International Journal of Applied Science and Engineering

# Machine learning-based soil classification: leveraging resistivity and CPT data for enhanced prediction accuracy

**Nurhasanah [1, 3*], Achmad Bakri Muhiddin [1], Abdul Rachman Djamaluddin [1], Muhammad Niswar [2]**

[1] *Department of Civil Engineering, Hasanuddin University, Gowa, Indonesia*
[2] *Department of Informatics Engineering, Hasanuddin University, Gowa, Indonesia*
[3] *Department of Physics, Tanjungpura University, Pontianak, Indonesia*

## ABSTRACT

Conventional methods provide reliable values of soil properties critical for geotechnical design purposes, but they struggle to handle the complexities of geotechnical data effectively. The growing intricacy of soil properties necessitates a more precise and effective data-driven methodology in geotechnical engineering. Applying advanced methodologies, including machine learning and integrated data, is essential to address these constraints and enhance the accuracy and efficiency of analytical techniques. The study investigates the efficacy of machine learning in enhancing soil classification performance and evaluates the impact of integrating resistivity and CPT data. A detailed dataset incorporating electrical resistivity and key CPT parameters—cone resistance, sleeve friction, friction ratio, and total friction was compiled for model training and testing. Techniques for soil type classification employing machine learning algorithms, such as K-Nearest Neighbours, Random Forest, and Extreme Gradient Boosting. The assessment of the performance of each algorithm was based on some metrics, including accuracy, precision, recall, and F1-score. The study found that the machine learning algorithm effectively identified soil types such as poorly graded sand and silty sand. The integration of resistivity and CPT data led to a marked improvement in classification performance. Random Forest and XGBoost outperform KNN in soil type classification, with Random Forest achieving the best accuracy, precision, recall, and F1 score results. This work highlights the benefits of combining resistivity and CPT data in soil classification and demonstrates Random Forest and XGBoost's superiority in handling intricate, multi-dimensional datasets. These findings suggest that this integrated approach can enhance the accuracy and efficiency of analytical techniques of geotechnical investigations.

***Keywords:*** Cone penetration testing, Data integration, Geotechnical engineering, Machine learning, Soil resistivity, Soil type classification.

**OPEN ACCESS**

**Received:** December 26, 2024
**Revised:** March 13, 2025
**Accepted:** March 30, 2025

**Corresponding Author:**
Nurhasanah
nurhasanah@physics.untan.ac.id



**Publisher:**
Chaoyang University of Technology
**ISSN:** 1727-2394 (Print)
**ISSN:** 1727-7841 (Online)

## 1. INTRODUCTION

Soil classification is a fundamental component of geotechnics, construction, and risk mitigation that underpins engineering projects' safety, efficiency, and sustainability. Two primary methods are electrical resistivity measurements and Cone penetration testing (CPT). Electrical resistivity is a non-invasive geophysical method that measures the resistance of soil to electrical current flow, providing continuous profiles of materials and understanding soil properties (Baker et al., 2015), and suitable for preliminary site investigations (Oyeyemi et al., 2020; Egwuonwu et al., 2022; Irawan et al., 2022; Ibitoye, 2023). CPT, conversely, is a prevalent in situ testing methodology that entails the insertion of a cone penetrometer into the soil to measure resistance, pore pressure, and other geotechnical parameters, analyzing soil behavior and allowing for detailed profiling of the soil layers (Mayne, 2007; Robertson, 2009; Fortier and Wu, 2012; Robertson, 2016). Both techniques provide valuable insight into soil properties, enabling

effective decision-making in engineering and land management (Jung et al., 2008; Tumay et al., 2008; Ural, 2018; Daniyal et al., 2023)

Traditional methods in geotechnical data analysis fail to effectively handle the complexities of geotechnical data, such as sparsity, non-linearity, and uncertainty. These methods struggle with complex problems, inefficiency, and human error due to reliance on historical data, inefficient analytical techniques, and insufficient visualization and digitization processes (Meng et al., 2012; Zhang et al., 2016; Ji et al., 2022; Dungca and Galupino, 2023). The increasing complexity of soil characteristics calls for a more accurate and efficient data-driven approach to geotechnical engineering. Advanced methodologies, such as machine learning and integrated modeling, can assist in overcoming the aforementioned limitations, thereby enhancing the accuracy and efficiency of geotechnical investigations (Song et al., 2013; Xu et al., 2022; Liu et al., 2023). By employing sophisticated analytical methodologies and integrating a multitude of datasets, engineers can achieve a more profound comprehension of soil characteristics, leading to improved risk management and more effective engineering solutions (Robertson, 2010; Fitzgerald and Ritchie, 2019; Nikooee et al., 2020; Nurhasanah et al., 2024).

Applying machine learning to resistivity and CPT data offers significant advantages, including analyzing large datasets, uncovering complex relationships, and enhancing predictive accuracy. This approach can revolutionize traditional soil analysis methods, enabling more informed decision-making in geotechnical engineering and environmental management (Hengl et al., 2015; Angelopoulou et al., 2020; Rauter and Tschuchnigg, 2021; Fletcher, 2023; Radočaj et al., 2023). The integration of classical methods like CPT with new technologies such as machine learning and big data provides exciting insights for the future of geotechnical research and practice (Oberhollenzer, 2021; Chala and Ray, 2023). Algorithms including K-Nearest Neighbours (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) have shown promising results in improving soil classification accuracy, effectively capturing complex soil data patterns, and enhancing model performance (Zafar and Haq, 2020; Zhang et al., 2020; Taher et al., 2021; Huang et al., 2022; Aydın et al., 2023; Chala and Ray, 2023; Kamarudin et al., 2023; Gao, 2024; Weng and Jia, 2024; Yang et al., 2024). Moreover, integrating data has been shown to significantly enhance model performance, providing a more comprehensive understanding of soil behavior and enabling better-informed engineering decisions (Fortier and Wu, 2012; Wang et al., 2013; Reale et al., 2018; Zhu et al., 2024).

Despite growing interest in machine learning applications in geotechnics, few studies have focused on integrating resistivity and CPT data using machine learning algorithms. This research seeks to fill this gap by offering a novel, data-driven approach to compare and optimize these methods for soil classification. Specifically, the research explores how machine learning algorithms can improve soil classification accuracy, assess the impact of integrating resistivity and CPT data, and evaluate the performance of KNN, RF, and XGBoost in soil classification tasks. The results will contribute to advancing data-driven geotechnical models that integrate resistivity and CPT data for more accurate and efficient soil classifications, with practical implications for construction safety, risk mitigation, and sustainable engineering practices.

## 2. MATERIALS AND METHODS

### 2.1 Data Collection and Data Features

The data was collected from the investigated regions, comprehensively depicting the subsurface conditions. Geoelectric resistivity, CPT, and borehole measurements were performed at the identical site, ensuring the data were gathered under consistent conditions. The data were directly collected from the field and laboratory analysis of the sample.

The team collected data for this study at the Bontoramba Sub-district, District of Somba Opu, Gowa Regency, South Sulawesi, covering an area from the northeast to the southwest (Fig. 1). This site was selected based on its geological characteristics, as mapped in the Ujung Pandang sheet (Sukamto and Supriatna, 1982), which indicates diverse soil conditions ideal for this study.

