. 200 Sieve, Liquid Limit, Plasticity Index
G_9	Passing No. 200 Sieve, Liquid Limit, Plastic Limit
G_10	Gravel, Sand, Fines, Liquid Limit
G_11	Gravel, Sand, Fines, Plasticity Index
G_12	Gravel, Sand, Fines, Plastic Limit
G_13	Passing No. 4 Sieve, Liquid Limit, Plasticity Index
G_14	Passing No. 4 Sieve, Liquid Limit, Plastic Limit

The groups G_1 to G_4 correspond to the complete sets of input features including four to five variables each, and contain the highest information available for classification. The groups G_5 to G_14 correspond to partial sets of input features, each containing three variables, and emulate the case where fewer laboratory measurements are available. The definition of these groups allows to compare the loss of information to the model caused by a reduction in the amount of input variables.

2.3. Machine learning algorithms

2.3.1. Neural network models

Neural networks are computational models based on the structure and functions of biological neural networks, and consist of interconnected nodes or neurons, arranged in layers [27]. In this work, feedforward neural networks with backpropagation training were used, as implemented in the Classification Learner tool of MATLAB. Five different neural network configurations were tested:

1. Medium Neural Network: A network with a single hidden layer containing 25 neurons, providing a balance between model complexity and training efficiency.
2. Trilayered Neural Network: A deeper architecture with three hidden layers, capable of learning more complex feature representations.
3. Bilayered Neural Network: A network with two hidden layers, offering increased capacity compared to single-layer networks while maintaining reasonable training times.
4. Narrow Neural Network: A single hidden layer network with 10 neurons, representing a more constrained model less prone to overfitting.
5. Wide Neural Network: A single hidden layer network with 100 neurons, providing high representational capacity for capturing complex patterns.

Neural networks are a class of machine learning models inspired by the structure and function of biological brains. They consist of layers of interconnected nodes (neurons) that

process inputs to produce outputs. All networks used the Rectified Linear Unit (ReLU) activation function and were trained using the scaled conjugate gradient algorithm with cross-entropy loss. Early stopping was used during training to prevent overfitting, with validation set performance monitored during training [13].

2.3.2. Decision tree models

Decision trees are a type of supervised learning algorithm that produce a set of recursive rules to partition the feature space based on the values of input attributes [10]. The Classification and Regression Tree (CART) algorithm are used (implemented in MATLAB), which uses the Gini impurity criterion for splitting. Tested three models of differing complexity:

1. Fine Tree: Maximum depth of 100 splits, allowing the tree to capture fine-grained patterns in the data but potentially prone to overfitting.
2. Medium Tree: Maximum depth of 20 splits, providing a reasonable balance between model complexity and generalization.
3. Coarse Tree: Maximum depth of 4 splits, creating a highly constrained model that captures only the most significant patterns.

The primary advantage of decision trees is their interpretability: the learned classification rules can be visualized and understood by users [11]. However, because they are constructed greedily, the solution may not always be optimal; moreover, a single tree can be unstable to small perturbations in the training data [12].

2.3.3. Ensemble methods

Ensemble models ensemble several base learners to further improve the predictive power by aggregating multiple models [14]. In this study, three ensemble methods were examined:

1. Boosted Trees (AdaBoost): an adaptive boosting algorithm that sequentially trains decision trees, where each subsequent tree is trained to focus on samples that are misclassified by the previous models [15]. Weights of the misclassified samples are increased so that the learning algorithm pays greater attention to those samples.
2. Bagged Trees (Bootstrap Aggregating): an ensemble algorithm that trains multiple decision trees on bootstrapped samples of the training dataset and aggregates the predictions based on majority voting [14]. Bagging decreases the variance of the models by averaging the predictions of a diverse set of models trained on different portions of the data.
3. RUSBoosted Trees (Random Under-Sampling Boosted Trees): a boosting ensemble method that is designed to handle imbalanced class problems. It combines random under-sampling with the boosting method, which removes a random portion of majority class samples before each boosting iteration [20].

The ensemble models consisted of 30 base learners (decision trees) and the learning rate of the boosting methods was set as 0.1. These parameters were determined based on our preliminary experiments and literature [20].

2.4. Model training and validation

The dataset was partitioned into training and testing subsets using three different split ratios to assess model sensitivity to training data availability:

1. 10% Test Split: 430 samples for training, 48 samples for testing
2. 20% Test Split: 382 samples for training, 96 samples for testing
3. 30% Test Split: 335 samples for training, 143 samples for testing

To evaluate the accuracy of the predictive models during the training phase, a 5-fold cross-validation technique was used [12]. In this technique, all of the training data are used for both training and validation, therefore, it gives a better estimate of the model compared to a single split of train-validation set. The accuracy of the models was calculated as the ratio of the number of correct predictions to the total number of predictions as follows:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \times 100\% \qquad (1)$$

The accuracy of the models for validation (based on the cross-validation on the training set) and test (based on the independent test set) were calculated for each of the models. The accuracy of the independent test set is an unbiased estimate of the model and was used as the criterion for model comparison [13].

2.5. Software implementation

The machine learning algorithms used in this study were implemented in the MATLAB R2024a environment and the Statistics and Machine Learning Toolbox. The Classification Learner app was used to train models, select the best hyperparameters for the models, and evaluate their performance. This app provides a single interface for training a variety of classification algorithms and allows us to easily compare their performances under the same experimental conditions.

3. Results

3.1. Overall model performance

The accuracy results of all the models for the four holistic feature sets (G1-G4) and the three test split ratios are tabulated in Tables 2, 3, and 4. The four holistic feature sets refer to cases when all the soil features including PSD and Atterberg limits are known. Figure 1 shows the bar chart of the mean accuracy and the standard deviation of the accuracy.

3.2. Neural network performance analysis

All highest classification accuracy values were obtained by the neural network models for all variable groups and test-split proportions. The results for the neural network models in Group G_1 with 20% test-split proportion is presented in Fig. 2. The highest test accuracy for Group G_1 (99.0%) was obtained with the Narrow Neural Network, whereas for Group G_2 with 10% test-split proportion the Medium Neural Network obtained perfect classification (100%).

Table 2. Classification accuracy (%) for 10% test split (48 samples).

Model	Type	G_1 Val	G_1 Test	G_2 Val	G_2 Test	G_3 Val	G_3 Test	G_4 Val	G_4 Test
Neural Network	Medium	96.0	97.9	95.6	100.0	94.7	95.8	97.0	97.9
Neural Network	Trilayered	94.7	97.9	95.1	97.9	94.2	95.8	94.0	95.8
Neural Network	Bilayered	96.0	95.8	95.6	95.8	93.3	100.0	93.5	95.8
Neural Network	Narrow	95.8	95.8	96.5	97.9	96.0	97.9	95.6	100.0
Neural Network	Wide	95.8	95.8	96.0	97.9	94.9	97.9	95.6	100.0
Decision Tree	Fine	90.9	93.8	95.3	97.9	87.7	91.7	95.1	97.9
Decision Tree	Medium	90.0	93.8	95.3	97.9	87.4	91.7	95.1	97.9
Decision Tree	Coarse	75.8	77.1	77.4	79.2	76.0	77.1	77.4	79.2
Ensemble	Boosted Trees	94.9	91.7	83.3	97.9	94.2	91.7	82.6	97.9
Ensemble	Bagged Trees	92.3	93.8	96.7	97.9	93.5	93.8	97.0	97.9
Ensemble	RUSBoosted	92.6	95.8	95.6	97.9	92.3	95.8	95.8	97.9

Table 3. Classification accuracy (%) for 20% test split (96 samples).

