Machine learning techniques for cohesive soil classification in construction in Vietnam

Danh Thanh Tran[1*], Dinh Xuan Tran[1], Vinh Hoang Truong[1]

[1]Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[*]Corresponding author: danh.tt@ou.edu.vn

ARTICLE INFO

DOI:10.46223/HCMCOUJS.tech.en.15.2.3816.2025

Received: October 25th, 2024

Revised: December 15th, 2024

Accepted: December 17th, 2024

Keywords:

geotechnical engineering; KNN; machine learning; soil classification; SVM

ABSTRACT

Accurate soil classification is imperative for determining land suitability for various construction projects in construction and geotechnical engineering. The physical and mechanical properties of soil significantly influence the design of foundations, the assessment of landslide risks, and the overall stability of structures. Recognizing the limitations of traditional soil classification methods, which are often labor-intensive and time-consuming, this research introduces machine learning as a transformative tool for enhancing soil classification processes. Utilizing K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms, this study analyzes 5,869 soil samples collected from 39 construction projects in Ho Chi Minh City, Vietnam, to evaluate the efficacy of machine learning techniques in classifying construction soils. The study identifies optimal strategies that significantly improve classification accuracy through a methodical investigation that includes varying training set sizes and integrating directly obtained and indirectly derived soil features. The findings underscore the importance of incorporating liquid and plastic limits and their derived indices, with the KNN model demonstrating superior performance in specific scenarios. This research highlights the potential of machine learning to revolutionize traditional soil classification methods. It provides foundational insights for future advancements in geotechnical engineering, aiming to achieve safer, more efficient, and sustainable construction practices.

1. Introduction

Soil classification is fundamental in the construction industry, providing critical insights for foundation design, risk assessment, and project cost estimation. In construction, soils are primarily categorized into cohesive and non-cohesive types. In Vietnam, soil classification commonly follows the TCVN 9362:2012 standard (Cong thong tin dien tu Bo Xay dung, 2012), alongside other internationally recognized systems such as USCS (Unified Soil Classification System), AASHTO (American Association of State Highway and Transport Officials), and ASTM (American Society for Testing and Materials). These classification systems generally rely on particle size distribution and Atterberg limits to determine soil properties (Casagrande, 1948; Das & Sobhan, 2013). Obtaining these values necessitates laboratory experiments, including sieve and sedimentation tests for particle

composition, tests for moisture content, and determining liquid and plastic limits. Despite their critical importance, these conventional methods are often time-consuming and labor-intensive, leading to a demand for more efficient approaches.

The rise of Artificial Intelligence (AI) across various sectors, from autonomous vehicles and facial recognition systems to virtual assistants and content recommendation systems, has opened up new possibilities for geotechnical engineering. Early AI applications in this field addressed diverse challenges, such as soil parameter prediction (Mollahasani et al., 2011; Nguyen et al., 2020; Pham et al., 2020; Pham, Mahdis, et al., 2021; Zhang, Wu, et al., 2021), pile load-bearing capacity estimation (Momeni et al., 2020; Pham et al., 2022; Singh & Walia, 2017; Tran et al., 2024), retaining wall design (Ghaleini et al., 2018; Gordan et al., 2019; Koopialipoor, Murlidhar, et al., 2020), TBM operational parameters (Koopialipoor, Fahimifar, et al., 2020; Ninić et al., 2017; Zhou et al., 2021), stratigraphy thickness prediction (Zhou et al., 2019), landslide susceptibility (Pham et al., 2016; Shirzadi et al., 2017; Wang et al., 2021; Xiao et al., 2018), and properties of rocks analysis (Armaghani et al., 2014; Karimpouli & Tahmasebi, 2019), leveraging techniques like neural networks and Machine Learning (ML) algorithms. Despite the initial scarcity of deep learning applications in geotechnical research, recent years have witnessed an increasing number of studies (Zhang, Li, et al., 2021), underscoring the potential of AI to advance geotechnical practices.

Applying ML to soil classification has emerged as a promising area for improving efficiency and accuracy. Early research by Cal (1995) using neural networks laid the groundwork for AI in soil classification. Subsequent studies have explored a variety of ML algorithms, including neural networks (Goktepe et al., 2010), K-Nearest Neighbor (KNN) (Carvalho & Ribeiro, 2019), SVM (Kovačević et al., 2010; Ma, 2005), and advanced ensemble methods like Multilayer Perceptron (MLP), Random Forest (RF), AdaBoost, Tree Modeling, Gradient Boosting, XGBoost, and LightGBM (Gambill et al., 2016; Kang et al., 2022; Nguyen et al., 2022; Pham, Nguyen, et al., 2021). These studies demonstrate ML's capability to handle complex soil datasets, offering highly accurate classifications aligned with established standards and real-world conditions. Notably, integrating operational parameters from Tunnel Boring Machines (TBM) and other advanced data sources has further expanded the potential of ML in soil classification, underscoring the versatility and adaptability of ML algorithms in addressing geotechnical challenges.