The resistivity method involves passing an electric current (I) through two electrodes in the ground, creating an electric field that moves through soil layers. The voltage difference (V) is measured by two potential electrodes and resistance below the surface can be calculated using Ohm's law (Eq.1) (Reynold, 2011). Apparent resistivity ($\rho_a$) in the Eq. 2 represents the resistivity value of subsurface material at a specific depth with K as a geometric factor that depends on the electrode configuration (e.g., Wenner or Schlumberger). By changing electrode distance, resistivity data can be obtained from different depths, creating a model of soil resistivity distribution.

$$R = \frac{V}{I} \quad (1)$$

$$\rho_a = K \frac{V}{I} \quad (2)$$

The Cone Penetration Test (CPT) is a method used to measure soil resistance and sleeve friction. It involves pushing a cone into the soil at a constant speed of 2 cm/s, determining cone resistance ($q_c$) and sleeve friction ($f_s$) (Robertson and Cabal, 2014). Denser soils have higher $q_c$ values, while softer soils have lower values. The CPT also measures sleeve friction, distinguishing coarse-grained from fine-grained soils. Additional parameters like friction ratio and total friction are considered to provide a more detailed description of soil characteristics. The friction ratio ($R_f$) is obtained from the comparison between the $f_s$ and $q_c$ values and is calculated using Eq. 3. The total friction ($T_f$) is obtained by summing the fs values multiplied by reading

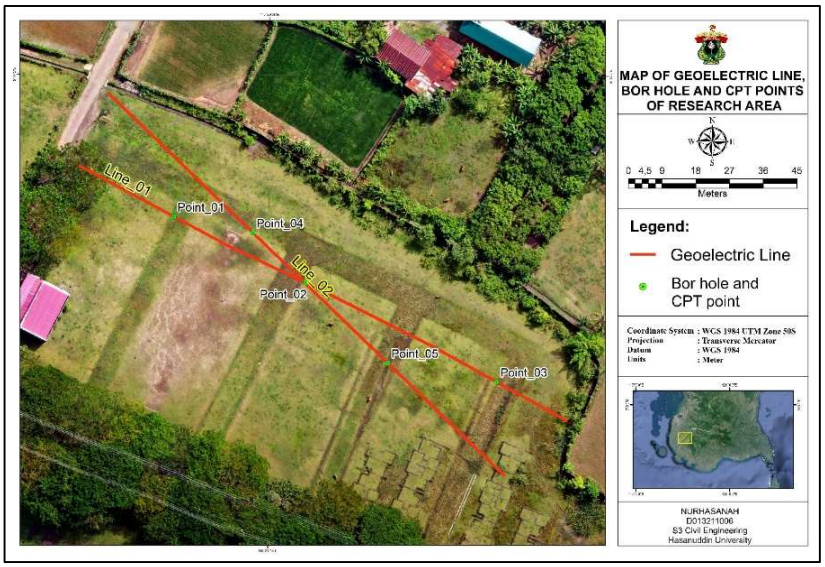


**Fig. 1**. Research site

interval and is calculated using Eq. 4 (SNI 2827, 2008).

$$R_f = \frac{f_s}{q_c} \times 100\% \quad (3)$$

$$T_f = f_s \times (\text{reading interval}) \quad (4)$$

The USCS classifies soils based on their physical properties, primarily grain size and plasticity. Soil samples from borehole measurement are analyzed in a laboratory to determine grain size distribution, with coarse-grained soils having over 50% larger particles than 0.075 mm. If refined grains are present, plasticity is tested by determining the liquid limits (LL) and plastic limits (PL) and calculating the plasticity index (PI). The results classify the soil based on the USCS diagram, with a two-letter code representing the soil type.

The features consist of parameters: resistivity value ($\rho$), cone resistance ($q_c$), sleeve friction ($f_s$), friction ratio ($R_f$), total friction ($T_f$), and soil type, recorded at every 20 cm of depth. The highest pressure achieved in the CPT measurement was 155 kg/cm$^2$ consequently, recorded depths at each point varied between 10.6 m and 12 m. The soil types obtained in this study consisted of Sandy Silt (ML), Poorly Graded Sand (SP), Sandy Lean Clay (CL), and Silty Sand (SM). The dataset collected for each feature consists of 272 data points. The dataset was randomly partitioned into two subsets, the first comprising 80% of the data for training purposes and the second comprising 20% for testing. This random division was carefully managed to maintain a balanced distribution between the two sets. This class imbalance poses challenges to the learning process and evaluation metrics.

Soil-type classification leverages both individual and integrated data features. In the case of individual data features, resistivity data is used as the input, with the corresponding soil type as the output. For the classification using CPT data, the input features include $q_c$, $f_s$, $R_f$, and $T_f$, with the soil type as the output. The integration of resistivity and CPT data combines the features $\rho$, $q_c$, $f_s$, $R_f$, and $T_f$ as inputs, with soil type data as the output, thereby utilizing integrated data features for enhanced classification.

## 2.2 Machine Learning Models

The machine learning algorithms employed for soil type classification are KNN, RF, and XGBoost. These algorithms are applied to individual and integrated data features to evaluate and compare their efficiency and accuracy in classifying soil types.

The KNN algorithm classifies a given sample according to the majority class of its nearest neighbors, identified using a distance metric (Euclidean distance). KNN model was employed with three neighbors (k) using the Euclidean distance metric and uniform weighting. RF represents an ensemble learning technique combining multiple decision trees, each trained on a randomly selected subset of the data and features. It uses majority voting for classification tasks. The RF algorithm was applied with 100 trees and excluded the splitting of subsets smaller than 5. XGBoost applies boosting techniques to improve weak learners sequentially. It minimizes a loss function using gradient descent and regularization to prevent overfitting. The RF algorithm was employed with 100 boosting stages.

Hyperparameters are important to improve model performance in classification tasks. GridSearchCV from scikit-learn is used for hyperparameter optimization. GridSearchCV tries all possible combinations to find the best combination based on accuracy.

## 2.3 Model Performance Evaluation

The classification model evaluation involves several steps and metrics to measure model performance, ensure accuracy, and detect potential overfitting or underfitting. The evaluation methods employed to measure classification

**Table 1.** Confusion matrix

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Positive | True Positive (TP) | False Negative (FN) |
| Actual Negative | False Positive (FP) | True Negative (TN) |

model's performance is confusion matrix, accuracy, precision, recall, and F1-score. This function helps understand how well the model predicts each class, especially when working with imbalanced datasets.

A confusion matrix is a tool employed for the evaluation of the classification model's performance, mainly to understand how well the model makes predictions for each class. It provides a visual representation of the true positives, true negatives, false positives, and false negatives, thus facilitating a more profound evaluation of the model's performance. The confusion matrix has four main components, as shown in Table 1.

The accuracy of a model is the percentage of correctly predicted outcomes out of the total number of predictions made. On the other hand, precision gauges the capacity of the model to predict the positive class accurately. Conversely, the recall metric is employed to evaluate the model's capacity to identify all positive instances from the total number of positive instances. The F1-score, which is the harmonic mean of precision and recall, gives a comprehensive evaluation by considering the balance between precision and recall. The aforementioned metrics can be calculated using Eqs. 5, 6, 7, and 8, with True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recal} \quad (7)$$

## 3. RESULTS AND DISCUSSION

### 3.1 Geological Conditions and Interpretation

The research location is in the Sombaopu Sub-district, as the Ujung Pandang geological map sheet indicates the area encompasses two primary geological formations: the Tpbv formation (Baturappe – Cindakko Volcanic Rock) and the Qac formation (Coastal Quarter Alluvium). The Qac formation comprises gravel, sand, clay, silt, and coral limestone, typically formed in river, beach, and deltaic environments. Alluvial deposits primarily stem from rock pieces sourced from the Lompobattang volcano. Conversely, the Tpbv formation is distinguished by lava, breccia, tuff, and conglomerate. The geological map sheet for Ujung Pandang is illustrated in Fig. 2 (Sukamto and Supriatna, 1982).