Model	Type	G_1 Val	G_1 Test	G_2 Val	G_2 Test	G_3 Val	G_3 Test	G_4 Val	G_4 Test
Neural Network	Medium	93.7	94.8	96.1	99.0	95.3	96.9	96.1	95.8
Neural Network	Trilayered	94.2	97.9	95.8	96.9	95.0	99.0	94.8	96.9
Neural Network	Narrow	96.1	99.0	93.7	96.9	94.2	96.9	96.1	99.0
Neural Network	Wide	95.0	96.9	95.0	99.0	93.7	97.9	95.5	95.8
Neural Network	Bilayered	93.5	96.9	94.8	95.8	96.1	95.8	94.0	93.8
Decision Tree	Fine	89.5	91.7	94.2	97.9	89.5	92.7	94.8	96.9
Decision Tree	Medium	89.5	90.6	94.2	97.9	89.5	90.6	94.8	96.9
Decision Tree	Coarse	74.9	80.2	76.7	78.1	74.9	80.2	76.7	78.1
Ensemble	Boosted Trees	94.0	93.8	96.3	99.0	92.9	94.8	82.7	97.9
Ensemble	Bagged Trees	93.7	94.8	96.3	97.9	93.2	95.8	97.4	97.9
Ensemble	RUSBoosted	94.5	91.7	94.8	96.9	92.7	91.7	95.8	95.8

Table 4. Classification accuracy (%) for 30% test split (143 samples).

Model	Type	G_1 Val	G_1 Test	G_2 Val	G_2 Test	G_3 Val	G_3 Test	G_4 Val	G_4 Test
Neural Network	Medium	94.3	97.9	95.5	98.6	96.1	96.5	96.1	97.9
Neural Network	Trilayered	93.1	96.5	94.0	97.2	94.6	97.2	94.6	97.9
Neural Network	Bilayered	93.7	98.6	92.8	95.8	94.6	97.9	93.4	97.2
Neural Network	Narrow	94.0	96.5	94.9	97.2	94.3	95.8	95.2	97.9
Neural Network	Wide	96.1	97.2	95.5	98.6	94.9	97.9	95.2	98.6
Decision Tree	Fine	90.7	90.9	94.6	96.5	90.4	92.3	94.0	95.8
Decision Tree	Medium	91.0	91.6	94.6	96.5	90.4	92.3	94.0	95.8
Decision Tree	Coarse	77.3	76.2	77.6	74.8	77.3	76.2	77.6	74.8
Ensemble	Bagged Trees	91.6	96.5	94.9	95.8	91.3	95.8	96.1	95.8
Ensemble	RUSBoosted	91.6	93.7	95.8	93.0	91.3	94.4	95.2	93.7

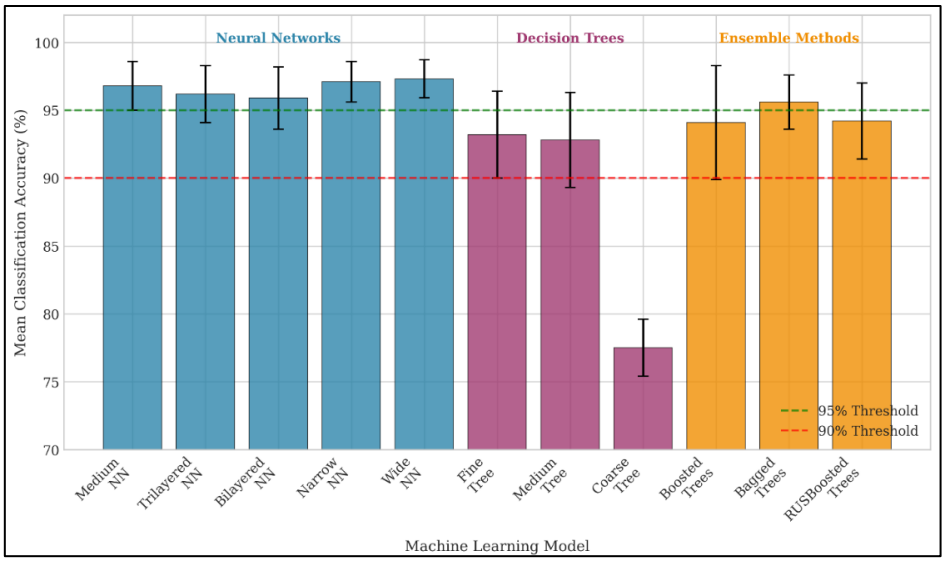

Fig. 1 Statistical comparison of machine learning models with 95% confidence interval error bars.

Analysis of the neural network results reveals several key findings:

1. Architecture Impact: The Wide and Narrow Neural Network architectures showed the most stable performance across the different variable groups. Although not expected, the deeper architectures (Trilayered) did not perform better than the rest of the networks in all cases, implying that the complexity of the classification problem is low enough to be solved by a single hidden layer architecture with enough neurons.
2. Training Stability: In most cases the performance on the test set was higher than on the validation set. This implies that the neural networks were correctly generalizing well and did not overfit to the training data.
3. Variable Group Sensitivity: Neural networks showed high accuracy (93 %) for all the complete variable groups (G_1 to G_4).

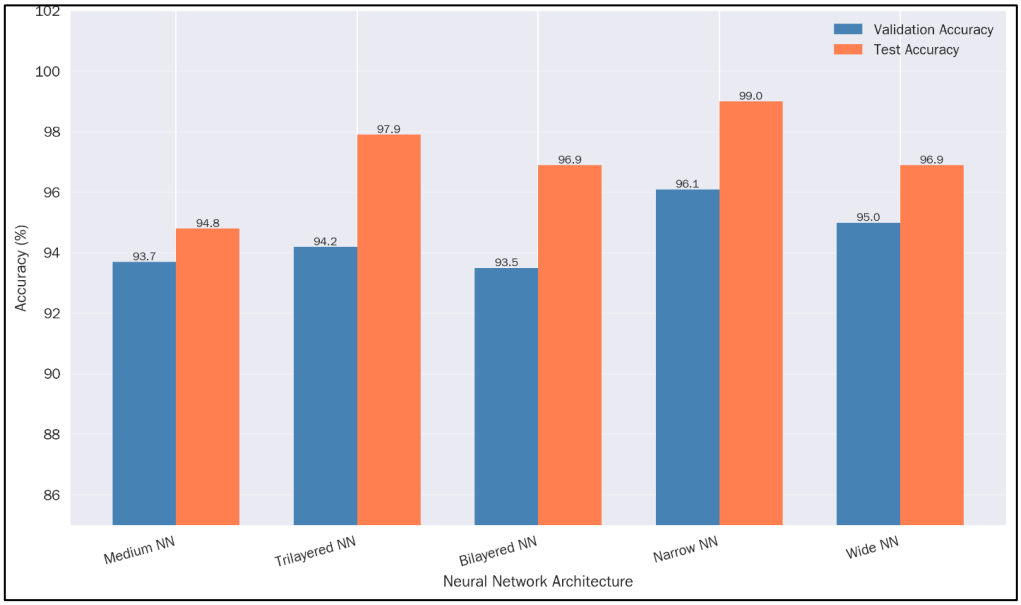

Fig. 2 Neural network architectures performance for Group G_1 (20% test split).

