Geomechanics and Engineering, *Vol. 35, No. 1 (2023) 67-80*
https://doi.org/10.12989/gae.2023.35.1.067

A study on data mining techniques for soil classification methods using cone penetration test results

Junghee Park[1], So-Hyun Cho[2], Jong-Sub Lee[3] and Hyun-Ki Kim *[2]

[1]*Department of Civil and Environmental Engineering, Incheon National University, Incheon 22012, Republic of Korea*
[2]*Department of Civil and Environmental Engineering, Kookmin University, Seoul 02707, Republic of Korea*
[3]*School of Civil, Environmental and Architectural Engineering, Korea University, Seoul 02841, Republic of Korea*

(*Received December 20, 2022, Revised August 7, 2023, Accepted August 26, 2023*)

Abstract. Due to the nature of the conjunctive Cone Penetration Test(CPT), which does not verify the actual sample directly, geotechnical engineers commonly classify the underground geomaterials using CPT results with the classification diagrams proposed by various researchers. However, such classification diagrams may fail to reflect local geotechnical characteristics, potentially resulting in misclassification that does not align with the actual stratification in regions with strong local features. To address this, this paper presents an objective method for more accurate local CPT soil classification criteria, which utilizes C4.5 decision tree models trained with the CPT results from the clay-dominant southern coast of Korea and the sand-dominant region in South Carolina, USA. The results and analyses demonstrate that the C4.5 algorithm, in conjunction with oversampling, outlier removal, and pruning methods, can enhance and optimize the decision tree-based CPT soil classification model.

Keywords: cone penetration test; data mining; decision tree model; machine learning; soil classification; stratification

1. Introduction

The cone penetration test (CPT) is one of the typical geotechnical field tests conducted on soft ground. Although this test method is useful for understanding various geotechnical engineering characteristics by measuring the tip resistance of the cone, the frictional resistance force on the circumferential surface of the sleeve, and the excessive pore water pressure of the surrounding ground during cone penetration, it is impossible to collect or observe soil samples directly while penetrating the cone, which becomes always a limitation for soil classification based on this test results. However, the development of the friction mantle cone and previous investigations have proposed various soil classification charts with CPT results (Begemann 1965, Douglas and Olsen 1981, Robertson and Campanella 1983, Robertson *et al*. 1986, Robertson 1990, Robertson and Wride 1998, Robertson 2009, Robertson and Cabal 2014, Robertson 2016).

In particular, the series of classification methods proposed by Robertson are the most widely used worldwide. The Robertson's CPT classification charts reflect the overall engineering characteristics of each soil and can classify soils very reasonably (Kim *et al*. 2008a, Kim *et al*. 2008b). However, there still exists inherent shortcomings in the CPT-based soil classification charts where they cannot consider the unique features of the local geomaterials. Many other soil classification methods have limitations associated with the bias of the data considered while developing each method (Park *et al*. 2022).

With the recent development of data acquisition and processing methods, data mining techniques have been proposed as an alternative to such traditional soil classification methods: decision trees, support vector machines, clustering, artificial neural networks, and so on (Odeh *et al*. 1992, Cal 1995, Rizzo *et al*. 1996, Najjar and Basheer 1996, Juang *et al*. 2001, Bhattacharya and Solomatine 2006, Das and Basudhar 2009, Bhargavi and Jyothi 2011, Cao and Wang 2013, Cai *et al*. 2018, Cao *et al*. 2018).

In this study, we propose an optimized CPT-based soil classification model trained with the regional CPT results at the target ground using a decision tree method, C4.5 algorithm (Quinlan 1986, Quinlan 2008). The parametric study is conducted focusing on the pretreatment and refinement of the training datasets and optimization of the trained model by pruning.

2. C4.5 decision tree algorithm

The C4.5 algorithm is a representative decision tree classification technique proposed by Quinlan (1986). The model is trained with the class-labeled data to perform classification based on specific criteria (Kim *et al*. 2022). This method selects the properties that split the data best on each to have lower complexity and uncertainty for a random decision where the complexity of the dataset is defined based on the concept of entropy. The Entropy for the data partition D, Info(D) is defined as

*Corresponding author, Pfofessor
E-mail: geotech@kookmin.ac.kr
[a]Assistant Professor
[b]Engineer
[c]Professor

ISSN: 2005-307X (Print), 2092-6219 (Online)

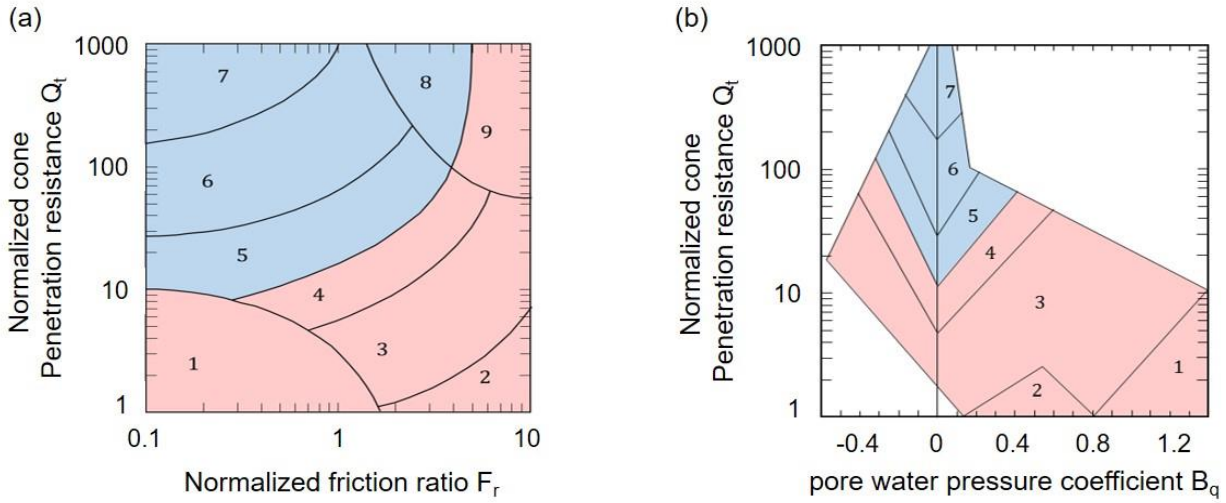

Fig. 1 Re-labeled Robertson's Classification chart (after Robertson 1990). (a) Q_t versus F_r and (b) Q_t versus B_q.

Table 1 Re-labeled classification index (after Robertson 1990)

	Soil Behavior Type	Classification
SBT1	Sensitive, fine grained	Fine-grained Soils (CH, CL, MH, ML) ; CLAYs
SBT2	Organic soils – clay	
SBT3	Clay – silty clay to clay	
SBT4	Silt mixtures – clayey silt to silty clay	
SBT9	Very stiff fine grained	
SBT5	Sand mixtures – silty sand to sandy silt	Coarse-grained Soils (SC, SM, SW, SP, GW, GP) ; SANDs
SBT6	Sands – clean sand to silty sand	
SBT7	Gravelly sand to dense sand	
SBT8	Very stiff sand to clayey sand	

$$Entropy(S) = Info(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (1)$$

where, Entropy(S) = entropy for the data set S; and p_i = the probability of elements of class i.

The indicators such as information gain, gain ratio and Gini index are the quantitative results of the degree of uncertainty reduction. The C4.5 algorithm applies the Gain Ratio, which is the ratio between the amount of information gained for a given split criterion and the split entropy for the decision, and the C4.5 algorithm has the advantage of being able to process continuous and discrete data together. Once again, the main purpose of the C4.5 algorithm is to build up a decision tree to perform classifications. This tree-like decision model involves three main features. Each internal node corresponds to a test on a characteristic, each branch to the outcome of the test, and each leaf node to a class label. There are four key steps in the C4.5 algorithm that involve attribute selection, tree construction, pruning and handling missing values (see Cho et al. 2023 for flow chart in C4.5 decision tree algorithm). The hyperparameters such as stopping parameters in tree construction, confidence level parameter for tree pruning, splitting criterion, subset splits and so on, are also optimized based on the parametric study.