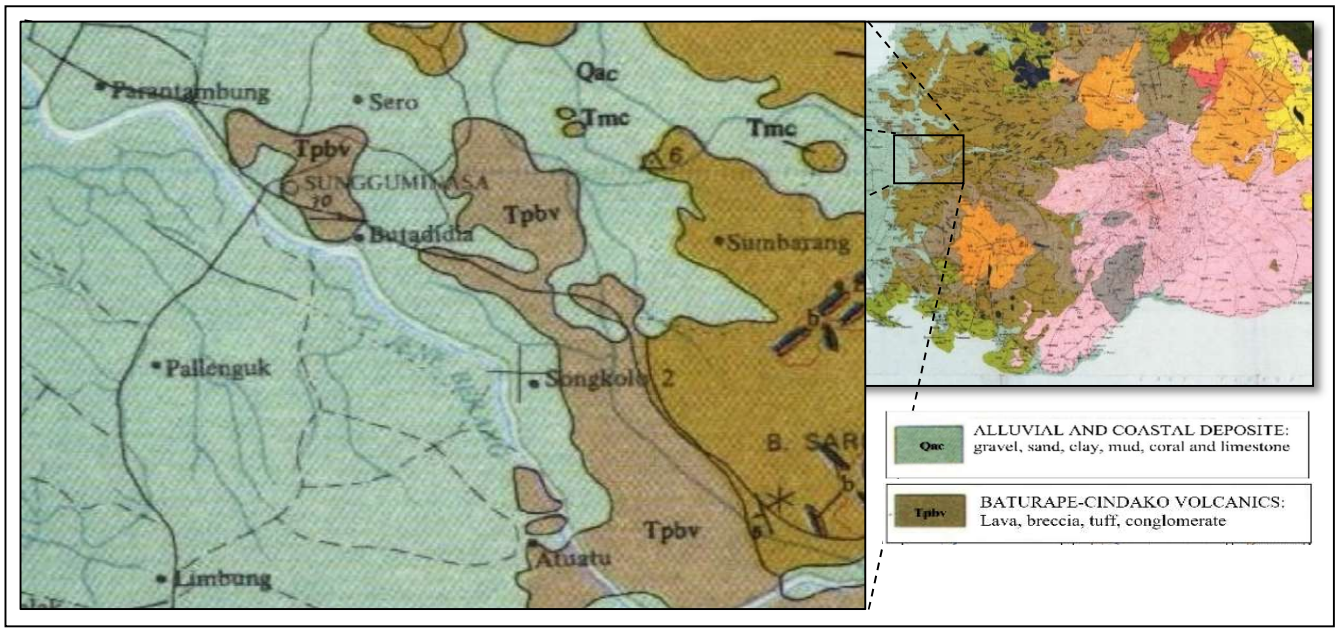


**Fig. 2.** Ujung Pandang geological map

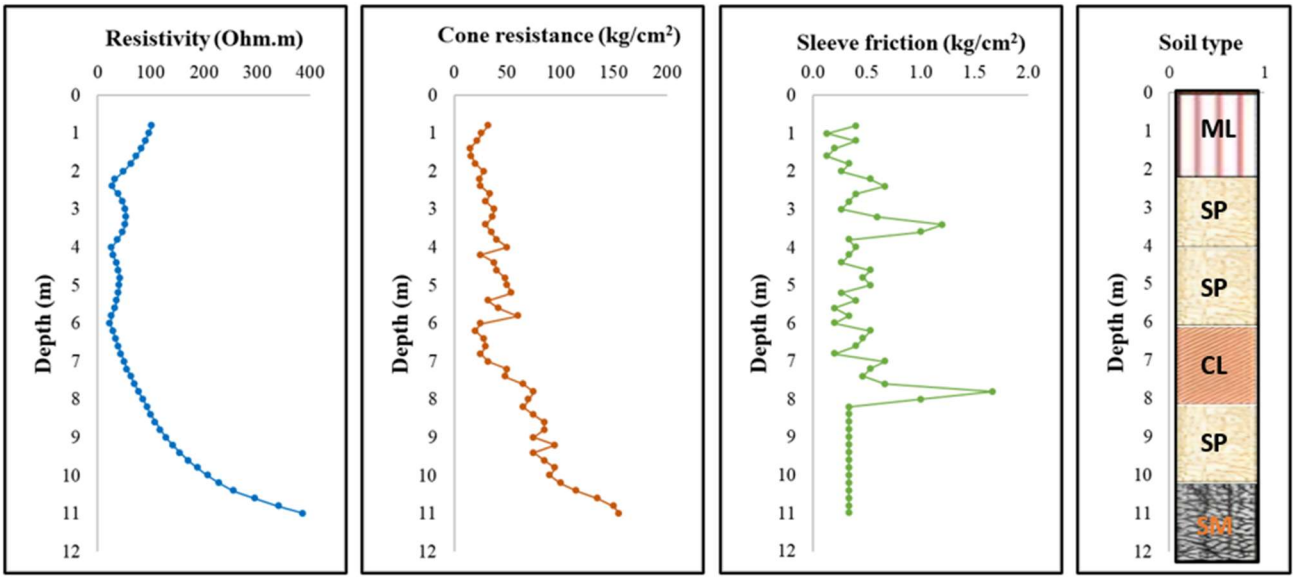


**Fig. 3.** The integration of resistivity, CPT, and soil type dataset at point 4

The integration of geoelectrical and geotechnical approaches enhances the understanding of subsurface conditions. A comparison of the resistivity value, cone resistance data, sleeve friction, and soil type in Fig. 3 indicates the emergence of data consistency at that depth. The resistivity and cone resistance values increase at 8–12 m depth. The USCS system categorizes various soil types displayed at different depths. These include Sandy Silt (ML) at 0–2 m, Poorly Graded Sand (SP) at 2–6 m and 9–10 m, Sandy Lean Clay (CL) at 7–8 m, and Silty Sand (SM) at 11–12 m.

A statistical summary of the dataset is presented in Table **2**, organized into 272 rows and five columns. As demonstrated in Fig**.** 4**.**, the frequency of soil type distribution in the dataset is illustrated. The distribution analysis showed that SP had the highest frequency and represented more than 50% of the dataset. The lowest frequency is the ML type, indicating an unbalanced dataset.

**Table 2.** Statistical summary of the datasets

| Parameter | Mean | Median | Min | Max | Range |
|---|---|---|---|---|---|
| ρ (Ω.m) | $2.0 \times 10^2$ | $8.6 \times 10^1$ | 6.0 | $1.3 \times 10^3$ | $1.3 \times 10^3$ |
| $q_c$ (Pa) | $4.9 \times 10^6$ | $3.5 \times 10^6$ | $7.9 \times 10^5$ | $1.5 \times 10^7$ | $1.5 \times 10^7$ |
| $f_s$ (Pa) | $4.6 \times 10^4$ | $3.3 \times 10^4$ | $6.6 \times 10^3$ | $5.3 \times 10^5$ | $5.3 \times 10^5$ |
| $R_f$ (%) | 1.2 | $8.0 \times 10^1$ | 0.149 | $1.2 \times 10^1$ | $1.2 \times 10^1$ |
| $T_f$ (N/m) | $2.6 \times 10^5$ | $2.4 \times 10^5$ | $8.0 \times 10^3$ | $7.7 \times 10^5$ | $7.6 \times 10^5$ |

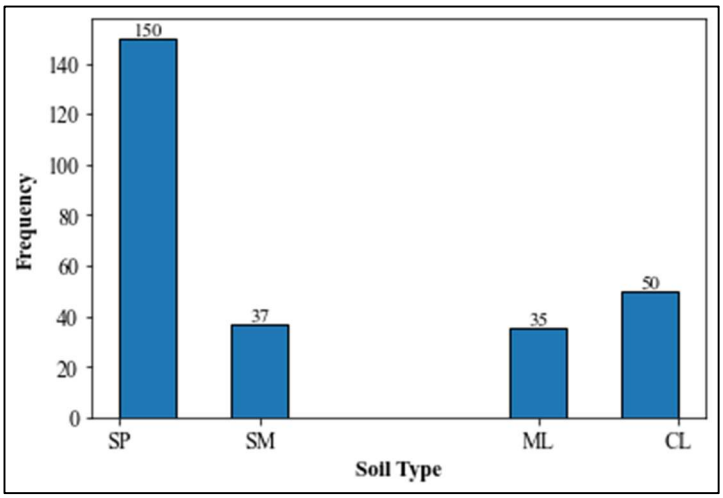


**Fig. 4.** Distribution of soil types considered for machine learning models

### 3.2 Performance on Individual Data Features

#### 3.2.1 Resistivity Data

The performance of the XGBoost, RF, and KNN algorithms in soil type classification was evaluated using resistivity data. The findings of the machine learning models are exhibited and examined through a confusion matrix and an assortment of performance metrics, including accuracy, precision, recall, and the F1-score. The aforementioned metrics can be observed in Fig. 5 and Table 3. As demonstrated by the confusion matrix in Fig. 5, the KNN algorithm exhibits a higher misclassification rate than both Random Forest and XGBoost, suggesting that its performance may be less robust for the given dataset. In contrast, the Random Forest model demonstrates relatively strong classification accuracy with moderate misclassification rates, outperforming KNN. XGBoost, however, achieves the fewest misclassifications, exhibiting superior classification performance and accuracy.