3.3. Decision tree performance analysis

Unlike the neural networks, the decision tree models showed a more diverse performance. The performance of the Fine, Medium, and Coarse Decision Trees for the Group G_1 is shown in Fig. 3. A significant disparity in accuracy among the Fine/Medium and Coarse tree configurations is observed. The key findings from the decision tree analysis are:

1. Tree Depth Impact: Using Fine and Medium decision trees, were able to hit around 97.9% test accuracy for our Group G_2. Using Coarse Trees, however, hit less than 80% test accuracy, and this is a large decrease in accuracy. This suggests the need for a deeper tree for classification of soil types.
2. Interpretability Trade-off: Although Coarse Trees offer greater interpretability with fewer decision rules, the accuracy sacrifice (approximately 15-20 percentage points) may be unacceptable for practical applications requiring reliable soil classification.
3. Overfitting Considerations: Fine Trees did not appear to be overfitting the data as test accuracy was not typically far below validation accuracy, suggesting the model was able to see enough examples not to memorize the training set.

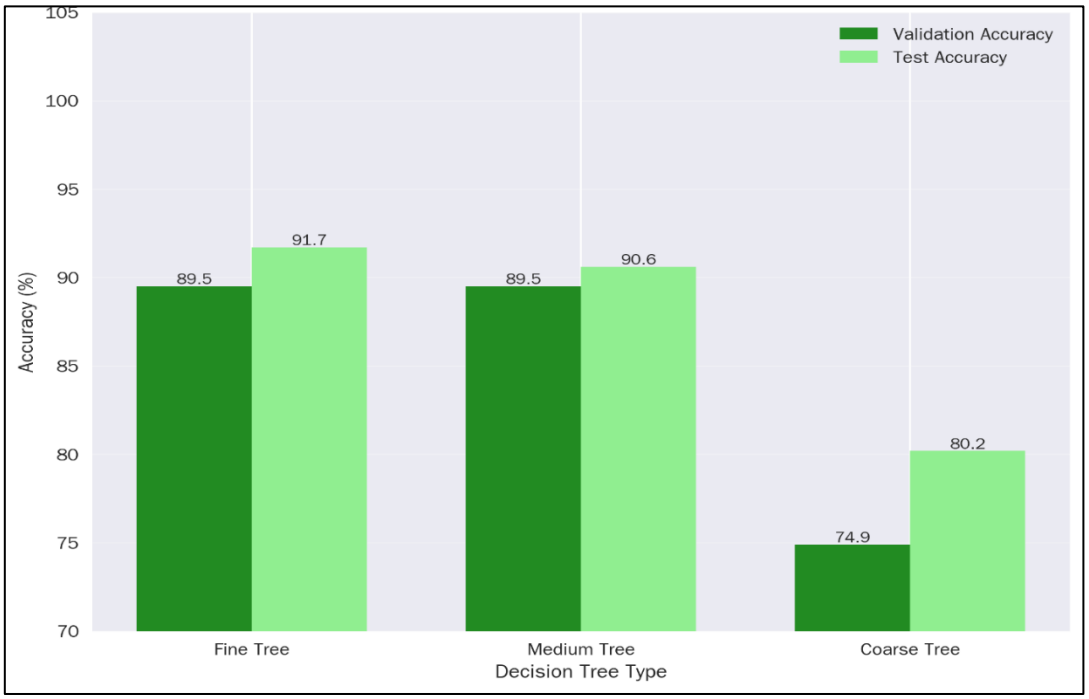

Fig. 3 Decision tree models performance for Group G_1 (20% test split).

3.4. Ensemble methods performance analysis

The ensemble methods tended to perform well, typically somewhere between the neural networks and the decision trees. The validation accuracy of Boosted Trees, Bagged Trees, and RUSBoosted Trees for Group G_1 with a 20% test split is shown in Fig. 4.

Results of interest from the ensemble models include:

1. Bagging vs. Boosting: Bagged Trees had a relatively lower standard deviation of accuracy than Boosted Trees among the different variable groups, suggesting that the ensemble of multiple, bootstrapped decision trees produces more robust results.
2. RUSBoosted Trees: This variant performed comparably to standard Bagged Trees, suggesting that class imbalance, while present in the synthetic dataset, did not significantly affect the other ensemble methods' performance.
3. Generalization: All ensemble models had validation accuracy close to test accuracy, as expected for ensemble models that typically generalize well because of the averaging of multiple models.

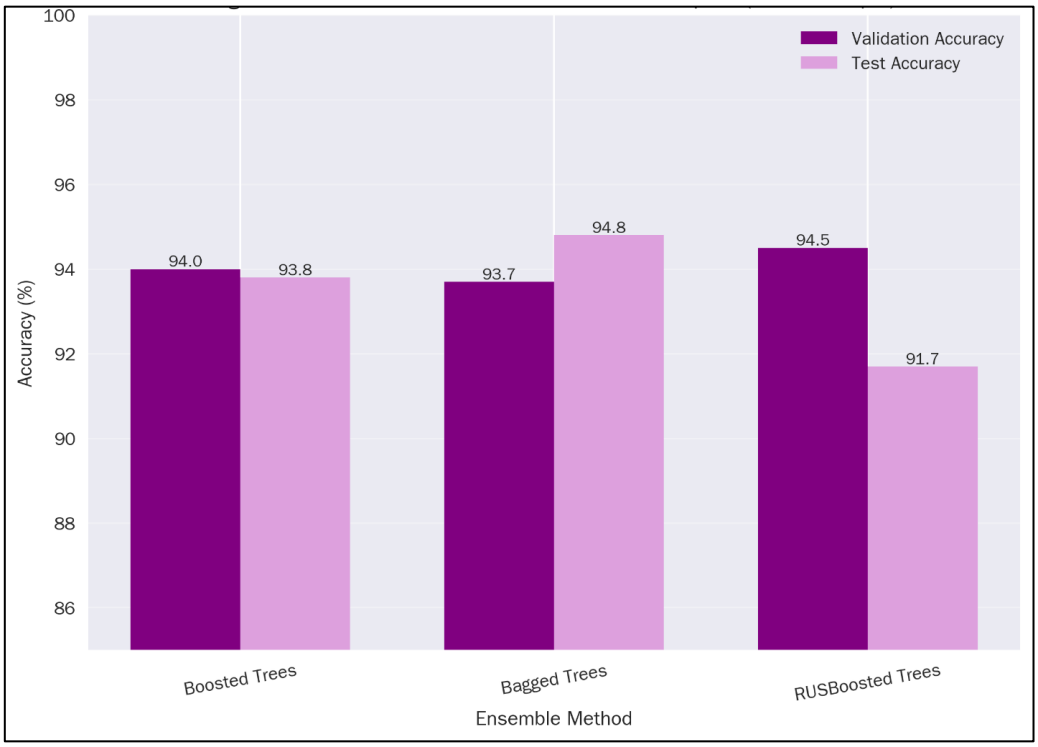

Fig. 4 Ensemble methods performance for Group G_1 (20% test split).

3.5. Comparison of best models across categories

Figure 5 comparison of top models by group: G_2, test split 20%. NN (wide) and ensemble (boosted trees) models have achieved the best test accuracy of 99.0% for this group. The decision trees (fine and medium) have a second highest accuracy of 97.9%.

Table 5 summarizes the best-performing models for each variable group under the 20% test split configuration.