3. Robertson's soil classification method and test results

Robertson (2009) modified the in-situ soil parameters obtained from CPTu tests to consider the effect of total and effective overburden surcharge; thereafter, he proposed three normalized factors such as normalized cone penetration resistance Q_t, normalized friction ratio F_r, and pore water pressure coefficient B_q

$$Q_t = \frac{q_t - \sigma_{vo}}{\sigma'_{vo}} \quad (2)$$

$$F_r = \frac{f_s}{q_t - \sigma_{vo}} \times 100[\%] \quad (3)$$

$$B_q = \frac{u - u_o}{q_t - \sigma_{vo}} \quad (4)$$

where q_t = corrected cone resistance, f_s = sleeve friction, u = measured porewater pressure, u_o= static porewater pressure,

Table 2 The range of the CPT measurements at Site A (Busan, South Korea). The minimum and maximum values of each parameter are shown in the table

SCS	Number of data points	Normalized cone penetration resistance Q_t	Normalized friction ratio F_r	Pore water pressure coefficient B_q
CH	5365	0.62 ~ 131.49 (5.57)	0.29 ~ 10.43 (2.31)	-0.08 ~ 3.86 (0.63)
CL	4951	0.24 ~ 8346 (15.36)	0.00 ~ 6.38 (2.11)	-0.08 ~ 3.62 (0.55)
SC	462	3.22 ~ 276.72 (38.8)	0.23 ~ 6.00 (2.00)	-0.05 ~ 0.87 (0.36)
SM	947	1.44 ~ 200.81 (37.02)	0.00 ~ 25.65 (1.89)	-0.61 ~ 1.06 (0.12)
SP	164	6.10 ~ 159.06 (61.29)	0.16 ~ 3.89 (1.39)	-0.01 ~ 0.81 (0.27)
SL (Shell Layer)	223	2.94 ~ 27.43 (16.94)	0.09 ~ 3.65 (0.60)	-0.02 ~ 0.65 (0.05)

Table 3 The range of the CPT measurements at Site B (Hollywood, South Carolina, USA). The minimum and maximum values of each parameter are shown in the table

USCS	Number of data points	Normalized cone penetration resistance Q_t	Normalized friction ratio F_r	Pore water pressure coefficient B_q
CH	1740	1.60 ~ 289.52 (17.6)	0.03 ~ 32.56 (2.23)	-0.03 ~ 0.82 (0.25)
SC-SM	2679	0.17 ~ 10439 (434.41)	0.00 ~ 1663 (5.11)	-4.87 ~ 0.64 (-0.01)
SM-SP	8499	0.52 ~ 4214 (82.31)	0.004 ~ 49.38 (1.16)	-0.18 ~ 0.27 (-0.002)
SP	3535	1.34 ~ 984.68 (215.4)	0.13 ~ 61.57 (1.13)	-0.57 ~ 0.61 (0.0004)

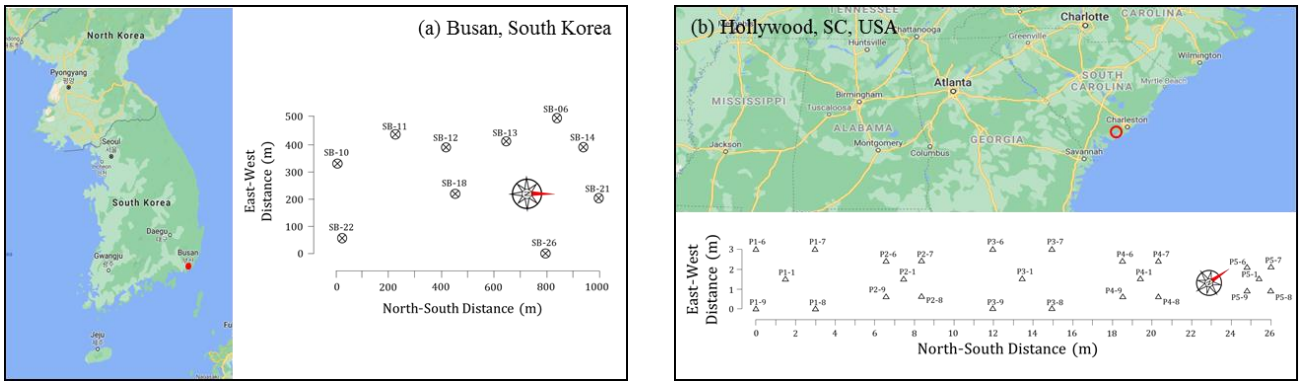

Fig. 2 Sites and CPT location. (a) Busan, South Korea and (b) Hollywood, South Carolina, USA

σ_{vo} = total overburden stress, σ'_{vo} = effective overburden stress. Then, these three parameters were used to produce the soil classification charts plotted in terms of Q_t versus F_r and Q_t versus B_q as shown in Fig. 1. These re-labeled Robertson's soil classification charts have been modified during a long time and used one of the most practical CPT-based soil classification methods in the world (Soleimani Fard and Goudarzy 2021, Arifuzzaman and Anisuzzaman 2022). The Robertson's chart classifies the soils with respect to the soil classes known as Soil Behavior Type (SBT) as listed in Table 1.

Most design standards usually provide the geotechnical design guidelines with respect to two soil types, sands and clays, or similarly coarse-grained soils and fine-grained soils. Therefore, this study attempts to propose a data-driven and machine learning-aided soil classification method within the modified Robertson's soil classification framework to achieve higher accuracy and reliability and to reflect regional characteristics of soils. This study uses the Soil Behavior Type SBTs in the revised diagrams as shown in Fig. 1 and Table 1 to relabel SBTs 1, 2, 3, 4, and 9 as CLAYs (or fine-grained soils) and SBTs 5, 6, 7, and 8 as SANDs (or coarse-grained soils).

The data used in this study were obtained from cone penetration test results at the following two construction sites: (a) Hwajeon, Busan, Korea defined as Site A (Figs, 2(a) and 2(b)) Hollywood, South Carolina, USA defined as Site B (Fig. 2(b)). For Site A, the total number of measurements is 12112 points which consists of 10316 points for clays (about 85%) and 1796 points for sands (about 15%); therefore, the Site A can be considered as a clay-dominant soil deposit. On the other hand, the total number of measurements for Site B is 16453 points that consist of 1740 points for clays (about 11%) and 14713 points for sands (about 89%). The Site B can be considered as a sand-dominant soil deposit. Detailed information on the data for Sites A and B are summarized in Tables 2 and 3.

Fig. 3 and Table 4 present the soil classification results according to Robertson's chart. For Site A (i.e., clay-dominant deposit), the results show a relatively lower classification accuracy of about 54% in Q_t-F_r chart and 58% in Q_t-B_q chart for sandy soils while the accuracy for clayey

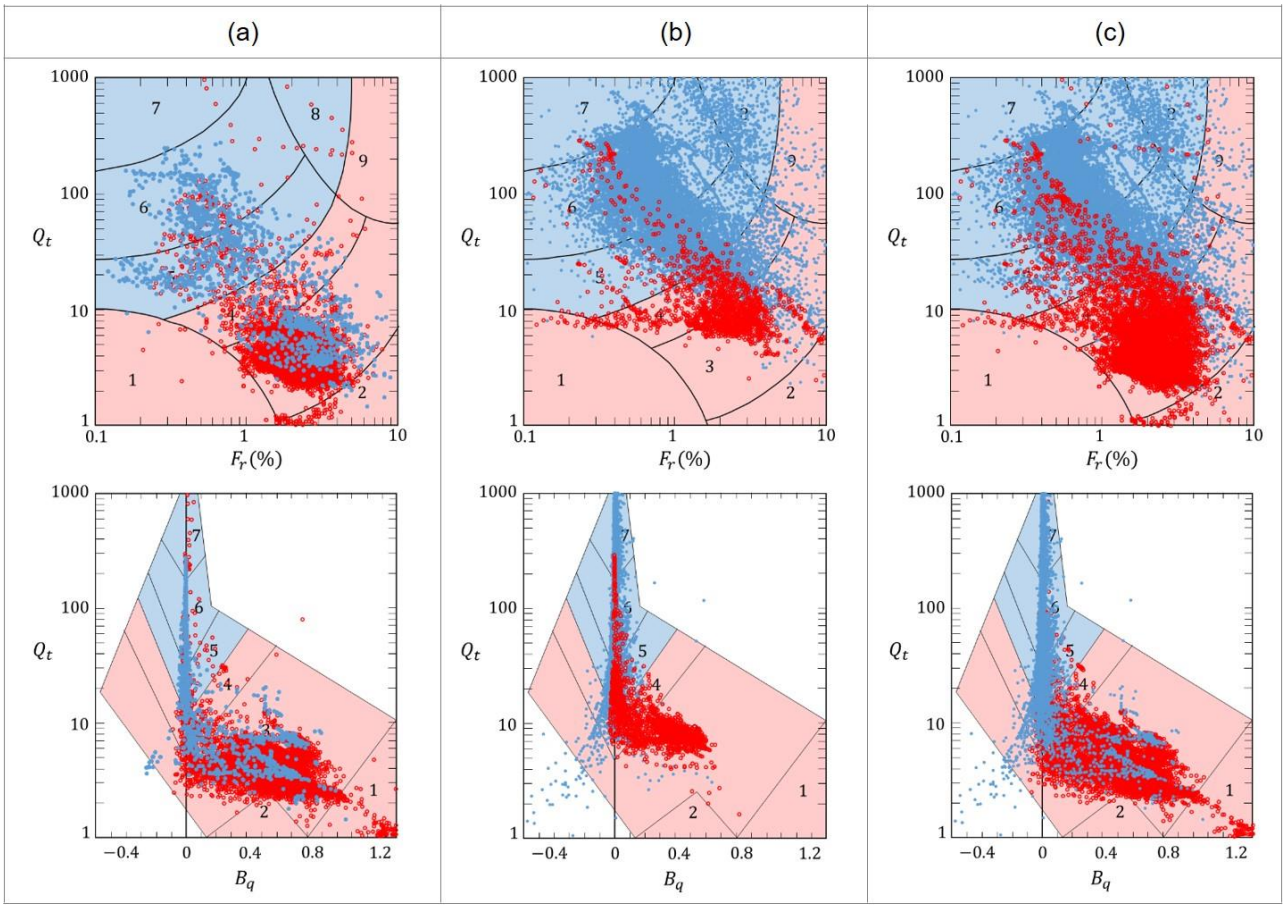

Fig. 3 Plotted data on re-labeled Robertson's classification chart. (a) Site A, (b) Site B, and (c) Site A & B. Note that red dots indicate the fine-grained soils and blue dots indicate the coarse-grained soils