The review highlights the growing but limited body of research utilizing AI and ML in geotechnical engineering, focusing on soil classification. The wide range of ML algorithms available demonstrates the potential for customized solutions tailored to specific research goals. However, despite the advancements, the application of ML in soil classification, especially using Vietnamese standards for construction projects (TCVN 9362:2012 - Cong thong tin dien tu Bo Xay dung, 2012), remains underexplored. This gap presents an opportunity for further research, aiming to harness the full capabilities of AI and ML to enhance the accuracy, efficiency, and cost-effectiveness of soil classification and, by extension, foundation design and risk management in construction projects.

In summary, integrating traditional soil classification methods with advanced Machine Learning (ML) and Artificial Intelligence (AI) techniques represents a pivotal shift toward more data-driven, analytical approaches in geotechnical engineering. This evolution promises to refine existing practices and unlock new opportunities for innovation and increased efficiency in construction and related fields.

2. Data collection and methods

This study applies machine learning algorithms to classify construction soils. Data for training and testing were collected from soil investigation reports across 39 construction projects in Ho Chi Minh City, Vietnam, focusing mainly on cohesive soils like clay and silty clay. 5,869 soil samples were analyzed, focusing on 13 soil parameters obtained directly from laboratory experiments and indirectly through calculations. These parameters include sample depth, clay content, moisture content, bulk unit weight, particle density, liquid limit, plastic limit, and others related to soil's physical properties. The study classifies soils into nine types per Vietnamese Standard TCVN 9362:2012 (Cong thong tin dien tu Bo Xay dung, 2012), ranging from silty sand to various clay types.

2.1. Data

Soil data in this research were sourced from investigation reports on 39 construction projects across different Ho Chi Minh City districts. The sampling data, which includes a depth ranging from 50m to 100m, revealed approximately 05 to 07 soil layers at each borehole. The primary focus was on cohesive soils, such as clay, silty clay, and silty sand, with a minor presence of non-cohesive soils like sand, which were not within the scope of this study. A total of 5,869 soil samples were collected and analyzed.

2.1.1. Data collection

The dataset comprises 5,869 soil samples, classified into nine different types according to the Vietnamese Standard TCVN 9362:2012 (Cong thong tin dien tu Bo Xay dung, 2012). The number of samples and the percentage of each soil type are as in Table 1. These figures illustrate the diversity of soil types encountered in the collected dataset, providing a comprehensive foundation for applying machine learning algorithms for soil classification.

Table 1

Number of Soil Samples

No	Soil types	Number of samples	Proportion (%)
1	Plastic silty sand	1,957	33.34
2	Semi-hard silty clay	219	3.73
3	Hard plastic silty clay	429	7.31
4	Soft plastic silty clay	236	4.02
5	Hard clay	569	9.70
6	Semi-hard clay	640	10.90
7	Hard plastic clay	790	13.46
8	Soft plastic clay	227	3.87
9	Liquid clay	802	13.67
	Total	5,869	100

Source. Data analysis result of the research

2.1.2. Input features

The study considers 13 input features representing specific soil characteristics from both laboratory experiments (Directly Obtained Features: X1, X2, X3, X4, X5, X6, X7) and calculations (Indirectly Obtained Features: X8, X9, X10, X11, X12, X13). The statistical summary of all input variables is presented in Table 2 below, which includes the mean, standard deviation, minimum, and maximum values.