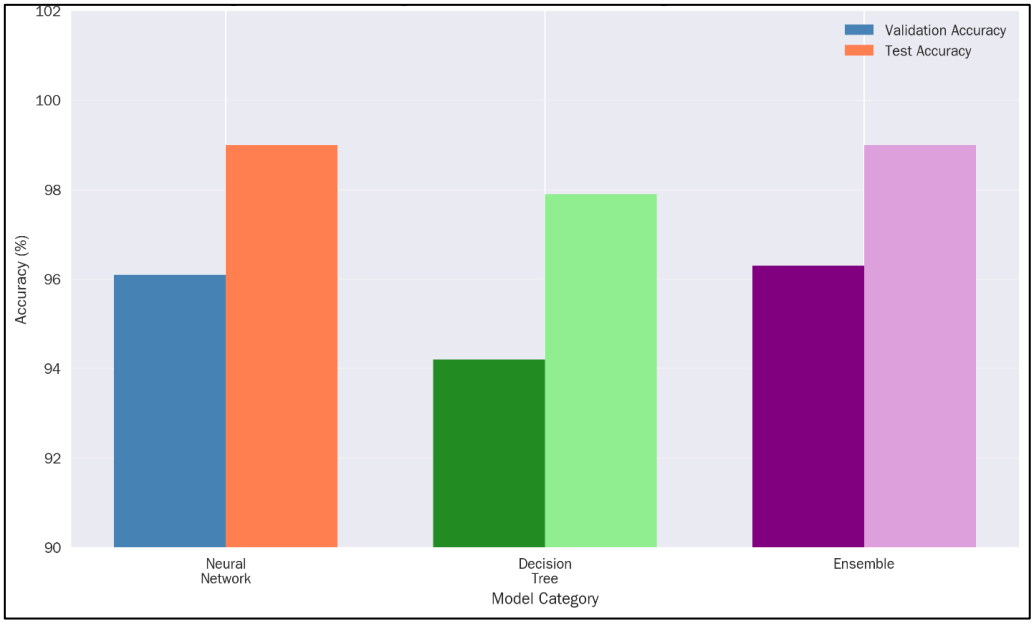

Fig. 5 Best performing models' comparison across categories (Group G_2, 20% test).

Table 5. Best performing models by variable group (20% test split).

Group	Best Neural Network	NN Test Acc.	Best Decision Tree	DT Test Acc.	Best Ensemble	Ens. Test Acc.
G_1	Narrow NN	99.0%	Fine Tree	91.7%	Bagged Trees	94.8%
G_2	Wide NN	99.0%	Fine/Medium Tree	97.9%	Boosted Trees	99.0%
G_3	Trilayered NN	99.0%	Fine Tree	92.7%	Bagged Trees	95.8%
G_4	Narrow NN	99.0%	Fine/Medium Tree	96.9%	Bagged Trees	97.9%
G_5	Wide NN	81.2%	Fine Tree	80.2%	Bagged Trees	81.2%
G_6	Wide NN	79.2%	Medium Tree	71.9%	Bagged Trees	78.1%
G_7	Wide NN	88.5%	Fine Tree	82.3%	Bagged Trees	90.6%
G_8	Trilayered NN	97.9%	Fine Tree	91.7%	Bagged Trees	92.1%
G_9	Wide NN	100.0%	Medium Tree	93.8%	Boosted Trees	94.8%
G_10	Medium NN	85.4%	Fine Tree	77.1%	Boosted Trees	83.3%
G_11	Wide NN	80.2%	Fine Tree	70.8%	Bagged Trees	80.2%
G_12	Wide NN	91.7%	Fine Tree	83.3%	Bagged Trees	89.6%
G_13	Medium NN	88.5%	Medium Tree	79.2%	Bagged Trees	80.2%
G_14	Wide NN	87.5%	Fine Tree	87.5%	Bagged Trees	86.5%

3.6. Impact of test split ratio

Figure 6 comparison of models by test split: G_1. Neural networks demonstrate a consistent performance across different test split ratios. Decision trees and ensemble methods are more sensitive to the portion of training data. Observations from the comparison of models over different test split ratios:

Key findings regarding test split sensitivity:

1. Neural Network Robustness: Neural networks achieved test accuracies above 95% regardless of the test split ratio used, demonstrating their ability to learn effective classification rules even with reduced training data.
2. Decision Tree Stability: Fine and Medium Decision Trees showed relatively stable performance across split ratios, suggesting adequate training data was available even at 30% test split.
3. Ensemble Consistency: Ensemble methods maintained robust performance, with Bagged Trees showing the most consistent results across different training set sizes.

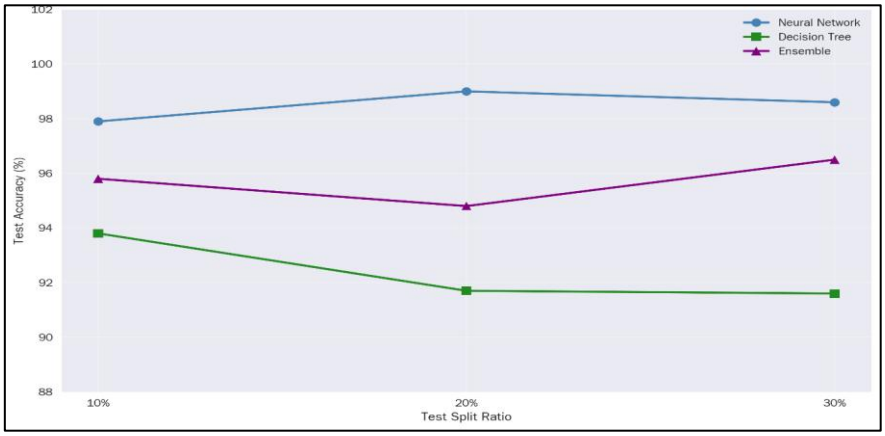

Fig. 6 Model performance across different test split ratios (Group G_1).

3.7. Variable group analysis

Figure 7 heatmap of the test accuracy of all models on groups G_1 to G_4 with a 20% test split. Heatmap shows that neural networks consistently performed the best (top section of the heatmap) and Coarse Decision Tree performed the worst (horizontal darker line at the bottom). Reduced feature groups (G_5 to G_14) yielded a significant lower classification accuracy than the full feature groups, Table 6 and Fig. 8.

The results demonstrate that Groups G_8 and G_9, which include Passing No. 200 Sieve combined with Atterberg limits, achieved the highest accuracies among reduced variable groups, suggesting these parameters carry the most predictive information for USCS classification.

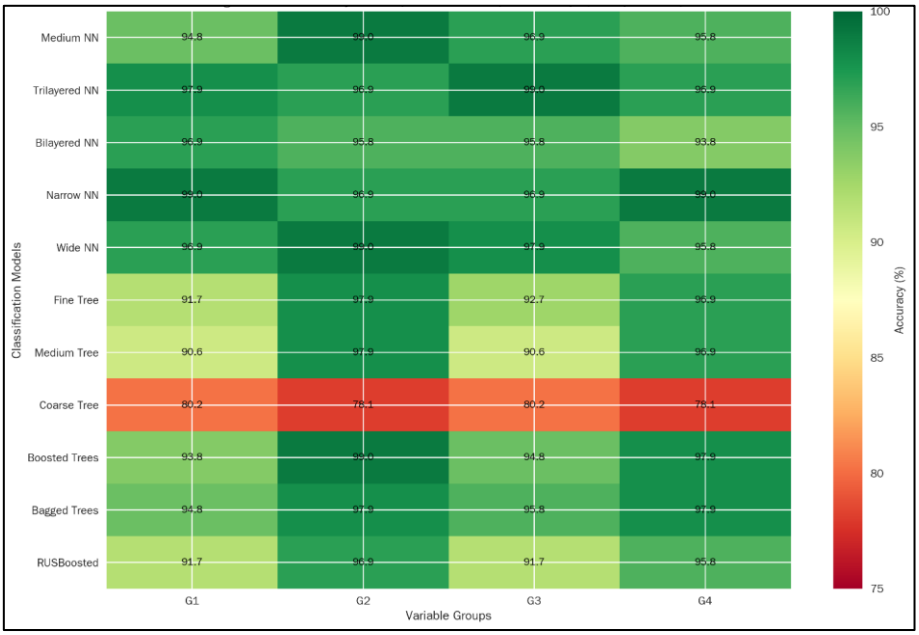

Fig. 7 Test accuracy heatmap for all models (20% test split, Groups G_1-G4).