soils in both charts is higher than 95% in comparison to sandy soils. On the other hand, for Site B (i.e., sand-dominant deposit), the classification accuracies for sandy and clayey soils are more similar to each other, about 80-to-90%. Clearly, the Q_t-B_q chart seems to better discriminate the sandy soils about 12% in comparison to the Q_t-F_r chart. Table 4 also summarizes the soil classification results for the dataset containing both Sites A and B where the sandy soil to clayey soil ratio becomes similar, about 42:58 (Tables 2 and 3). The results indicate that the classification results become more accurate and balanced for the dataset composed of similar proportions with respect to soil type.

4. Parametric studies for decision tree optimization

As shown above, the soil classification results using the conventional classification charts with the results of tests from the sites with typical regional characteristics can mislead the actual soil type. Therefore, this study proposes a data-driven soil classification method combined together with the C4.5 classification algorithm that reflects the local characteristics of the target area; then, its performance is reviewed and verified.

Securing and selecting appropriate training data is critical for the machine learning process. Thus, a parametric study was conducted to identify the most optimized combinations of the measurements and indices obtained from the cone penetration test results for the data-driven model training. Our previous study reveals that the combinations of Robertson's normalized factors such as Q_t, F_r and B_q are more appropriate input parameters for the model training in comparison to using the combinations of the measurements of tip resistance, side friction and excess porewater pressure (see details in Cho 2021, Cho *et al.* 2023). This study proposes an optimized data-driven soil classification model by examining a number of data refinement methods and a pruning option – Oversampling, outlier removal and controlling minimum number of instances per leaf nodes based on the model accuracy and the tree model size simultaneously. An optimized model is defined as a trained model showing acceptable accuracy even with a small-sized decision tree to avoid the overfitting problem in this study.

4.1 Data refinement methods

This section describes improvements in classification models through data preprocessing. The first method, oversampling, is one of the typical class balancing techniques, which are only necessary in the case of caring about the minority classes (Bai *et al.* 2021, Lee *et al.* 2022,

Table 4 Classification result using re-labeled Robertson's classification chart

Site(s)	Chart	CLAYs	SANDs	Mean accuracy
Site A	Q_t - F_r chart	97%	54%	91%
	Q_t - B_q chart	96%	58%	90%
Site B	Q_t - F_r chart	85%	84%	84%
	Q_t - B_q chart	77%	96%	94%
Sites A & B	Q_t - F_r chart	95%	82%	88%
	Q_t - B_q chart	93%	92%	93%

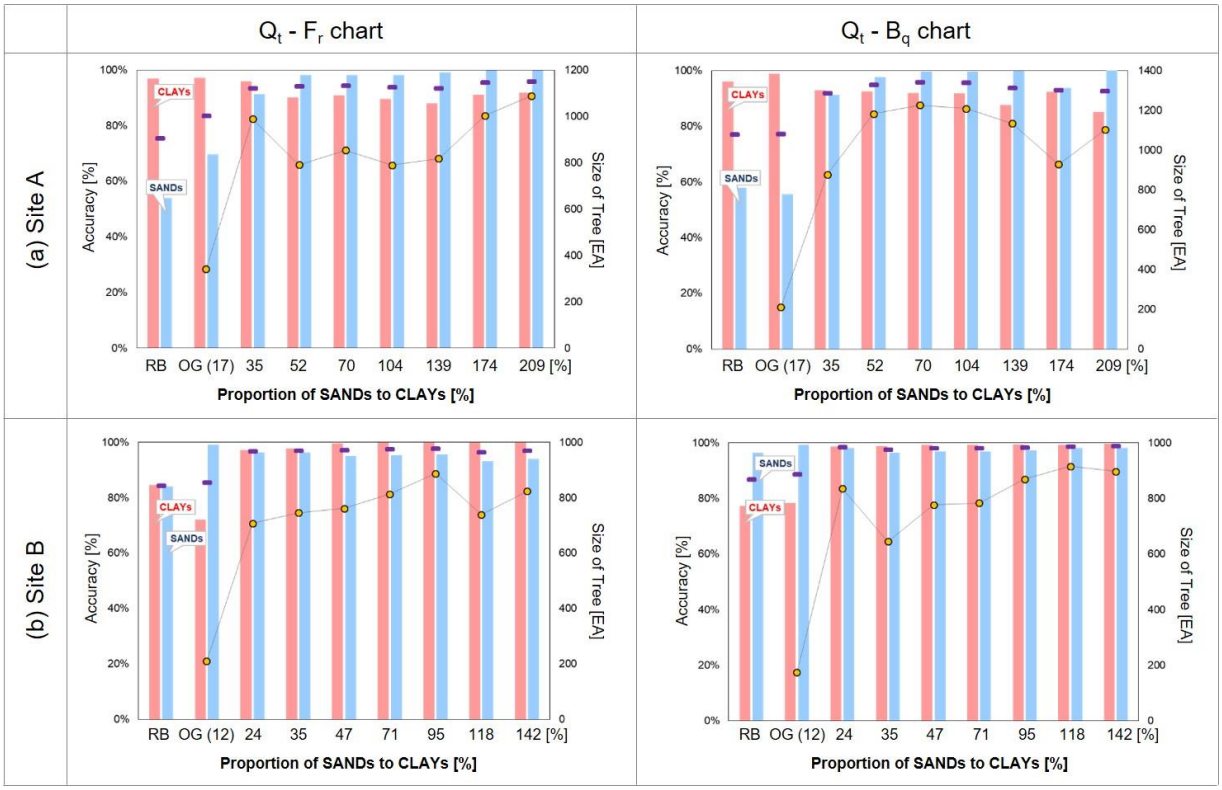

Fig. 4 Result of data refinement in Q_t - F_r chart and Q_t - B_q chart - Oversampling. (a) Site A and (b) Site B. The classification accuracy of CLAY class is indicated by the red bar, the classification accuracy of SAND class by the blue bar, and the average classification accuracy is presented by the purple dots. Note: RB = Robertson's chart, OG = C4.5 algorithm application to original dataset, and 35% = oversampling rate

Kwak and Ko 2022). The second method, unsupervised outlier detection, is examined. Outliers are defined as events or observations significantly different from the majority of the data (Kim and Kim 2019), and the model performance is evaluated regarding the ignorance rate of the outliers.

4.1.1 Oversampling

Oversampling is making copies of a minority class to attain the same number of examples as the majority class has the same contributions of each class to model training without obtaining more data in the minority class. Oversampling is considered as an alternative to resolve class imbalance problem if it is difficult to set the appropriate class weights (Ma and He 2013, Demir and Şahin 2022)

In the case of Site A where CLAY deposits are dominant, the number of SAND class data points is just 17% of that of CLAY class data points (i.e., the SAND class is the minority in Site A). By contrast, SAND deposits are dominant in Site B and the number of CLAY class data points is about 12% of that of SAND class data points (i.e., the CLAY class is the minority in Site B). By applying the oversampling technique to the minority class data, the class balance is adjusted more similarly and model performance is evaluated accordingly.