The accuracy values for all three algorithms are relatively low, suggesting limited performance on the resistivity dataset (Table 3). KNN achieved the highest accuracy at 0.62, outperforming both XGBoost and RF. All models exhibit low precision, recall, and F1 scores, indicating they struggle with several classes. The weighted average, which accounts for class support, scores higher than the macro averages, suggesting that the models perform better on larger classes but encounter difficulties with smaller ones.

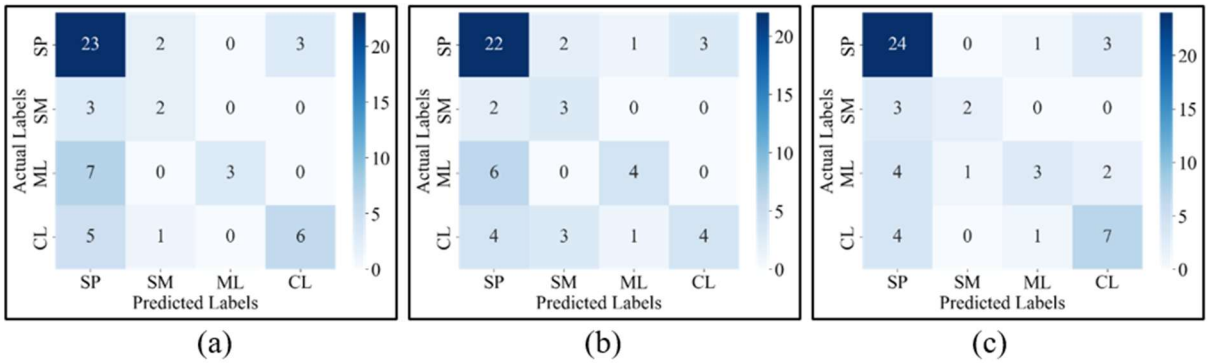


**Fig. 5.** Confusion matrix of machine learning model based on resistivity data: (a) KNN; (b) RF; (c) XGBoost

**Table 3.** The classification report of the KNN, RF, and XGBoost algorithms was assessed using resistivity data

| Classification report of KNN, RF, and XGBoost Algorithm | | | | | |
|---|---|---|---|---|---|
| Model | | Precision | Recall | F1-score | Support |
| KNN | SP | 0.61 | 0.82 | 0.70 | 28 |
| | SM | 0.40 | 0.40 | 0.40 | 5 |
| | ML | 1.00 | 0.30 | 0.46 | 10 |
| | CL | 0.67 | 0.50 | 0.57 | 12 |
| | Accuracy | - | - | 0.62 | 55 |
| | Macro avg | 0.67 | 0.51 | 0.53 | 55 |
| | Weighted avg | 0.67 | 0.62 | 0.60 | 55 |
| RF | SP | 0.65 | 0.79 | 0.71 | 28 |
| | SM | 0.38 | 0.60 | 0.46 | 5 |
| | ML | 0.67 | 0.40 | 0.50 | 10 |
| | CL | 0.57 | 0.33 | 0.42 | 12 |
| | Accuracy | - | - | 0.60 | 55 |
| | Macro avg | 0.57 | 0.53 | 0.52 | 55 |
| | Weighted avg | 0.61 | 0.60 | 0.59 | 55 |
| XGBoost | ML | 0.60 | 0.30 | 0.40 | 10 |
| | SP | 0.69 | 0.86 | 0.76 | 28 |
| | CL | 0.58 | 0.58 | 0.58 | 12 |
| | SM | 0.67 | 0.40 | 0.50 | 5 |
| | Accuracy | - | - | 0.60 | 55 |
| | Macro avg | 0.63 | 0.54 | 0.56 | 55 |
| | Weighted avg | 0.65 | 0.65 | 0.63 | 55 |

### 3.2.2 Cone Penetration Test Data

The performance of the XGBoost, RF, and KNN algorithms in soil type classification was evaluated by analyzing CPT data. The predictive capabilities of the models are generally satisfactory, with Class SP and Class SM demonstrating consistent accuracy across all three models (Fig. 6). The KNN algorithm has the highest error rate among the models, struggling particularly with distinguishing between the SP and SM classes. Random Forest outperforms KNN, with fewer misclassifications and strong performance overall, although it faces some difficulty between the SM and ML classes. XGBoost leads with the fewest errors and the most reliable classification results across all classes, providing the most robust performance for this dataset.

As illustrated in Table 4, all algorithms enhanced performance when evaluated on CPT data, with RF attaining the highest level of accuracy. XGBoost exhibited superior performance in more extensive classes, while KNN exhibited the lowest accuracy. The minimal discrepancy between macro and weighted averages suggests that Random Forest effectively addresses class imbalances across smaller and larger classes. XGBoost exhibited superior performance in classes with more support, while KNN continued to demonstrate low precision and recall, as evidenced by its suboptimal macro averages. The weighted averages for KNN are higher, indicating better performance on larger classes