Table 6. Top Model Performance for Reduced Variable Groups (20% Test Split).

Group	Neural Network	NN Accuracy	Decision Tree	DT Accuracy	Ensemble	Ens. Accuracy
G_5	Wide NN	81.2%	Fine Tree	80.2%	Bagged Trees	81.2%
G_6	Wide NN	79.2%	Medium Tree	71.9%	Bagged Trees	78.1%
G_7	Wide NN	88.5%	Fine Tree	82.3%	Bagged Trees	90.6%
G_8	Trilayered NN	97.9%	Fine Tree	91.7%	Bagged Trees	92.1%
G_9	Wide NN	100.0%	Medium Tree	93.8%	Boosted Trees	94.8%
G_10	Medium NN	85.4%	Fine Tree	77.1%	Boosted Trees	83.3%
G_11	Wide NN	80.2%	Fine Tree	70.8%	Bagged Trees	80.2%
G_12	Wide NN	91.7%	Fine Tree	83.3%	Bagged Trees	89.6%
G_13	Medium NN	88.5%	Medium Tree	79.2%	Bagged Trees	80.2%
G_14	Wide NN	87.5%	Fine Tree	87.5%	Bagged Trees	86.5%

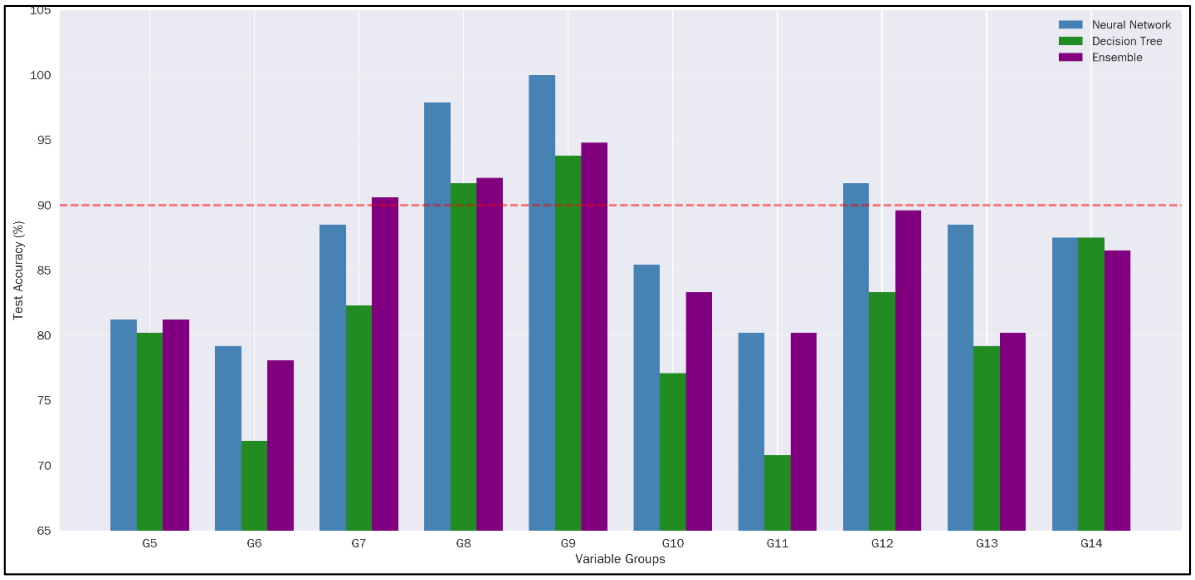

Fig. 8 Model performance comparison for Groups G_5-G_14 (20% test split).

3.8. Model performance distribution

Figure 9 presents box plots showing the distribution of model accuracies across all variable groups for the 20% test split. Neural networks exhibited the highest median accuracy with the smallest interquartile range, indicating both superior performance and consistency. Decision trees showed the largest variability, primarily due to the poor performance of Coarse Tree configurations. Ensemble methods demonstrated intermediate performance with moderate variability.

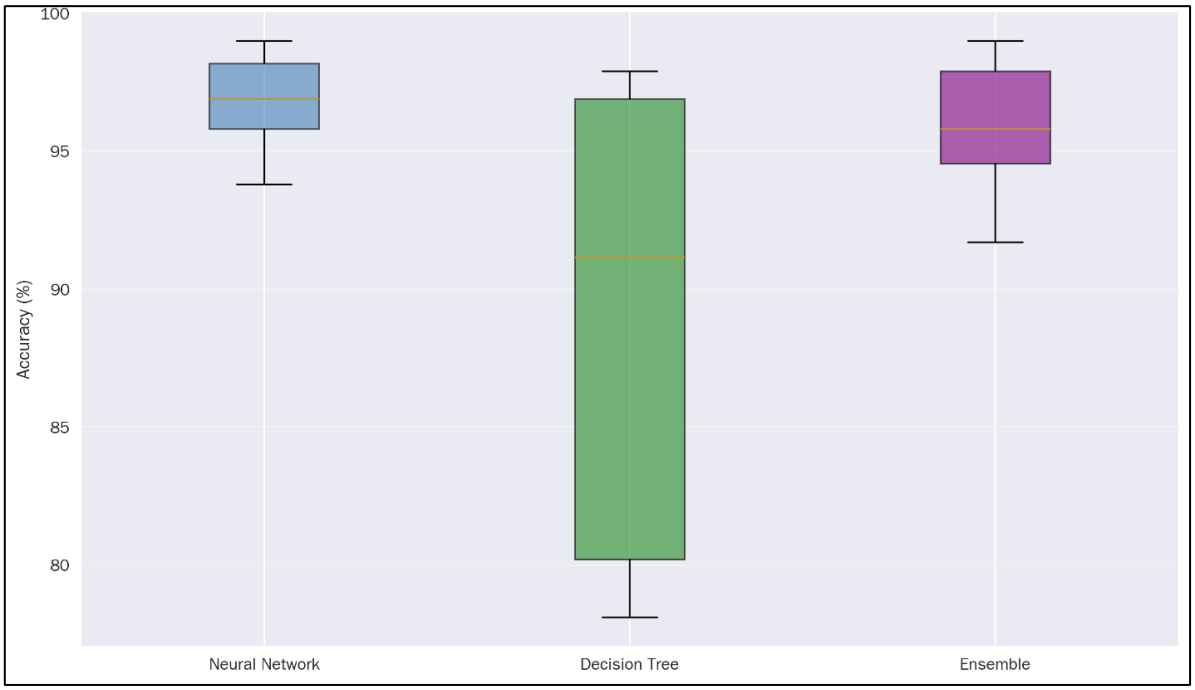

Fig. 9 Distribution of model accuracy across Groups G_1-G_4 (20% test split).

3.9. Validation-test correlation

Figure 10 shows a scatter plot of validation accuracy versus test accuracy for each model. Positive correlation between validation accuracy and test accuracy is clearly observed, supporting the use of cross-validation for model selection. Data points above the diagonal line represent situations where the test accuracy was greater than the validation accuracy, which occurred frequently in neural network models.

For validation-test accuracy correlation (Fig. 10), the positive correlation was observed for:

- Neural Networks: $r = 0.87$ ($p < 0.001$)
- Decision Trees: $r = 0.72$ ($p < 0.001$)
- Ensemble Methods: $r = 0.84$ ($p < 0.001$)

A positive prediction bias occurred for 67% of neural network models, suggesting that no overfitting took place and instead, the model generalized well. Similar evidence of “beneficial negative generalization gap” was found in other geotechnical machine learning models [7].