Fig. 4(a) shows the results of classification by applying C4.5 algorithm to the Site A data. The classification accuracy of the CLAY class is indicated by the red bar, the classification accuracy of the SAND class by the blue bar, and the average classification accuracy is presented by the purple dots. The size of tree is an indicator of the complexity of the classification tree of each model and this can be identified by yellow-colored circles. In the case of Qt-Fr chart for Site A, when applied to Robertson's

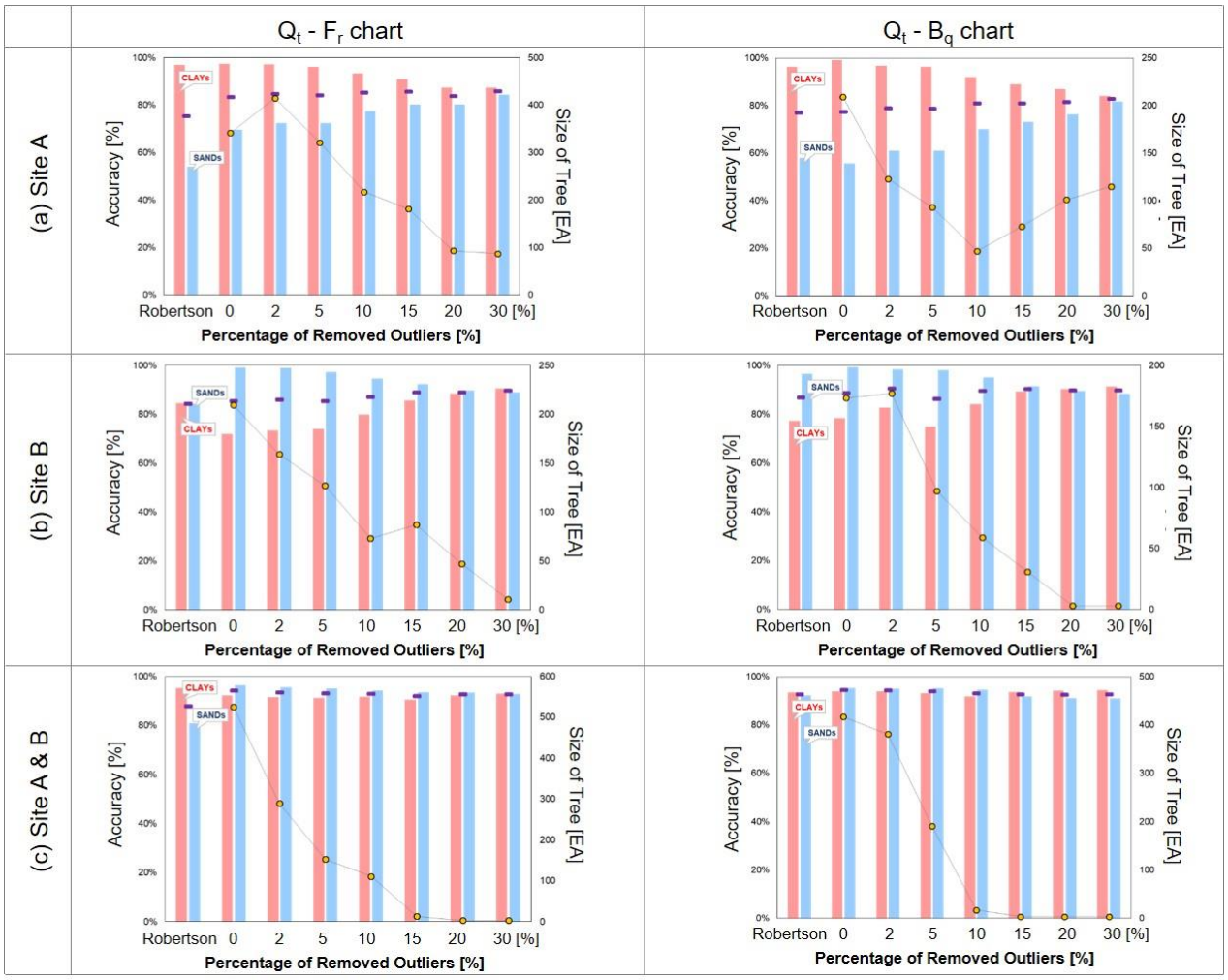

Fig. 5 Result of data refinement in Qt - Fr chart and Qt - Bq chart - Outlier removal. (a) Site A and (b) Site B, and (c) Sites A & B. The classification accuracy of CLAY class is indicated by the red bar, the classification accuracy of SAND class by the blue bar, and the average classification accuracy is presented by the purple dots. Note: % = outlier removal rate

classification chart, the classification accuracy for the SAND class data is found to be 54% (Table 4); however, when classified using C4.5 algorithm, it increases to 70% (Qt-Fr chart in Fig. 4(a)). When the class balance is changed by oversampling of the SAND class data up to 35% of the CLAY class data in the original data, the classification accuracy of the SAND class increases from 70% to 90%. When applied to the Qt-Bq chart, using C4.5 algorithm to the original dataset, there is no significant difference from Robertson's chart in terms of the classification accuracy; however, oversampling of the SAND class data by 35% of the CLAY class data increases the SAND class accuracy from 56% to 91% (Table 4 and Fig. 4(a)).

In the same way as Site A, the results at Site B are presented in Fig. 4(b). Both of the Qt-Fr chart and the Qt-Bq chart show no significant performance improvement when C4.5 algorithm is applied to the original data; yet, the classification accuracy of the CLAY class increases up to more than 95% by oversampling of the CLAY class data to more than 24% of the SAND class data (Table 4 and Fig. 4(b)). According to the results for Sites A and B, oversampling of the minority class significantly improves the trained model performance with only about 30-to-40% of the majority class.

4.1.2 Outlier removal

As mentioned above, the events or the observations significantly different from the majority of the data can be identified as outliers, and the model performance can be improved by ignoring the detected outliers in the model training. Therefore, the model performance is examined in terms of the removal rate of the outliers where the outliers are defined based on the Euclidian distance of the data points from the mean point of each class in the classification chart.

Fig. 5(a) shows the evolution of the classification accuracy and classification tree complexity with respect to the outlier removal rate in the original data of Site A, i.e., the CLAY dominant deposit. As the outlier removal rate increases, the classification accuracy of the SAND class is gradually improved; however, the accuracy of the CLAY class tends to decrease in both Q_t-F_r chart and Q_t-B_q chart.

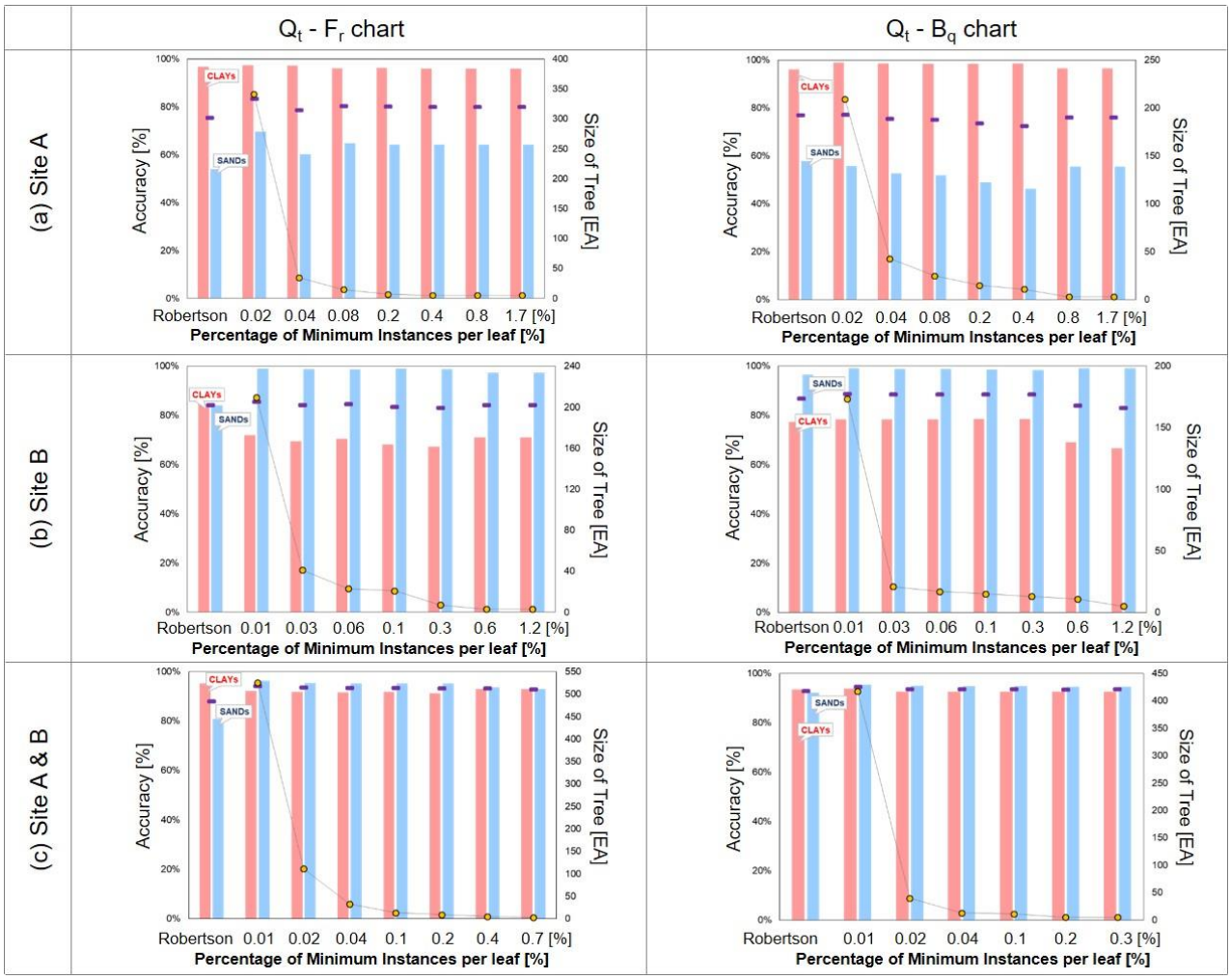

Fig. 6 Effects of model tree pruning - Controlling MNI. (a) Site A and (b) Site B, and (c) Sites A & B. The classification accuracy of CLAY class is indicated by the red bar, the classification accuracy of SAND class by the blue bar, and the average classification accuracy is presented by the purple dots. Note: % = outlier removal rate

In the case of Site B, i.e., the SAND dominant deposit, the classification accuracy of the CLAY class, the minority at Site B, is improved by ignoring the outliers, and the accuracy of the SAND class tends to decrease as shown in Fig. 5(b). Fig. 5(c) shows that in the case of Sites A & B, combining the data from Site A and Site B balances the soil classes similarly; yet, the outlier removal does not change classification accuracy significantly. However, it is confirmed that the complexity of the trained tree model decreases remarkably as the outlier removal rate increases. Such a simplification of the tree model by the outlier removal can also be observed in the cases of Sites A and B. Comparing the classification results of those three cases abovementioned in Figs. 5(a)-5(c), it is possible to obtain the trained tree model with improved performance by removing only about 10-15% of outliers.