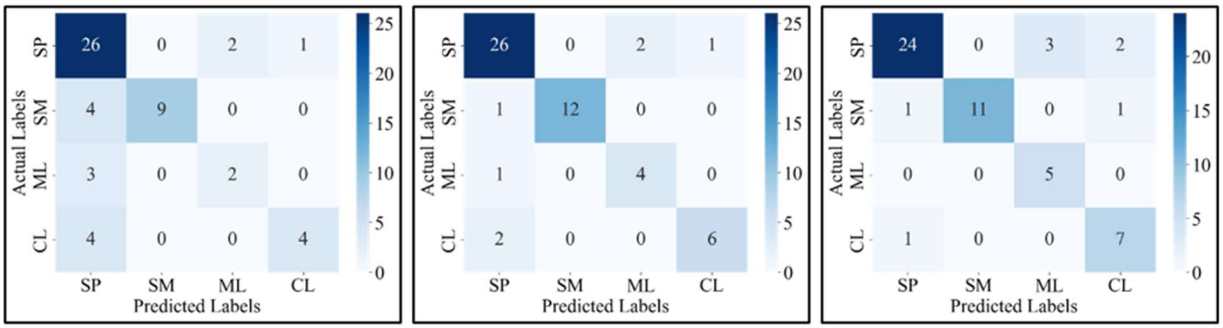


**Fig**. 6. Confusion matrix of machine learning model based on CPT data: (a) KNN; (b) RF; (c) XGBoost

**Table 4.** The classification report of the KNN, RF, and XGBoost algorithms was assessed using CPT data

| Classification report of KNN, RF, and XGBoost Algorithm | | | | | |
|---|---|---|---|---|---|
| Model | | Precision | Recall | F1-score | Support |
| KNN | SP | 0.70 | 0.90 | 0.79 | 29 |
| | SM | 1.00 | 0.69 | 0.82 | 13 |
| | ML | 0.50 | 0.40 | 0.44 | 5 |
| | CL | 0.80 | 0.50 | 0.62 | 8 |
| | Accuracy | - | - | 0.75 | 55 |
| | Macro avg | 0.75 | 0.62 | 0.67 | 55 |
| | Weighted avg | 0.77 | 0.75 | 0.74 | 55 |
| RF | SP | 0.87 | 0.90 | 0.88 | 29 |
| | SM | 1.00 | 0.92 | 0.96 | 13 |
| | ML | 0.67 | 0.80 | 0.73 | 5 |
| | CL | 0.86 | 0.75 | 0.80 | 8 |
| | Accuracy | - | - | 0.87 | 55 |
| | Macro avg | 0.85 | 0.84 | 0.84 | 55 |
| | Weighted avg | 0.88 | 0.87 | 0.87 | 55 |
| XGBoost | ML | 0.62 | 1.00 | 0.77 | 5 |
| | SP | 0.92 | 0.83 | 0.87 | 29 |
| | CL | 0.70 | 0.88 | 0.78 | 8 |
| | SM | 1.00 | 0.85 | 0.92 | 13 |
| | Accuracy | - | - | 0.85 | 55 |
| | Macro avg | 0.81 | 0.89 | 0.83 | 55 |
| | Weighted avg | 0.88 | 0.85 | 0.86 | 55 |

### 3.3 Performance of Integrated Data

The present study utilized the integration of resistivity and CPT data to assess how combining these datasets enhances the algorithms' predictive performance, as demonstrated in Fig. 7. KNN has the highest misclassification rate, requiring further optimization, especially for distinguishing similar classes. Random Forest performs well with high accuracy and few misclassifications, outperforming other models. XGBoost, similar to Random Forest in accuracy, excels in differentiating classes, yielding minimal errors and robust performance. The results are summarised in Table 5. The random forest model outperforms the other algorithms in predicting soil types based on resistivity and CPT data, achieving an accuracy of 93%. In comparison, the XGBoost model achieves an accuracy of 91%, while the KNN model reaches 75% and exhibits many misclassifications. RF also leads in both macro averages and weighted averages, demonstrating robust accuracy and consistent performance across all classes. This makes it particularly effective for datasets with imbalanced class distributions.

### 3.4 A Comparison of Machine Learning Models' Performance

The comparative analysis of algorithms across all datasets reveals clear distinctions in their performance for soil type classification. The comparison highlights the performance differences among the three algorithms (XGBoost, Random Forest, and KNN) and the impact of utilizing different data types. Fig. 8 illustrates each algorithm's accuracy, and Figs. 9, 10, and 11 present their precision, recall, and F1 score, respectively.

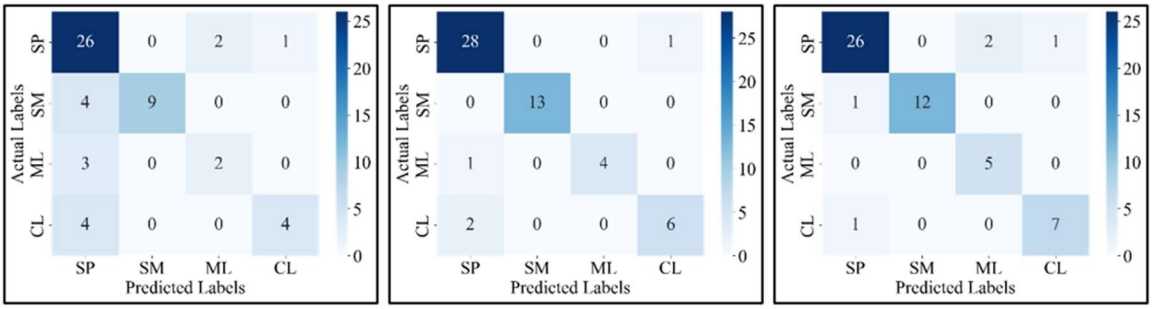


**Fig. 7.** Confusion matrix of machine learning model based on resistivity and CPT data: (a) KNN; (b) RF; (c) XGBoost

**Table 5.** Classification report of the KNN, RF, and XGBoost algorithms was assessed using resistivity data and CPT data

| Classification report of KNN, RF, and XGBoost Algorithm | | | | | |
|---|---|---|---|---|---|
| Model | | Precision | Recall | F1-score | Support |
| KNN | SP | 0.70 | 0.90 | 0.79 | 29 |
| | SM | 1.00 | 0.69 | 0.82 | 13 |
| | ML | 0.50 | 0.40 | 0.44 | 5 |
| | CL | 0.80 | 0.50 | 0.62 | 8 |
| | Accuracy | - | - | 0.75 | 55 |
| | Macro avg | 0.75 | 0.62 | 0.67 | 55 |
| | Weighted avg | 0.77 | 0.75 | 0.74 | 55 |
| RF | SP | 0.90 | 0.97 | 0.93 | 29 |
| | SM | 1.00 | 1.00 | 1.00 | 13 |
| | ML | 1.00 | 0.80 | 0.89 | 5 |
| | CL | 0.86 | 0.75 | 0.80 | 8 |
| | Accuracy | - | - | 0.93 | 55 |
| | Macro avg | 0.94 | 0.88 | 0.91 | 55 |
| | Weighted avg | 0.93 | 0.93 | 0.93 | 55 |
| XGBoost | ML | 0.71 | 1.00 | 0.83 | 5 |
| | SP | 0.93 | 0.90 | 0.91 | 29 |
| | CL | 0.88 | 0.88 | 0.88 | 8 |
| | SM | 1.00 | 0.92 | 0.96 | 13 |
| | Accuracy | - | - | 0.91 | 55 |
| | Macro avg | 0.88 | 0.92 | 0.90 | 55 |
| | Weighted avg | 0.92 | 0.91 | 0.91 | 55 |

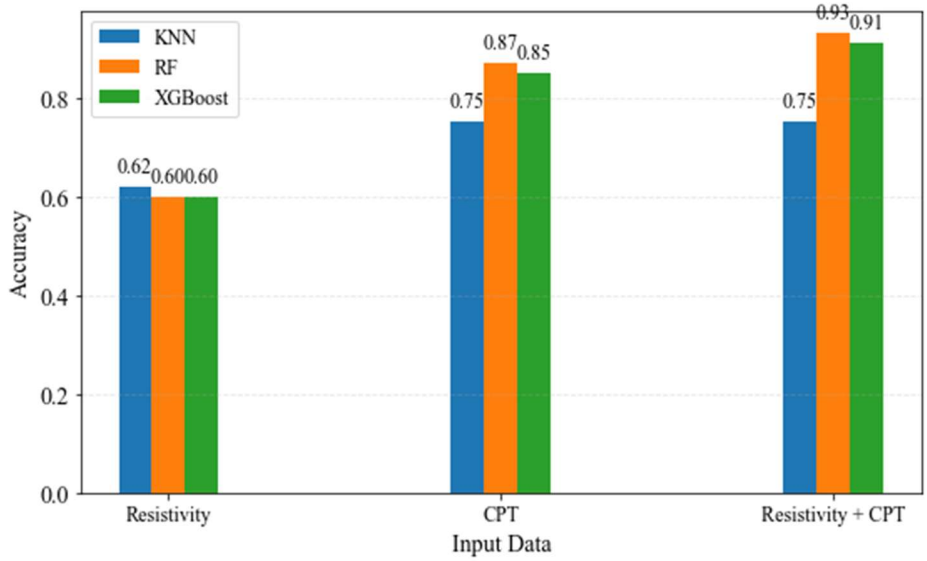


**Fig. 8.** The graph of accuracy machine learning model based on resistivity and CPT data

As shown in Fig. 8, RF model consistently outperformed the other algorithms across all three datasets, achieving the highest accuracy in both Set 2 (CPT data) and Set 3 (combined resistivity and CPT data). The XGBoost model followed closely, demonstrating slightly varying accuracy values (green). In contrast, the KNN model achieved the highest accuracy with the resistivity dataset but performed poorly with both the CPT and combined datasets (blue).