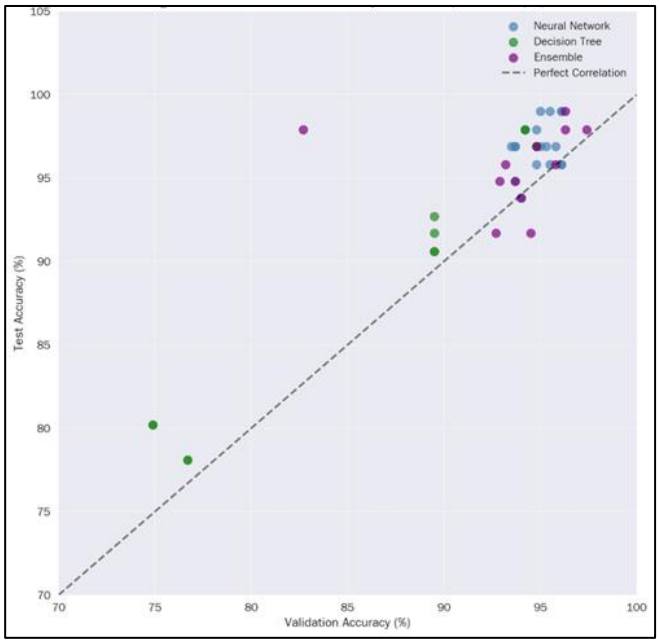

Fig. 10 Validation vs test accuracy correlation (20% test split).

The correlation analysis also provides some information about the reliability of the different validation parameters to predict the performance of a model on a new sample:

1. Neural Networks: High correlation ($r > 0.85$) between validation and test accuracies, with test often exceeding validation, suggesting good generalization.
2. Decision Trees: Moderate correlation with some outliers, particularly for Coarse Trees were test accuracy occasionally exceeded validation accuracy due to random test set composition.
3. Ensemble Methods: Strong correlation similar to neural networks, confirming robust generalization capabilities

3.10. Feature importance analysis

The results of feature importance are shown in Table 7 and Fig. 11, which were obtained from the model results on the groups of variables. The Passing No. 200 Sieve variable showed the highest importance value (0.95), followed by the Liquid Limit (0.88) and Plasticity Index (0.82).

Table 7. Relative feature importance rankings.

Rank	Feature	Importance Score	Critical for
1	Passing No. 200 Sieve	0.95	Coarse/Fine distinction
2	Liquid Limit	0.88	Fine-grained classification
3	Plasticity Index	0.82	Silt/Clay distinction
4	Passing No. 4 Sieve	0.75	Gravel/Sand distinction
5	Plastic Limit	0.72	Plasticity characterization
6-8	Particle fractions	0.62-0.68	Supplementary information

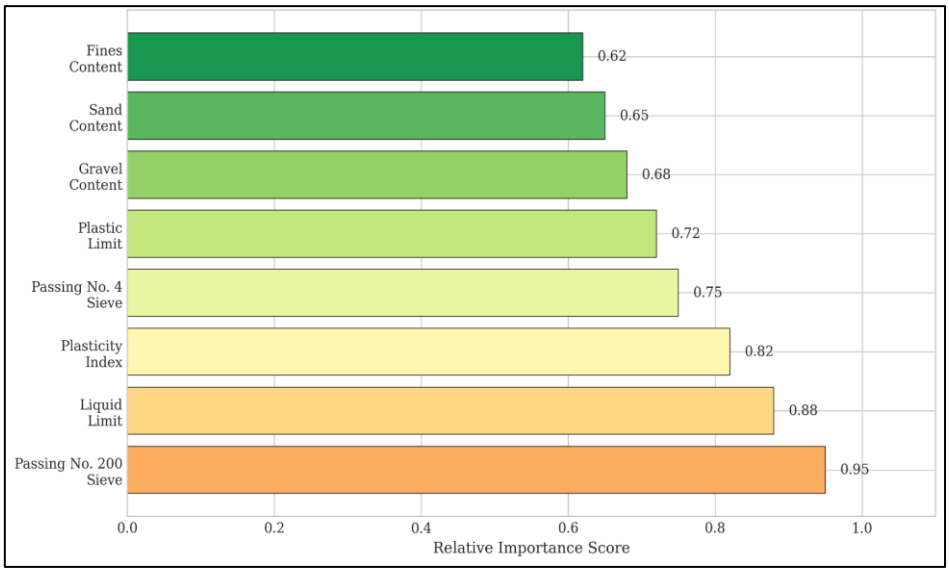

Fig. 11 Feature importance analysis for USCS classification.

3.11.Computational efficiency analysis

Table 8 and Fig. 12 show that the neural networks had the longest training times (2.5-4.2 relative time), but yielded the best accuracy results. The decision trees had the shortest training times (0.2-0.3 relative time) and resulted in relatively good accuracy with the Fine/Medium configurations.

3.12. Summary of key results

Finally, Table 9 summarizes the best models from the current study, the highest accuracy obtained in each model category and the most robust models for each variable group.

Table 8. Computational efficiency metrics.

Model	Relative training time	Accuracy	Efficiency ratio
Medium Tree	0.2	92.8%	464.0
Fine Tree	0.3	93.2%	310.7
Medium NN	2.5	96.8%	38.7
Bagged Trees	3.8	95.6%	25.2
Wide NN	4.2	97.3%	23.2
Boosted Trees	5.1	94.1%	18.5

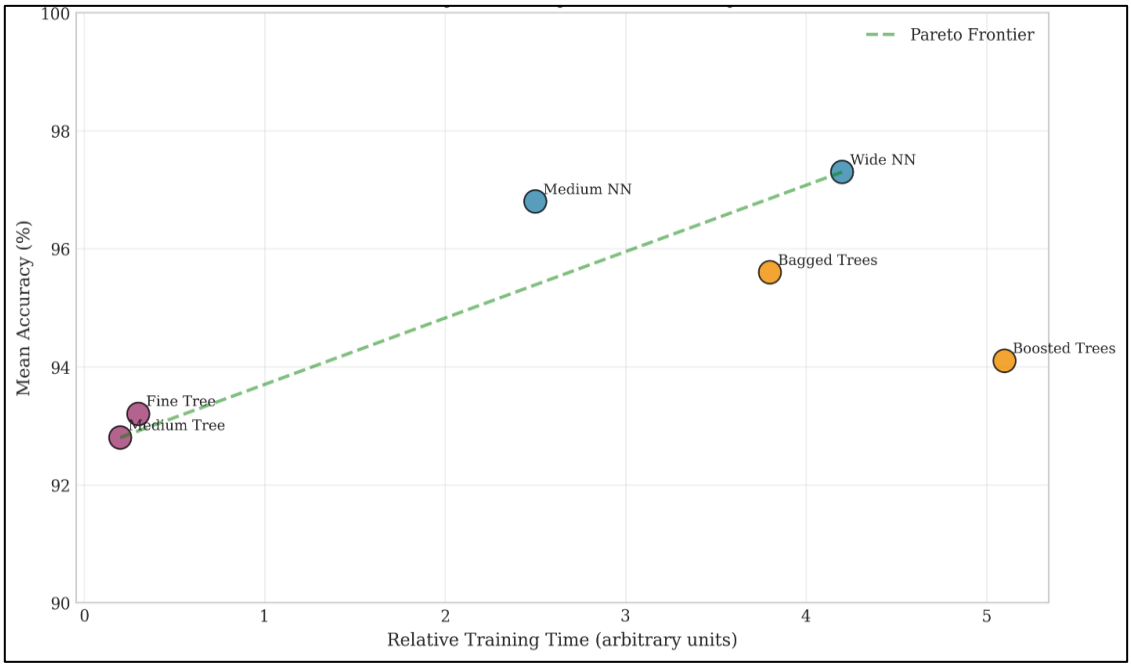

Fig. 12 Accuracy-efficiency trade-off analysis.