4.2 Controlling minimum number of instances for leaf nodes – Tree pruning

The previously discussed data preprocessing methods confirm improvement in the trained model performance. However, the trained models with higher accuracy often require a higher complexity of the classification tree, which can be indicated in terms of the tree size. In general, the larger size of the classification tree leads to a more complex classification process which may overfit the model to the training data. In this context, model optimization through pruning is considered to prevent overfitting of the trained models.

The complexity and classification accuracy of the classification tree are examined simultaneously by controlling the minimum number of instances per leaf node, hereinafter MNI, and the performance results are shown in Fig. 6(a) for site A, Fig. 6(b) for Site B and Fig. 6(c) for Sites A & B. It can be observed that the size of the trained tree model illustrated by yellow-colored circles decreases dramatically to less than 50 by increasing MNI to only about 0.04% of the number of the total data points without significant loss in the model accuracy.

4.3 Model optimization

As discussed in the previous section, the data-driven soil

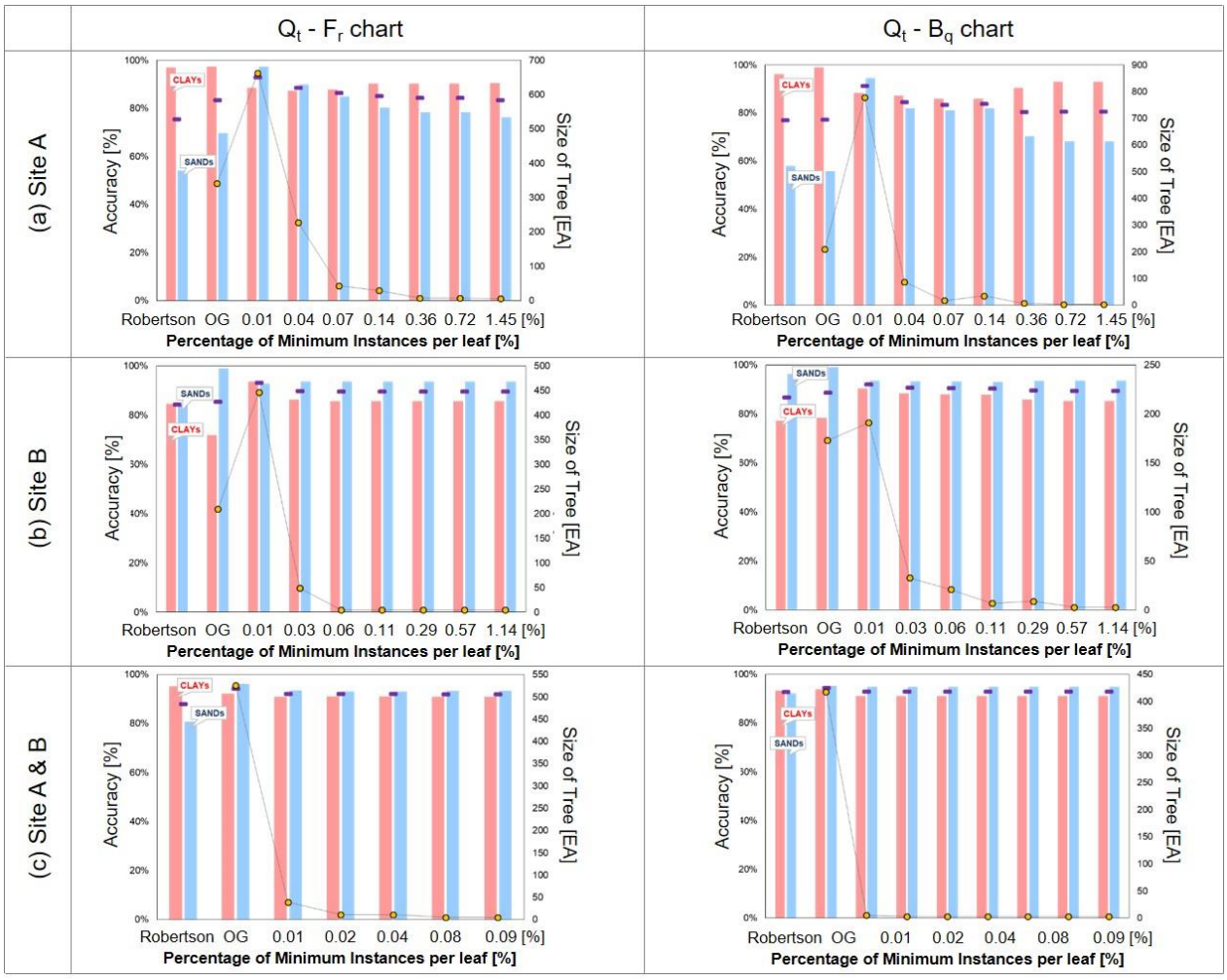

Fig. 7 Result of the model optimization conducted combining oversampling, outlier removal, and pruning method: (a) Site A and (b) Site B, and (c) Sites A & B. The classification accuracy of CLAY class is indicated by the red bar, the classification accuracy of SAND class by the blue bar, and the average classification accuracy is presented by the purple dots

classification model can be improved and optimized using the data preprocessing and pruning methods. First, oversampling the minority class to no less than 30% of the majority class can improve classification accuracy by greater than 90% when the class balances in the data are unequal. Second, this study explores the effect of the outlier removal rate on model performance and demonstrates that classification accuracy can be improved to 90% or more by ignoring the outermost 10-15% of the training data. Third, it is recommended to simplify the classification tree structure through pruning methods because the tree structure can be too much biased to the training data by overfitting. The minimum number of instances per leaf node merely about 0.1% of the total data points is enough in this case to reduce the size of the model tree to less than 50.

Those three conditions are applied together to optimize the model training and the results of optimizing the classification model through this preprocessing are presented in Fig. 7(a) for site A, Fig. 7(b) for site B and Fig. 7(c) for site A & B. Class equalization of classes on training datasets eliminate outliers and can improve class classification accuracy with relatively small amounts, and a model with lower tree complexity and higher classification accuracy can be proposed through minimum instance count adjustment. In Site A, the classification accuracy of sand layers, which was 50-to-60% when classified by Robertson's chart, can be achieved by optimization, and despite the relatively simple structural classification tree through MNI adjustment, the average accuracy was maintained without much difference. In site B, model optimization also allows models with simple classification trees to be obtained while maintaining and improving overall accuracy. In the case of Sites A & B where class equalization was achieved by combining data from the two regions, an optimal classification model was obtained by eliminating 12% of the outliers and MNI adjustments.