The analysis revealed that all three algorithms exhibited relatively low accuracy when applied to the resistivity dataset alone, indicating limited predictive capability. However, performance improved when using the CPT dataset, and the integration of both resistivity and CPT data further enhanced the performance of the random forest and XGBoost models. In contrast, KNN showed no substantial improvement with the combined dataset, suggesting that including resistivity data did not notably benefit its predictive performance.

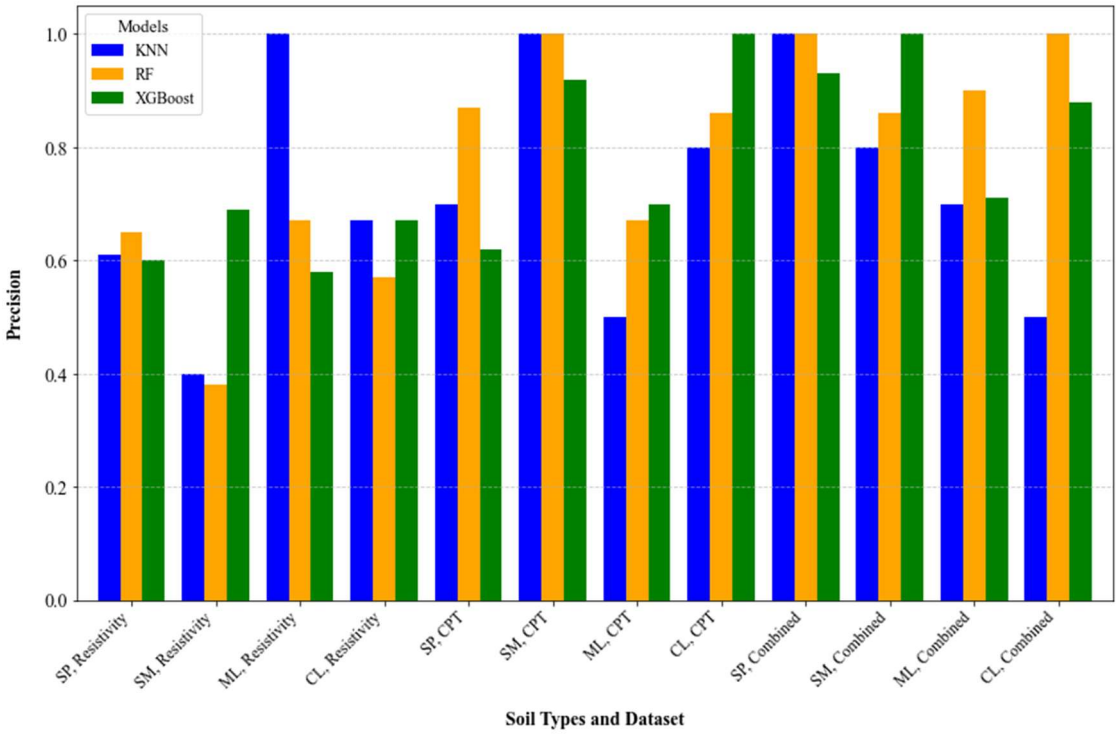


**Fig. 9.** The graph of precision machine learning model based on resistivity and CPT data

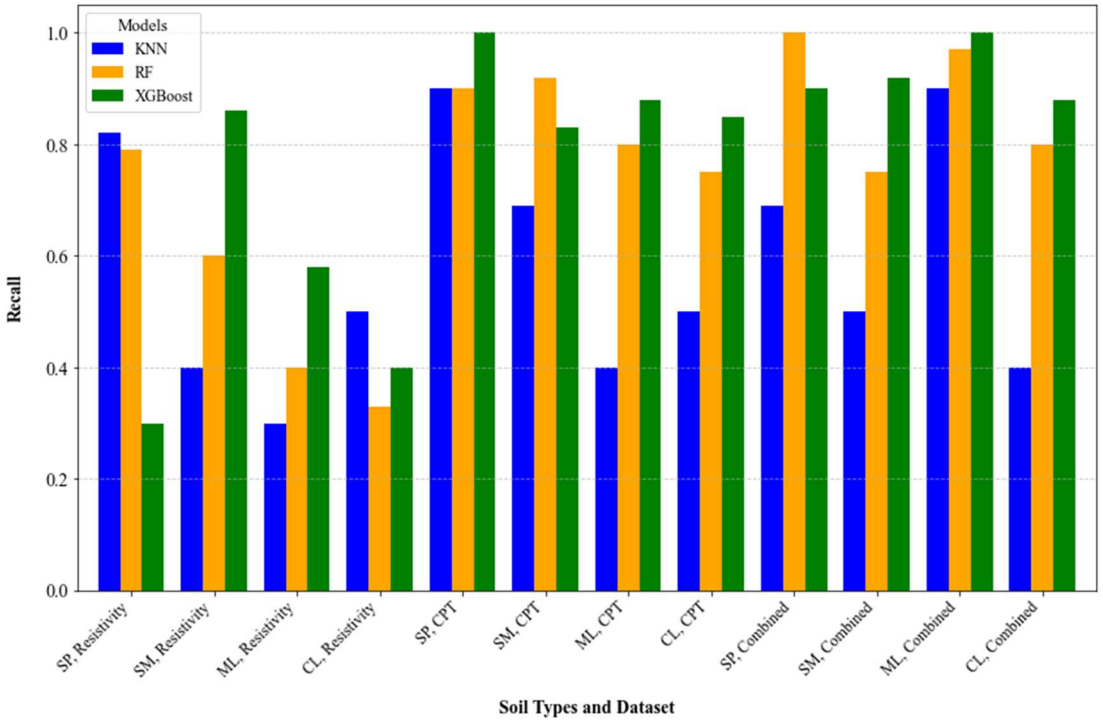


**Fig. 10.** The graph of recall machine learning model based on resistivity and CPT data

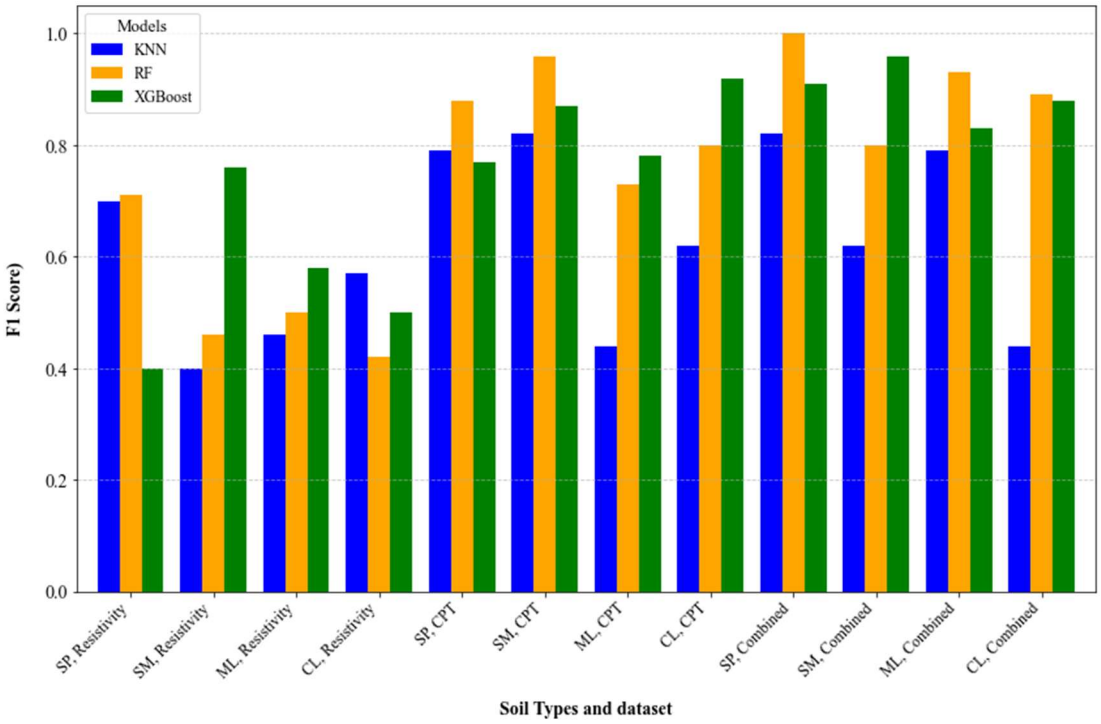


**Fig. 11.** The graph of F1-score machine learning model based on resistivity and CPT data

As illustrated in Fig. 9, the RF algorithm exhibits superior precision stability across most classes. The XGBoost algorithm shows a slight precision discrepancy, while the KNN algorithm demonstrates the lowest precision. These findings suggest that both RF and XGBoost are proficient in predicting positive classes and reducing false positives.

Fig. 10 highlights that XGBoost achieves high recall across most classes, while RF demonstrates superior recall for more prominent classes (SP) across various datasets. In contrast, KNN shows the lowest recall in nearly all classes.

This suggests that XGBoost excels in class recognition, RF is particularly effective at identifying larger classes, and KNN encounters difficulties in class recognition.

As shown in Fig. 11, Random Forest achieves the highest F1-score for most of its primary classes in both the CPT and combined datasets, with XGBoost following closely, albeit with a slight difference. In contrast, KNN exhibits the weakest performance. These results indicate that RF and XGBoost maintain a more balanced performance, striking a favorable equilibrium between prediction accuracy and class detection. Conversely, KNN demonstrates limited proficiency in class prediction and detection, reflecting challenges in capturing the complexity inherent in the CPT and combined datasets.

This analysis shows that using various data types, the RF and XGBoost algorithms outperform KNN in soil type classification. Random Forest consistently delivers the best results in accuracy, precision, recall, and F1-score, followed by XGBoost, which provides highly competitive results. While KNN performs reasonably well with the resistivity dataset, it performs poorly than other datasets and struggles with handling more complex data. Therefore, for datasets involving combined resistivity and CPT data, Random Forest and XGBoost are recommended as the preferred algorithms for soil type classification.

The analysis based on the impact of utilizing different data types revealed that all three algorithms had relatively low accuracy, precision, recall, and F1-scores when applied to the resistivity dataset alone, indicating limited predictive capability. However, performance improved with the CPT dataset, and integrating resistivity and CPT data further enhanced the performance of XGBoost and Random Forest. KNN showed no significant improvement with integrated data. The combination of resistivity and CPT datasets, which provide complementary information from both datasets, such as mechanical soil properties and resistivity measurements, markedly improved predictive performance and classification precision, demonstrating the advantages of leveraging multiple data sources. The integration of CPT with resistivity data allows for a more nuanced understanding of sediment types, as shown by Goebel and Knight, who utilized co-located CPT and electrical resistivity measurements to classify sediment types into coarse and fine-grain-dominated the materials (Goebel and Knight, 2021). This resistivity-to-sediment-type transformation captures the inherent uncertainties associated with variable water salinity and content, thereby enhancing the reliability of subsurface assessments. This approach aligns with the emphasized need for a comprehensive understanding of soil liquefaction potential by integrating both geotechnical and geophysical data, which allows for a more accurate evaluation of spatial distributions and induced surface settlements (Yang et al., 2023). Moreover, Duffy et al. utilized CPT parameters in conjunction with gradient boosting methods to refine the assessment of soil compressibility, underscoring the importance of integrating diverse datasets for improved predictive modeling (Duffy et al., 2020).