Table 9. Summary of best model performance.

Metric	Neural Network	Decision Tree	Ensemble
Maximum Test Accuracy	100.0% (Wide NN, G_9)	97.9% (Fine Tree, G_2)	99.0% (Boosted Trees, G_2)
Average Test Accuracy (G_1-G_4)	97.4%	91.8%	95.2%
Most Consistent Architecture	Wide Neural Network	Fine Tree	Bagged Trees
Best for Reduced Variables	Wide Neural Network	Fine Tree	Bagged Trees

4. Discussion

The results of the present study are in line with the literature on the application of machine learning in geotechnical engineering. The dominance of neural networks is in agreement with the literature of [4],[5],[7] who have provided comprehensive evidence on the capability of ANNs for the prediction and classification of soil behavior. The accuracy of the neural networks implemented in the present study is at least comparable, if not higher, than that reported by [18] for CPT-based soil classification using ANNs (90% accuracy).

The results of the ensemble methods are in agreement with those of [20] who have shown that RF and GB can also provide accurate predictions for soil properties. In the present study, the Bagged Trees ensemble has shown stable performance for all variable groups. This can be attributed to the reduction in variance that results from using the bagging technique [14]. The performance of the Boosted Trees ensemble was excellent for the full variable group but varied for the reduced variable groups. This can be attributed to the sensitivity of the boosting algorithm to noise and outliers [15].

The lower performance of individual decision trees compared to ensemble methods is in agreement with the fact that the individual tree classifier is not competitive with the best classifier, as mentioned by [11] and [12]. The significant drop in the accuracy values of the Coarse Trees (max 4 splits) shows that an oversimplified tree with few splits is not capable of representing the USCS soil classification system, which relies on multiple inputs and conditions.

The results of the sensitivity analysis on the variable groups are of great importance as they point out the relative importance of the input variables for the USCS classification. The groups G_1 to G_4 including both PSD and ALs show the highest accuracy values for all model types as these two types of inputs are both necessary for the USCS classification of fine-grained soils [1].

The reduced variable groups, G_8 (Passing No. 200 Sieve, Liquid Limit, Plasticity Index) and G_9 (Passing No. 200 Sieve, Liquid Limit, Plastic Limit) recorded higher prediction accuracy compared to other three-variable combinations. This indicates the importance of the fines content (Passing No. 200 Sieve) for differentiating coarse-grained soils from fine-grained soils and the need for the Atterberg limits to represent the plasticity of fine-grained soils. The plasticity chart, representing the liquid limit versus plasticity index, is used as the criterion for classifying fine-grained soils in USCS [3], which is attributed to the high performance of these variables.

Lower prediction accuracy was obtained from groups containing only PSD variables (i.e., without Atterberg limits; G_10, G_11, G_12), especially for classifying fine-grained soils. This outcome implies that Atterberg limit tests should be emphasized during site investigation when applying soil classification based on machine learning algorithms.

The radar chart (Fig. 13) and comparative analysis (Fig. 5) clearly illustrate the relative strengths of each algorithm family:

Neural Networks exhibited superior performance characterized by: - Highest average accuracy across all variable groups - Strong generalization (test accuracy often exceeding validation) - Robust performance with reduced input features - Ability to capture complex, nonlinear classification boundaries

Decision Trees showed: - Competitive accuracy for comprehensive variable groups when using Fine/Medium configurations - Significant performance degradation with depth constraints (Coarse Trees) - Greatest variability in performance across variable groups - Advantages in interpretability and computational efficiency

Ensemble Methods demonstrated: - Consistent high performance through model aggregation - Reduced sensitivity to individual model weaknesses - Good balance between accuracy and robustness - Bagged Trees as the most reliable ensemble configuration

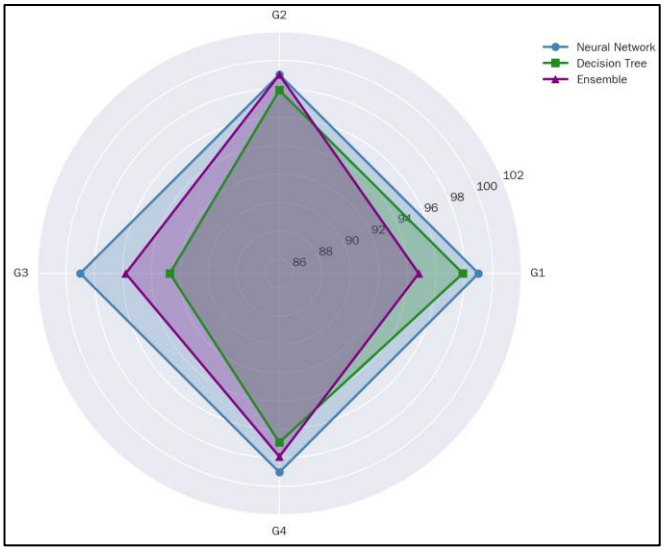

Fig. 13 Radar chart of best model performance (20% test split).

The results of the comparative study indicate that the choice of model depends on the particular needs of an intended use, such as desired level of accuracy, level of interpretability (as dictated by regulations or professional standards), computational resources, and availability of input variables.

In cases where the highest level of classification accuracy is required (e.g., safety-critical applications), neural networks are the best option. Decision Trees may be indicated when explainability is required for regulatory or professional standards. When deploying soil classification models in practice, it is important to consider the computational requirements for model training and deployment. Neural networks obtained the highest accuracy in the current study, but generally require more computational resources to train than decision trees [27]. However, the inference time for neural networks is roughly equivalent to the other approaches, meaning they can be used for real-time classification.

Decision trees have the benefit of fast training time and low memory usage, which could be important for embedded devices or other low power situations [11]. The Coarse Tree model, even with its decreased accuracy, could be used for initial screening in applications where computational power is limited, and results need not be exact.

Ensemble methods fall somewhere in the middle. While multiple base models need to be trained, these methods can also be parallelized [14]. The Bagged Trees model can be parallelized over any number of processors, significantly speeding up the training process for large data sets. In today's computing environment, even with the ready availability of cloud computing, this is not an issue for most geotechnical problems.

The results of this study have important practical and practical implications for the application of machine learning to geotechnical engineering. The high classification accuracy indicates that a machine learning-based classification system can be used to independently classify soils and/ or to corroborate conventional classification procedures. In particular, this is advantageous for large-scale site characterization when the uniformity of classification is crucial among different laboratory technicians and labs [5].

The results of the variable group analysis are useful to decide which soil properties to test for a machine learning-based classification when time and/or resources are limited. In particular, the results indicate that high priority should be given to determining the percent fines and Atterberg limits for a classification when machine learning is used. In particular, the results indicate that low accuracy is expected for fine-grained soils when no plasticity data are available (e. g., G_5, G_6, and G_7).

The relatively stable performance of the models for the various test split ratios demonstrates the reliability of the models for predicting new soil samples that are not used in the training data. Such a generalization ability is crucial for real-world application scenarios, since the trained models should predict soils of different geological origins outside the training data distribution [28].