Table 2

Statistical Summary of Input Variables

	Variable	Mean	Standard deviation	Min	Max
X1	Sample depth	31.07	20.90	0.50	99.80
X2	Clay content	29.04	18.38	1.80	85.00
X3	Moisture content	31.36	22.25	11.40	118.93
X4	Bulk unit weight	1.92	0.19	1.34	2.33
X5	Particle density	2.69	0.04	2.53	3.40
X6	Liquid limit	39.44	17.28	13.80	102.59
X7	Plastic limit	21.91	8.37	7.60	63.80
X8	Dry unit weight	1.51	0.31	0.64	2.06
X9	Void ratio	0.90	0.57	0.40	3.13
X10	Porosity	44.15	11.11	28.60	75.80
X11	Degree of saturation	90.99	6.49	58.30	100.00
X12	Plasticity index	17.53	10.43	2.83	51.40
X13	Liquidity index	0.47	0.45	-0.71	2.75

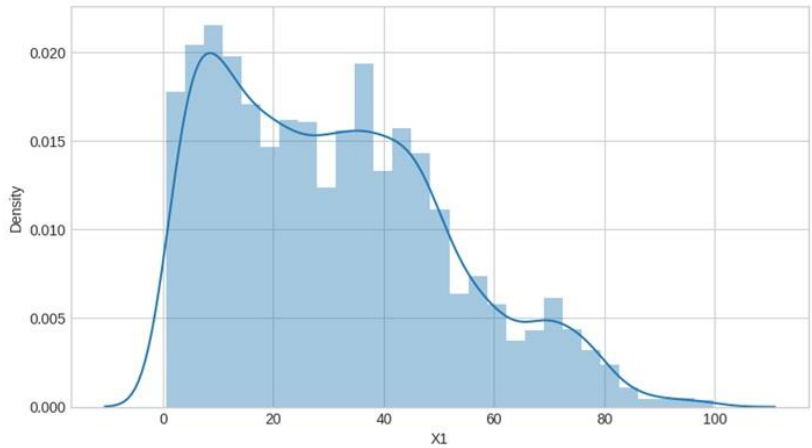
Source. Data analysis result of the research

To clearly understand the data distribution, we have included histograms for each input variable (Figure 1 to Figure 13). These histograms illustrate the frequency distribution of each variable, highlighting the range, central tendency, and variability within the dataset. Below are the directly obtained features.

(1) Sample depth (X1): The depth at which the soil sample was taken, ranging from 0.5m to 99.8m. This parameter is crucial as it reflects the soil layer's position, which could influence its properties due to varying pressure and environmental conditions.

Figure 1

Histogram of Sample Depth (X1)

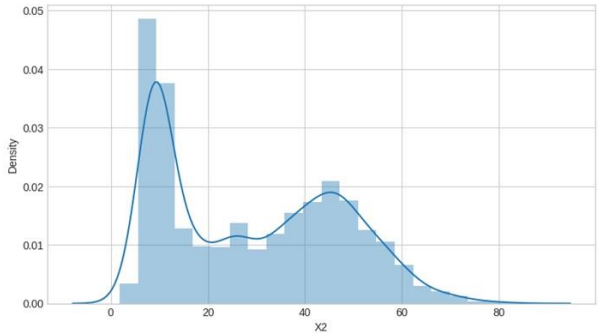


Source. Data analysis result of the research.

(2) Clay content (X2): The percentage of particles in the soil sample smaller than 0.002mm, indicating the presence of clay. It ranges from 1.8% to 85%, highlighting the soil's fine-grained nature and its potential impact on compressibility and plasticity.

Figure 2

Histogram of Clay Content (X2)

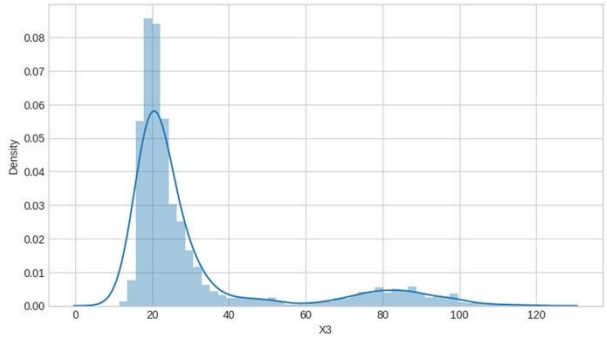


Source. Data analysis result of the research

(3) Moisture content (X3): Represents the water content in the soil, affecting its strength and compaction. It varies from 11.4% to 118.93%, showcasing the samples' wide range of water saturation levels.

Figure 3

Histogram of Moisture Content (X3)

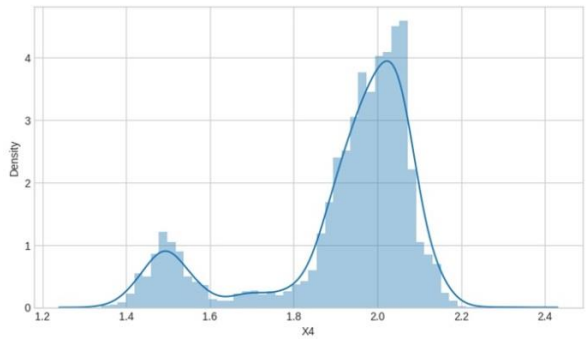


Source. Data analysis result of the research

(4) Bulk unit weight (X4): The weight of the soil per unit volume, which includes the weight of solids and the weight of the voids (air and water), ranging from 1.34 to 2.33 g/cm³. It's indicative of the soil's density and compaction level.