Although the integration of resistivity and CPT data substantially improved classification accuracy across all algorithms, class imbalance and the limited size of the dataset may have impacted the predictions and limited generalizability. Larger datasets encompassing a broader range of soil conditions would allow for a more robust evaluation. The poor performance of KNN suggests it is susceptible to high dimensionality and noise within the integrated dataset, posing a potential challenge for practical applications.

## 4. CONCLUSION

This research highlights that RF and XGBoost outperform KNN in soil type classification, with RF achieving the best accuracy, precision, recall, and F1-score results. While KNN performs well with the resistivity dataset, it struggles with other and more complex data. Integrating resistivity and CPT data enhances the performance of RF and XGBoost, while KNN shows no significant improvement. These findings emphasize enhancing analytical techniques' predictive accuracy and efficiency by utilizing the RF and XGBoost algorithms and integrating diverse data types, such as resistivity and CPT. This has practical implications for geotechnical applications, such as construction planning, foundation design, and environmental risk mitigation.

## DECLARATION OF COMPETING INTEREST

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

## ACKNOWLEDGMENT

The author gratefully acknowledges the financial support from the Direktorat Jenderal Pendidikan Tinggi Riset dan Teknologi through the Penelitian Pascasarjana (PPS) 2024. Special thanks are also due to the Soil Mechanics Laboratory, Department of Civil Engineering, Hasanuddin University for providing the necessary facilities.

## REFERENCES

Angelopoulou, T., Balafoutis, A., Zalidis, G., Bochtis, D. 2020. From laboratory to proximal sensing spectroscopy for soil organic carbon estimation – A review. Sustainability, 12, 443.

Aydın, Y., Işıkdağ, Ü., Bekdaş, G., Nigdeli, S.M., Geem, Z.W. 2023. Use of machine learning techniques in soil classification. Sustainability, 15, 2374.

Baker, H., Gabr, A., Djeddi, M. 2015. Geophysical and geotechnical techniques: complementary tools in studying subsurface features. International Conference on Engineering Geophysics, 15–18.

Chala, A.T., Ray, R.P. 2023. Assessing the performance of machine learning algorithms for soil classification using cone penetration test data. Applied Sciences, 13, 5758.

Chala, A.T., Ray, R.P. 2023. Machine learning techniques for soil characterization using cone penetration test data. Applied Sciences, 13, 8286.

Daniyal, M., Sohail, G.M. Rashid, H.M.A. 2023. GIS-based mapping of geotechnical and geophysical properties of Lahore soils. Environmental Earth Science, 82, 540.

Duffy, K., Siderius, K., Long, M. 2020. Parameterisation of the Koppejan settlement prediction model using cone penetration testing and gradient boosting. Proceedings of the International Association of Hydrological Sciences, 382, 443–447.

Dungca, J., Galupino, J. 2023. Developing nomographs for the unit weight of soils. Buildings, 13, 2315.

Egwuonwu, G.N., Okwonna, I.A., Okpala, P.K. 2022. Geoelectrical and geotechnical investigations for development of superstructures at Nkpologwu proposed judiciary site, Anambra Basin, Southeastern Nigeria. Physical Science International Journal, 26, 47–58.

Fitzgerald, D.M., Ritchie, A.I.M. 2019. Geophysical methods for subsurface characterization: A review. Geophysical Journal International, 218, 1543–1561.

Fletcher, R. 2023. Machine learning mapping of soil apparent electrical conductivity on a research farm in Mississippi. Agricultural Sciences, 14, 915–924.

Fortier, R.M., Wu, W. 2012. Penetration rate-controlled electrical resistivity and temperature piezocone penetration tests in warm ice-rich permafrost in Northern Quebec, Canada. In Proceedings of the 2012 ASCE Geo-Engineering Conference, 1–10.

Gao, W. 2024. The application of machine learning in geotechnical engineering. Applied Sciences, 14, 4712.