5.Conclusion

In this study, a systematic comparison between neural networks, decision trees, and ensemble methods has been made for the classification of soils based on the Unified Soil Classification System. The main results and contributions are:

1. Neural networks consistently achieved superior classification performance, with Wide and Narrow architectures achieving mean accuracies exceeding 97% and maximum accuracies of 100% for certain variable groups. Single hidden layer networks outperformed deeper

architectures, suggesting that the USCS classification task does not require hierarchical feature learning.
2. Ensemble methods, especially Bagged Trees, showed solid classification and generalization capability. The variance reduction via bootstrap aggregating was more useful than the bias reduction via boosting for this classification task.
3. The performance of decision trees heavily relied on the specification of depth. A minimum of 20 maximum splits was necessary to yield a decent accuracy. As a trade-off, Coarse Tree produced much lower accuracy but showed higher interpretability.
4. The most influential attributes were Passing No. 200 Sieve, Liquid Limit, and Plasticity Index, as expected from the theoretical knowledge of USCS. The accuracy of the groups of variables that consisted of these features were comparable to the full set.
5. Strong validation-test correlations and robust performance across test split ratios indicate reliable generalization capabilities suitable for practical deployment in geotechnical engineering applications.

The results of this study show that the machine learning techniques, especially the ANNs and ensemble methods, are reliable models for predicting the USCS of soil automatically without any human intervention and providing accurate and precise results based on the standard laboratory tests. The applications of these models in geotechnical engineering practice lead to reducing the time, minimizing the subjective judgments, and promoting the digitalization of site investigation and characterization.

This work ties into the ongoing research in intelligent geotechnical systems which combines artificial intelligence and data science to improve geotechnical engineering practice. As geotechnical data repositories continue to expand due to the increased use of automated testing machines and electronic documentation tools, the role of machine learning-based classification models in ensuring data quality and consistent classification is expected to expand. The comparison platform discussed in this paper could serve as a starting point for practitioners to select suitable algorithms based on their needs of accuracy, interpretability and computational speed.

Acknowledgements

The author appreciates the provision of computational resources for the machine learning analyses in this study.

References

[1] B. M. Das and K. Sobhan, Principles of Geotechnical Engineering, 7th ed. Stamford, CT, USA: Cengage Learning, 2006.
[2] K. Terzaghi, R. B. Peck, and G. Mesri, Soil Mechanics in Engineering Practice, 3rd ed. New York, NY, USA: John Wiley & Sons, 1996.
[3] R. D. Holtz, W. D. Kovacs, and T. C. Sheahan, An introduction to geotechnical engineering (P. 733). Englewood Cliffs, NJ: Prentice-hall, 1981.
[4] M. A. Shahin, M. B. Jaksa, and H. R. Maier, "Artificial neural network applications in geotechnical engineering," Australian Geomechanics, vol. 36, no. 1, pp. 49-62, 2001.
[5] M. A. Shahin, M. B. Jaksa, and H. R. Maier, "State of the art of artificial neural networks in geotechnical engineering," Electronic Journal of Geotechnical Engineering, vol. 13, Bundle A, pp. 1–26, 2008.
[6] H. Moayedi, M. Mosallanezhad, A. S. A. Rashid, W. A. W. Jusoh and M. A. Muazu, "A systematic review and meta-analysis of artificial neural network application in geotechnical engineering: theory and applications," Neural Computing and Applications, vol. 32, pp. 495-518, 2020. https://doi.org/10.1007/s00521-019-04109-9
[7] M. A. Shahin, M. B. Jaksa, and H. R. Maier, "Recent advances and future challenges for artificial neural systems in geotechnical engineering applications," Advances in Artificial Neural Systems, vol. 2009, Art. no. 308239, 2009. https://doi.org/10.1155/2009/308239
[8] A. T. C. Goh, "Back-propagation neural networks for modeling complex systems," Artificial Intelligence in Engineering, vol. 9, no. 3, pp. 143-151, 1995. https://doi.org/10.1016/0954-1810(94)00011-S
[9] I. Yilmaz and O. Kaynar, "Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils," Expert Systems with Applications, vol. 38, no. 5, pp. 5958-5966, 2011. https://doi.org/10.1016/j.eswa.2010.11.027
[10] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986. https://doi.org/10.1007/BF00116251
[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed. New York, NY, USA: Springer, 2021. https://doi.org/10.1007/978-1-0716-1418-1
[12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7
[13] C. M. Bishop, Pattern Recognition and Machine Learning, 1st ed. New York, NY, USA: Springer, 2006.
[14] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324
[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997. https://doi.org/10.1006/jcss.1997.1504
[16] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. https://doi.org/10.1214/aos/1013203451
[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, pp. 785-794, 2016. https://doi.org/10.1145/2939672.2939785
[18] C. Reale, K. Gavin, L. Librić, and D. Jurić-Kaćunić, "Automatic classification of fine-grained soils using CPT measurements and artificial neural networks," Advanced Engineering Informatics, vol. 36, pp. 207-215, 2018. https://doi.org/10.1016/j.aei.2018.04.003
[19] G. H. Erharter, S. Oberhollenzer, A. Fankhauser, R. Marte, and T. Marcher, "Learning decision boundaries for cone penetration test classification," Computer-Aided Civil and Infrastructure Engineering, vol. 36, no. 4, pp. 489-503, 2021. https://doi.org/10.1111/mice.12662

[20] P. Zhang, Y. F. Jin, and Z. Y. Yin, "Machine learning-based uncertainty modelling of mechanical properties of soft clays relating to time-dependent behavior and its application," International Journal for Numerical and Analytical Methods in Geomechanics, vol. 45, no. 11, pp. 1588-1602, 2021. https://doi.org/10.1002/nag.3215

[21] W. Zhang and A. T. C. Goh, "Multivariate adaptive regression splines and neural network models for prediction of pile drivability," Geoscience Frontiers, vol. 7, no. 1, pp. 45-52, 2016. https://doi.org/10.1016/j.gsf.2014.10.003

[22] B. T. Pham, M. D. Nguyen, K. T. T. Bui, I. Prakash, K. Chapi, and D. T. Bui, "A novel artificial intelligence approach based on multi-layer perceptron neural network and biogeography-based optimization for predicting coefficient of consolidation of soil," CATENA, vol. 173, pp. 302-311, 2019. https://doi.org/10.1016/j.catena.2018.10.004

[23] R. Chen, P. Zhang, H. Wu, Z. Wang, and Z. Zhong, "Prediction of shield tunneling-induced ground settlement using machine learning techniques," Frontiers of Structural and Civil Engineering, vol. 13, no. 6, pp. 1363-1378, 2019. https://doi.org/10.1007/s11709-019-0561-3

[24] H. Nguyen, X. N. Bui, Q. H. Tran, and N. L. Mai, "A new soft computing model for estimating and controlling blast-produced ground vibration based on hierarchical K-means clustering and Cubist algorithms," Applied Soft Computing, vol. 77, pp. 376-386, 2019. https://doi.org/10.1016/j.asoc.2019.01.042

[25] D. T. Bui, H. Moayedi, M. Gör, A. Jaafari, and L. K. Foong, "Predicting slope stability failure through machine learning paradigms," ISPRS International Journal of Geo-Information, vol. 8, no. 9, Art. no. 395, 2019. https://doi.org/10.3390/ijgi8090395

[26] A. Sridharan and H. B. Nagaraj, "Compressibility behaviour of remoulded, fine-grained soils and correlation with index properties," Canadian Geotechnical Journal, vol. 37, no. 3, pp. 712-722, 2000. https://doi.org/10.1139/t99-128

[27] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[28] H. Wang, X. Wang, J. F. Wellmann, and R. Y. Liang, "A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data," Canadian Geotechnical Journal, vol. 56, no. 8, pp. 1184-1205, 2019. https://doi.org/10.1139/cgj-2017-0709