The local data-driven CPT soil classification charts are prepared in Fig. 8 based on the parametric study results. As summarized in the tables, these data-driven tree models can classify soils more correctly in comparison to the conventional Robertson's charts. It is worth to noting that the areas for the minority class tend to be larger than

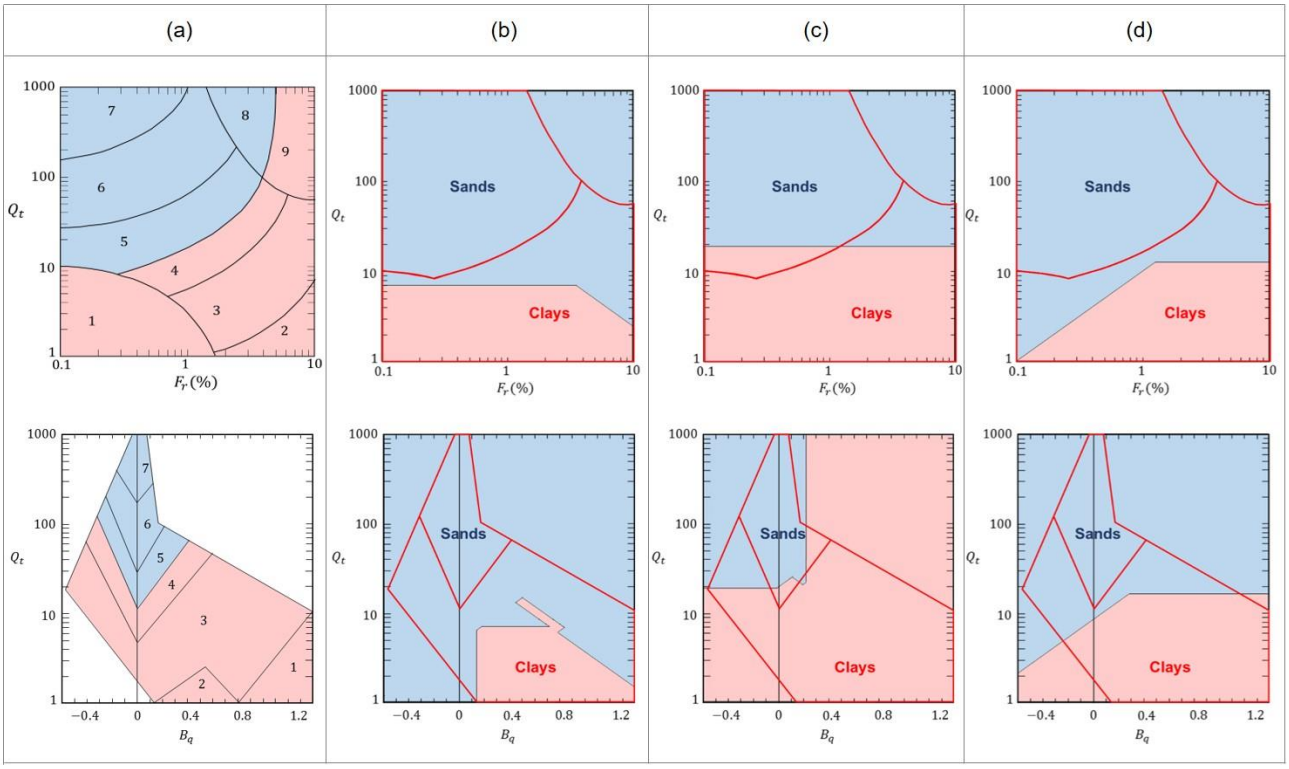

Fig. 8 CPTu-based soil classification charts. (a) Robertson's charts, and local data driven CPTu soil classification charts based on the parametric study results. (b) Site A, (c) Site B, and (d) Sites A & B

Table 5 Results of optimum classification model

Site(s)	Model	Q_t –F_r chart				Q_t-B_q chart			
		Size of tree	CLAYs	SANDs	Average accuracy	Size of tree	CLAYs	SANDs	Average accuracy
Site A	Original (C4.5)	341	97%	70%	83%	209	99%	56%	77%
	Optimum	29	90%	86%	85%	35	86%	81%	84%
	Robertson	N/A	97%	54%	75%	N/A	96%	58%	77%
Site B	Original (C4.5)	209	72%	99%	85%	173	78%	99%	89%
	Optimum	49	86%	94%	90%	33	88%	93%	91%
	Robertson	N/A	85%	84%	84%	N/A	77%	96%	87%
Sites A & B	Original (C4.5)	525	92%	96%	94%	417	94%	95%	95%
	Optimum	11	91%	93%	92%	5	91%	95%	93%
	Robertson	N/A	95%	81%	88%	N/A	93%	92%	93%

Robertson's charts because all of the approaches used in this study such as oversampling, outlier removal, and MNI control, cause the minor class data to be overestimated and increase the accuracy of the minority class more than the majority class to improve the overall classification accuracy.

In the case of Site A (Fig. 8(b)), where the boring data reports that the stiff clay was rarely found, there are few spots where negative excess pore water pressure was measured during cone penetration; hence, the data-driven classification chart sets the sand area a little bit larger. Overall, the results of tree model training utilizing class-balanced data, as shown in Fig. 9 and Table 5, indicate a higher similarity to Robertson's charts, implying that Robertson's charts may be used for soil classification in most circumstances.

4.4 Multi-variate classification (Q_t - F_r - B_q)

Robertson proposed two classification diagrams such as Q_t-F_r chart and Q_t-B_q chart, each separately considering the effect on cone penetration resistance and pore water pressure. However, when the classification results of each chart are compared, it is confirmed that the same soil is

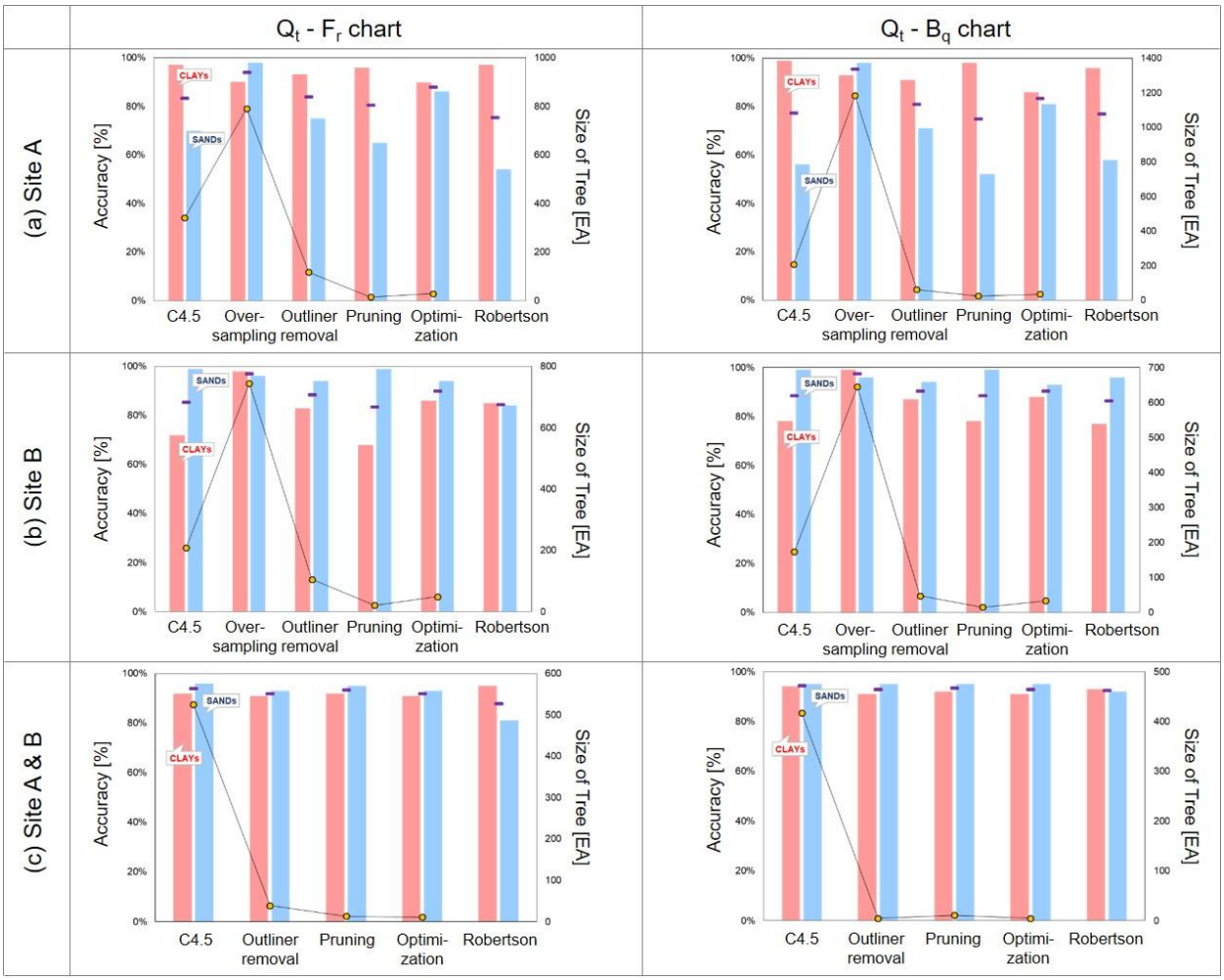

Fig. 9 Comparison between model optimization and other methods. (a) Site A and (b) Site B, and (c) Sites A & B. The classification accuracy of CLAY class is indicated by the red bar, the classification accuracy of SAND class by the blue bar, and the average classification accuracy is presented by the purple dots

often classified into different groups. Thus, we reviewed the performance of a three-dimensional classification model that considered factors for both diagrams simultaneously and compared them with two-dimensional analyses. Like the previous analyses, three-dimensional models also perform optimization of the model, identifying the impact of the data preprocessing process.