Goebel, M., Knight, R. 2021. Recharge site assessment through the integration of surface geophysics and cone penetrometer testing. Vadose Zone Journal, 20, 1–18.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Jesus, J.M.D., Tamene, L., Tondoh, J.E. 2015. Mapping soil properties of Africa at 250 m resolution: Random Forests significantly improve current predictions. PLoS ONE, 10, e0125814.

Huang, L., Liu, Y., Huang, W., Dong, Y., Ma, H., Wu, K., Guo, A. 2022. Combining Random Forest and XGBoost methods in detecting early and mid-term winter wheat stripe rust using canopy level hyperspectral measurements. Agriculture, 12, 74.

Ibitoye, F.P. 2023. Geophysical investigations for design parameters related to geotechnical engineering. Avantgarde reliability implications in civil engineering. IntechOpen, 1–17.

Irawan, L., Pradana, I., Panoto, D., Darmansyah, A. 2022. Identification of subsurface materials in landslide-susceptible areas in the Pacet-Trawas road corridor using the geoelectrical resistivity method. 3rd International Conference on Geography and Education (ICGE). KNE Social Sciences, 219–230

Ji, X., Lu, X., Guo, C., Pei, W., Xu, H. 2022. Predictions of geological interface using relevant vector machine with borehole data. Sustainability, 14, 10122.

Jung, B.C., Gardoni, P., Biscontin, A., 2008. Probabilistic soil identification based on cone penetration tests. Géotechnique, 58, 591–603.

Kamarudin, F., Budiman, F., Winarno, S., Kurniawan, D. 2023. Optimizing classification algorithms using soft voting: a case study on soil fertility dataset. Journal Teknologi Informasi Dan Pendidikan, 16, 255–268.

Liu, H., Chen, Y., Zhao, L., Liu, W. 2023. Research on geotechnical data interpolation and prediction techniques. In 2023, the 4th International Conference on Management Science and Engineering Management (ICMSEM 2023). Atlantis Press, 1788–1795.

Mayne, P.W. 2007. Cone penetration testing: A review of the state of the art. Geotechnical Testing Journal, 30, 1–16.

Meng, F.Z., Li, S.J., Zhang, Z.H. 2012. Determination of mechanical parameters of reservoir landslide based on back analysis using evolutionary artificial network. Applied Mechanics and Materials, 170, 729–734.

Nikooee, E., Mirghafari, R., Habibagahi, G., Khorassani, A.G., Nouri, A.M. 2020. Determination of soil-water retention curve: an artificial intelligence-based approach. E3s Web of Conferences, 195, 02010.

Nurhasanah, Muhiddin, A.B., Djamaluddin, A.R., Niswar, M. 2024. determining the depth of hard soil layers using geoelectric resistivity and cone penetration test methods (case study: Kelurahan Bontoramba Kecamatan Somba opu kabupaten gowa). Journal Ilmu Pendidikan Fisika, 9, 142–151.

Oberhollenzer, S., Premstaller, M., Marte, R., Tschuchnigg, F., GeorgH. Erharter, G., Marcher, T. 2021. Cone penetration test dataset Premstaller Geotechnik. Data in Brief, 34, 106618.

Oyeyemi, K.D., Olofinnade, O.M., Aizebeokhai, A.P., Sanuade, O.A., Oladunjoye, M.A., Ede, A.N., Adagunodo, T.A., Ayara, W.A. 2020. Geoengineering site characterization for foundation integrity assessment. Cogent Engineering, 7, 1711684.

Radočaj, D., Jurišić, M., Tadić, V. 2023. The effect of bioclimatic covariates on ensemble machine learning prediction of total soil carbon in the Pannonian Biogeoregion. Agronomy, 13, 2516.

Rauter, S., Tschuchnigg, F. 2021. CPT data interpretation employing different machine learning techniques. Geosciences, 11, 265.

Reale, C., Gavin, K., Librić, L., Kaćunić, D.J. 2018. Automatic classification of fine-grained soils using CPT measurements and artificial neural networks. Advanced Engineering Informatics, 36, 207–215.

Robertson, P.K. 2009. Interpretation of cone penetration test–a unified approach. Canadian Geotech, Journal the NRC Research Press, 46, 1337–1355.

Robertson, P.K. 2016. Cone penetration testing in geotechnical engineering. Geotechnical Engineering Handbook. Koerner, 2nd ed. New York: McGraw-Hill.

Robertson, P.K. 2010. Interpretation of cone penetration test data. Geotechnical Engineering Handbook, 2nd ed. New York: McGraw-Hill.

Shao, W., Yue, W., Zhang, Y., Zhou, T., Zhang, Y., Dang, Y., Wang, H., Feng, X., Chao, Z. 2023. The application of machine learning techniques in geotechnical engineering: A review and comparison. Mathematics, 11, 3976.

Song, Q.H., Li, X. L., Li, Y.Y., Wu, Y. 2013. The practical probability analysis methods for stability of landslide. Applied Mechanics and Materials, 423, 1308–1311.

Standar Nasional Indonesia. 2008. Cara uji penetrasi lapangan dengan alat sondir. Badan Standardisasi Nasional, SNI 2827.

Sukamto, R., Supriatna, S. 1982. Peta Geologi Bersistem Lembar Ujungpandang. Benteng, dan Sinjai Sulawesi, Pusat Penelitian Dan Pengembangan Geologi, Bandung.

Taher, K.I., Abdulazeez, A.M., Zebari, D.A. 2021. Data mining classification algorithms for analyzing soil data. Asian Journal of Research in Computer Science, 8, 17–28.

Tumay, M.T., Farsakh, M.A., Zhang, Z. 2008. From theory to implementation of a CPT-Based probabilistic and fuzzy soil classification. From Research to Practice in Geotechnical Engineering, 259–276.

Ural, N. 2018. The Importance of Clay in Geotechnical Engineering. InTech.

Wang, Y., Huang, K., Cao, Z. 2013. Probabilistic identification of underground soil stratification using cone penetration tests. Canadian Geotechnical Journal, 50, 766–776.

Weng, K., Jia, M. 2024. Predicting base resistance of super-long piles using a random forest model: a case study from ho chi minh city. IOP Conference Series: Earth and Environmental Science, 1337, 012035.

Xu, J., Wang, Y., Zhang, L. 2022. Fusion of geotechnical and geophysical data for 2d subsurface site characterization using multi-source Bayesian compressive sampling. Canadian Geotechnical Journal. 59, 1756–1773.

Yang, H., Liu, Z., Yan, Y., Li, Y., Tao, G. 2023. Adaptive fusion sampling strategy combining geotechnical and geophysical data for evaluating two-dimensional soil liquefaction potential and reconsolidation settlement. Applied Sciences, 13, 5931.

Yang, Y., Song, X., Zhang, S., Hu, J., Ruan, M., Zeng, D., Luo, H., Wang, J., Wang, Z. 2024. Correlation analysis and prediction of the physical and mechanical properties of coastal soft soil in the Jiangdong New District, Haikou, China. Advances in Civil Engineering, 2024, 9985210.

Zafar, N., Haq, I.U. 2020. Traffic congestion prediction based on estimated time of arrival. Plos One, 15, 0238200.

Zhang, J., Wu, C., Wang, L., Mao, X., Wu, Y. 2016. The work flow and operational model for geotechnical investigation based on BIM. in IEEE Access, 4, 7500–7508, 2606158.

Zhang, M., Shi, W., Xu, Z. 2020. Systematic comparison of five machine-learning models in classification and interpolation of soil particle size fractions using different transformed data. Hydrology and. Earth System Sciences 24, 2505–2526.

Zhu, F., Zhu, C., Lu, W., Fang, Z., Li, Z., Pan, J. 2024. Soil classification mapping using a combination of semi-supervised classification and stacking learning (SSC-SL). Remote Sensing, 16, 405.