4.4.1 Treatment #1; Oversampling

The performance review through class equalization in the two regions is shown in Figs. 10(a) and 10(b). First, the oversampling leads to high accuracy in both 2D and 3D analyses (see Figs. 9(a) and 9(b) for 2D and Figures 10a and 10b for 3D analyses). For 3D analyses, in the case of Site A, the classification accuracy of the sand layer increased from about 75% to about 98% (Fig. 10(a)). Similarly, for Site B, the classification accuracy of the clay layer changed from ~80% to 99% (Fig. 10(b)). However, it is developed into a tree of very high complexity as well as high classification accuracy and requires appropriate processing. Likewise, it can be confirmed that even if only about 30% of the dominant data is secured, sufficient numerical accuracy is provided.

4.4.2 Treatment #2; Outlier removal

The review of performance improvement due to the removal of outliers in the 3D analysis is shown in Fig. 10 (see Fig. 10(a) for Site A; Figure 10a for Site B; and Fig. 10(c) for Sites A & B). 3D Analyses conducted as part of this study indicate that as the outlier removal rate increases, the complexity of the tree tends to decrease, and the overall classification performance also shows improved values. Like the two-dimensional analysis, approximately 12% of the removal is determined by the appropriate removal rate.

4.4.3 Pruning methods: MNI (Minimum number of instances per leaf)

The result of optimization with the adjustment of the minimum number of instances of each leaf is shown in Fig. 10 (see Fig. 10(a) for Site A; Fig. 10(a) for Site B; and Fig. 10(c) for Sites A & B). Interestingly, the previous analyzes showed that the simpler the tree structure became by adjusting certain factors, the more accurate the classification of the vulnerable, in which case the classification of the sand layer tends to decrease as the tree structure becomes simpler. For site A and site B, when the minimum number of instance allocations is set to

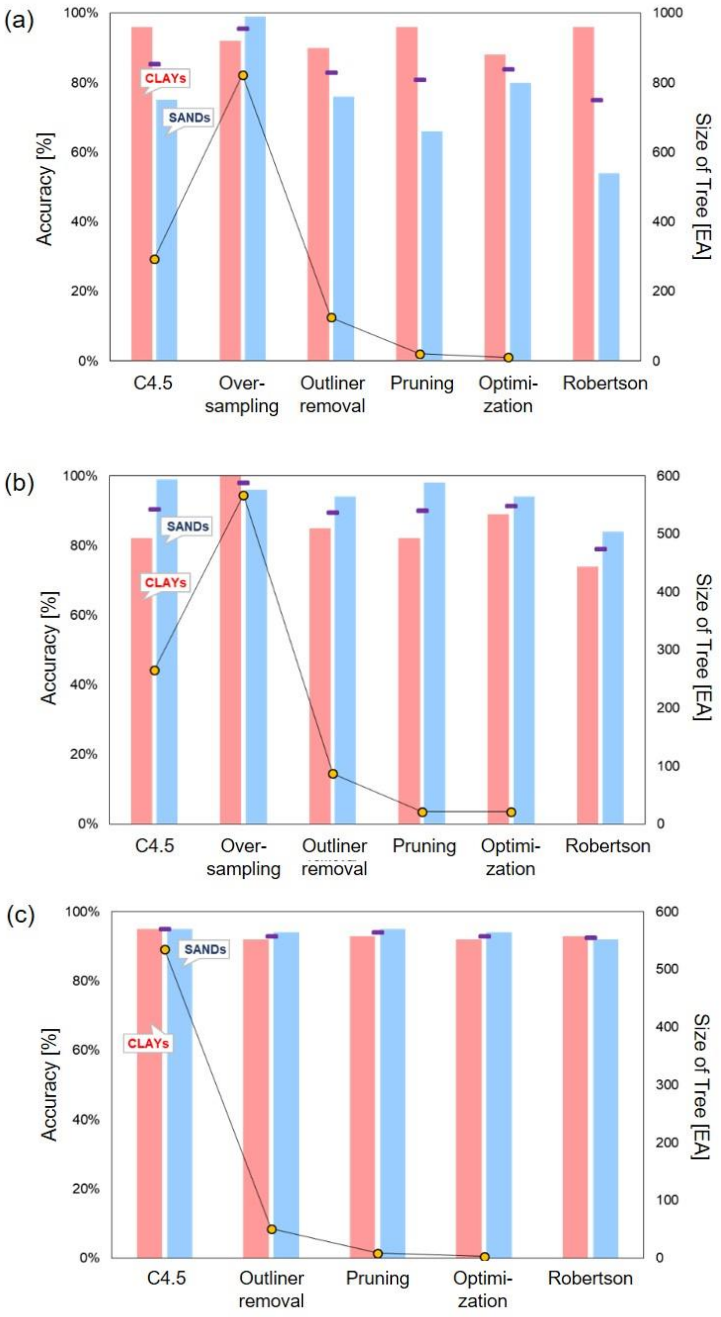

Fig. 10 Result of 3D classification model optimization. (a) Site A, (b) Site B, and (c) Sites A & B

approximately 0.1% of the total data, low complexity tree structures can be obtained while maintaining overall accuracy.

Just utilizing data mining techniques for two-dimensional analysis, such as Robertson's classification chart, which is a traditional classification method, can confirm improved classification performance. However, when three factors are considered at the same time, the classification process is a little more complicated, but it can be confirmed that the classification accuracy is excellent and improved. Similarly, the 3D model was also optimized and the performance of each model was compared.

4.4.4 Optimum classification model

Previously, for the optimization of the 3D classification model, the performance of data preprocessing and pruning was reviewed and results are shown in Fig. 10. For each Site A and Site B where classes were unevenly distributed, data preprocessing was performed such as class equalization and outlier removal, and the performance was reviewed by adjusting the minimum number of instances. In the case of Sites A & B, where classes are evenly distributed, only outliers were removed and the performance was assessed according to the minimum instance adjustment.

Table 6 Comparison of classification model

Dataset	Classification method	Size of tree		CLAYs		SANDs		Average accuracy	
		Before	After	Before	After	Before	After	Before	After
Site A	Robertson	N/A		96%		54%		75%	
	C4.5 (2D; F_r)	341	29	97%	90%	70%	86%	83%	85%
	C4.5 (2D; B_q)	209	35	99%	86%	56%	81%	77%	84%
	C4.5 (3D)	293	11	96%	88%	75%	80%	86%	84%
Site B	Robertson	N/A		74%		84%		79%	
	C4.5 (2D; F_r)	209	49	72%	86%	99%	94%	85%	90%
	C4.5 (2D; B_q)	173	33	78%	88%	99%	93%	89%	91%
	C4.5 (3D)	265	21	82%	89%	99%	94%	91%	91%
Site A & B	Robertson	N/A		93%		81%		87%	
	C4.5 (2D; F_r)	525	11	92%	91%	96%	93%	94%	92%
	C4.5 (2D; B_q)	417	5	94%	91%	95%	95%	95%	93%
	C4.5 (3D)	535	3	95%	92%	95%	94%	95%	93%

Table 6 compares the original result and optimization result of each classification model. Utilizing C4.5 algorithms, the performance of the proposed classification models showed better accuracy than that of Robertson's classification chart, and the classification performance of the three-dimensional model was superior to the two-dimensional analysis that separately considers the impact on friction and clearance water pressure. Optimization not only improves classification accuracy for strata, which showed relatively low classification performance but also provides an appropriate size classification tree to confirm a concise yet accurate classification method.

5. Conclusions

In this paper, it is confirmed that an objective classification method considering the characteristics of the target area can be presented by using the cone penetration test data conducted in the field and C4.5, which is a representative classification algorithm. When classifying the strata as a result of the cone penetration test, it was often applied to the classification diagram proposed by existing researchers to roughly confirm the stratum distribution in the area. However, due to the limitations of existing charts that could not reflect various locality characteristics, engineering judgments could be erroneous due to results that did not fit the actual site. Therefore, in this paper, an analysis was conducted using the C4.5 classification algorithm to classify the two regions where the distribution of the soil is highly unequal, and the following results were obtained.

(1) When applying the C4.5 classification algorithm to data organized by factors considering the shape of Robertson's classification charts based on the measurements of the cone penetration test, the classification accuracy was not improved or even lower. Thus, it was intended to propose an optimal classification model by pre-treating the training data appropriately. The fit of the model was evaluated by simultaneously considering the classification accuracy for each class and the complexity of the classification tree.

(2) The performance of preprocessing through class equalization was reviewed by increasing the number of data in a specific stratum. The sand layer data of site A, where the clay layer is dominant, and the clay layer data of site B, where the sand quality is dominant, were increased and applied to the analysis. As a result, it was confirmed that a classification model with adequate and improved performance can be obtained even if only a value corresponding to about 30% of the dominant class is secured.

(3) By checking the distribution of each soil layer, data points that are far from the center point were defined as outliers, and the degree of improvement in classification performance was confirmed by removing the outliers. In general, as the removal amounts of defined outliers increased, the classification accuracy for the vulnerable group increased. The most optimal classification results were shown when approximately 12% of the outliers were removed from the overall data.

(4) To prevent overfitting of the complicated model due to data preprocessing, the performance of the model by adjusting the minimum number of instances per leaf, one of the pruning methods, was reviewed. It can be seen that the size of the classification tree is rapidly reduced even if only a value corresponding to about 0.04% of the total data is set as the minimum number of instances. As a result of examining the model performance according to the change of the minimum instance, an appropriate classification model is provided when a value corresponding to about 0.1% of the total data is set.

(5) As above, when constructing a classification model using the C4.5 algorithm, it is possible to propose a classification method that reflects the characteristics of the target area because it provides classification results that reflect the characteristics of the data used. In addition, stratum classification can be performed quickly and with high accuracy, and it has the advantage of intuitively grasping what type of soil layer the area is composed of.

(6) In addition, the performance of the classification method that considers the effect of circumferential friction force and the effect of pore water pressure simultaneously without considering separately was reviewed. When the three factors were considered simultaneously, rather than considering the effects of each factor separately, the classification performance with higher accuracy was shown, and by optimization, a concise classification tree can be obtained while maintaining the accuracy. In this way, when stratum classification is performed using the C4.5 classification algorithm for field test data, it is possible to propose a stratum classification method that reflects the characteristics of the target area.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1088527).

References

Arifuzzaman and Anisuzzaman, M., (2022), "An initiative to correlate the SPT and CPT data for an alluvial deposit of Dhaka city", *Int. J. Geo-Eng*., **13**(1), 5. https://doi.org/10.1186/s40703-021-00170-3.

Bai, X.D., Cheng, W.C., Ong, D.E. and Li, G. (2021), "Evaluation of geological conditions and clogging of tunneling using machine learning", *Geomech. Eng*., **25**(1), 59-73. https://doi.org/10.12989/gae.2021.25.1.059.

Begemann, H.K.S. (1965), "The friction jacket cone as an aid in determining the soil profile", *Proceedings of the 6th International Conference on Soil Mechanics and Foundation Engineering*, ICSMFE, **1**, 17-20.

Bhargavi, P. and Jyothi, S. (2011), "Soil classification using data mining techniques: a comparative study". *Int. J. Eng. Trends Technol*., **2**(1), 55-59.

Bhattacharya, B. and Solomatine, D.P. (2006), "Machine learning in soil classification", *Neural Networks*, **19**(2), 186-195. https://doi.org/10.1016/j.neunet.2006.01.005.

Cai, Y., Li, J., Li, X., Li, D. and Zhang, L. (2018), "Estimating soil resistance at unsampled locations based on limited CPT data", *B. Eng. Geol. Environ*., **78**, 3637-3648. https://doi.org/10.1007/s10064-018-1318-2.

Cal, Y. (1995), "Soil classification by neural-network", *Adv. Eng. Softw*., **22**(2), 95-97. https://doi.org/10.1016/0965-9978(94)00035-H

Cao, Z. and Wang, Y. (2013), "Bayesian approach for probabilistic site characterization using cone penetration tests", *J. Geotech. Geoenviron. Eng*., **139**(2), 267-276. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000765.

Cao, Z., Zheng, S., Li, D. and Phoon, K. (2018), "Bayesian identification of soil stratigraphy based on soil behaviour type index". *Can. Geotech. J*., **56**(4), 570-586. https://doi.org/10.1139/cgj-2017-0714.

Cho, S. (2021), "A study on data mining techniques for soil classification method using cone penetration test results", Master's thesis, Kookmin University, South Korea.

Cho, S., Kim, H.S. and Kim, H. (2023). "Locally specified CPT soil classification based on machine learning techniques", *Sustainability*, **15**(4), 2914.

Das, S.K. and Basudhar, P.K. (2009), "Utilization of self-organizing map and fuzzy clustering for site characterization using piezocone data", *Comput. Geotech*., **36**(1-2), 241-248. https://doi.org/10.1016/j.compgeo.2008.02.005.

Farhadi, M.S. (2019), "An integrated optimization-game theory model for CPT-based subground stratification", 2019 TC304 Student Contest.

Demir, S. and Sahin, E.K. (2022). "Comparison of tree-based machine learning algorithms for predicting liquefaction potential using canonical correlation forest, rotation forest, and random forest based on CPT data", *Soil Dyn. Earthq. Eng*., **154**, 107130. https://doi.org/10.1016/j.soildyn.2021.107130.

Douglas, B.J. and Olsen, R.S. (1981), "Soil classification using electric cone penetrometer", Symposium on Cone Penetration Testing and Experience, Geotechnical Engineering Division, ASCE, St. Louis, Missouri, (Missouri, 1981), 209-227

Juang, C.H., Jiang, T. and Christopher, R.A. (2001), "Three-dimensional site characterization: neural network approach", *Geotechnique*, **51**(9), 799-809. https://doi.org/10.1680/geot.2001.51.9.799.

Kwak, N.S. and Ko, T.Y. (2022), "Machine learning-based regression analysis for estimating Cerchar abrasivity index", *Geomech. Eng*., **29**(3), 219-228. https://doi.org/10.12989/gae.2022.29.3.219.

Kim, C.H., Im, J.C. and Kim, Y.S. (2008), "Study on the applicability of CPT based soil classification chart", *KSCE J. Civil Environ. Eng. Res.*, **28**(5), 293-301 (in Korean).

Kim, C.H., Im, J.C., Kim, Y.S. and Joo, N.A. (2008). "New soil classification system using cone penetration test", *J. Korean Geotech. Soc.*, **24**(10), 57-70.

Kim, H.S. and Kim, H.K. (2019). "Optimizing site-specific geostatistics to improve geotechnical spatial information in Seoul, South Korea", *Arab. J. Geosci*., **12**, 1-20. https://doi.org/10.1007/s12517-018-4171-5.

Kim, Y., Hong, J., Shin, J. and Kim, B. (2022), "Shield TBM disc cutter replacement and wear rate prediction using machine learning techniques", *Geomech. Eng*., **29**(3), 249-258. https://doi.org/10.12989/gae.2022.29.3.249.

Lee, J.S., Park, J., Kim, J. and Yoon, H.K. (2022), "Study of oversampling algorithms for soil classifications by field velocity resistivity probe". *Geomech. Eng*., **30**(3), 247-258. https://doi.org/10.12989/gae.2022.30.3.247.

Ma, Y. and He, H. (2013), "Imbalanced learning: foundations, algorithms, and applications", University of Rhode Island: Kingston, RI, USA, 2013.

Najjar, Y.M. and Basheer, I.A. (1996), "Neural network approach for site characterization and uncertainty prediction", *Geotechnical Special Publication, ASCE*, **58**(1), 134-148.

Odeh, I.O.A., Chittleborough, D.J. and McBratney, A.B. (1992), "Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships", *Soil Sci. Soc. Am. J*., **56**(2), 505-516.

Park, J., Lee, J.S., Jang, B.S., Min, D.H. and Yoon, H.K. (2022), "A comprehensive laboratory compaction study: Geophysical assessment". *Geomech. Eng*., **30**(2), 211-218. https://doi.org/10.12989/gae.2022.30.2.211.

Quinlan, J.R. (1986), "Induction of decision trees. Machine Learning", **1**(1), 81-106

Quinlan, J.R. (2008), "Top 10 algorithms in data mining", *Knowl. Inf. Syst*., **14**(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2.

Rizzo, D.M., Lillys, T.P. and Dougherty, D.E. (1996), "Comparisons of site characterization methods using mixed data", *Geotechnical Special Publication, ASCE*, **58**(1), 157-179.

Robertson, P.K. (1990), "Soil classification using the cone penetration test". *Can. Geotech. J*., **27**(1), 151-158. https://doi.org/10.1139/t90-014.

Robertson, P.K. (2009), "Interpretation of cone penetration tests –

a unified approach", *Can. Geotech. J*., **46**(11), 1337-1355. https://doi.org/10.1139/T09-065.

Robertson, P.K. (2016), "Cone penetration test –based soil behaviour type classification system – an updated". *Can. Geotech. J*., **53**(12), 1910-1927. https://doi.org/10.1139/cgj-2016-0044.

Robertson, P.K. and Cabal, K.L. (2014), Guide to Cone Penetration Testing 6th Edition.

Robertson, P.K. and Campanella, R.G. (1983), "SPT-CPT correlations", *J. Geotech. Div*. *ASCE*, **109**(11), 1449-1460.

Robertson, P.K. and Wride, C.E. (1998), "Evaluating cyclic liquefaction potential using the cone penetration test". *Can. J. Geotech*., **35**(3), 442-459. https://doi.org/10.1139/t98-017.

Robertson, P.K., Campanella, R.G., Gillespie, D. and Greig, J. (1986), "Use of piezometer cone data", *Proceedings of the America Society of Civil Engineers*, In-Situ 86 Specialty Conference, Blacksburg, Virginia.

Soleimani Fard, H. and Goudarzy, M. (2021), "Influence of surcharge on cone penetration test results and the inspection of various approaches for capturing its effect: a case study", *Int. J. Geo-Eng*., **12**(1), 17. https://doi.org/10.1186/s40703-021-00146-3.

IC