Figure 4

Histogram of Bulk Unit Weight (X4)

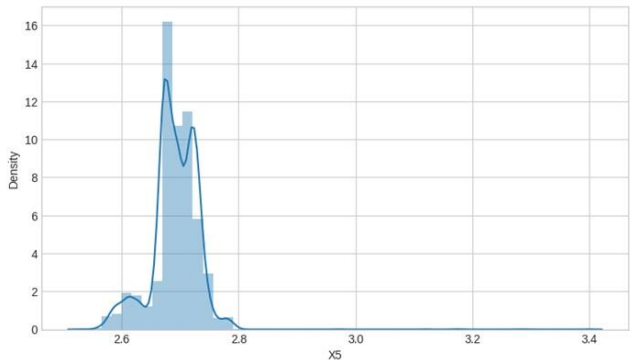


Source. Data analysis result of the research

(5) Particle density (X5): Reflects the density of the soil particles, excluding the pore spaces, ranging from 2.53 to 3.4.

Figure 5

Histogram of Particle Density (X5)

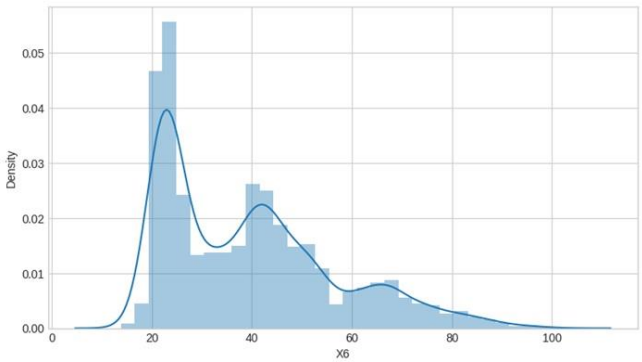


Source. Data analysis result of the research

(6) Liquid limit (X6): The moisture content at which soil changes from a plastic to a liquid state, varying from 13.8% to 102.59%.

Figure 6

Histogram of Liquid Limit (X6)

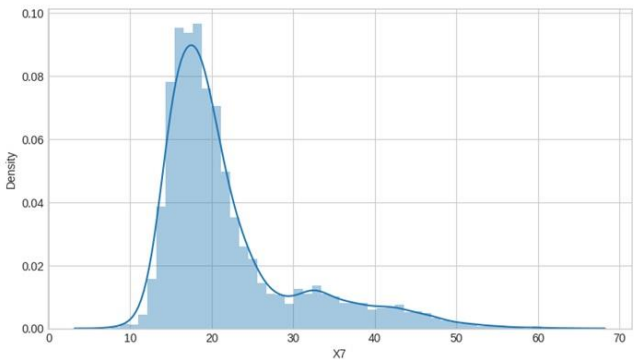


Source. Data analysis result of the research

(7) Plastic limit (X7): The moisture content at which soil changes from semi-solid to plastic, ranging from 7.6% to 63.8%.

Figure 7

Histogram of Plastic Limit (X7)



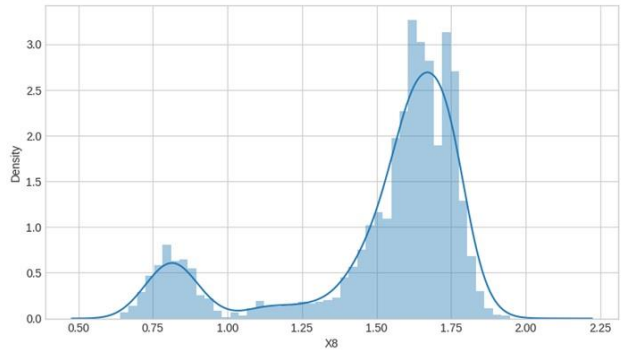
Source. Data analysis result of the research

Below are the indirectly obtained features.

(8) Dry unit weight (X8): Calculated from the bulk unit weight and moisture content, this indicates the weight of solids per unit volume, essential for compaction and stability analyses.

Figure 8

Histogram of Dry Unit Weight (X8)

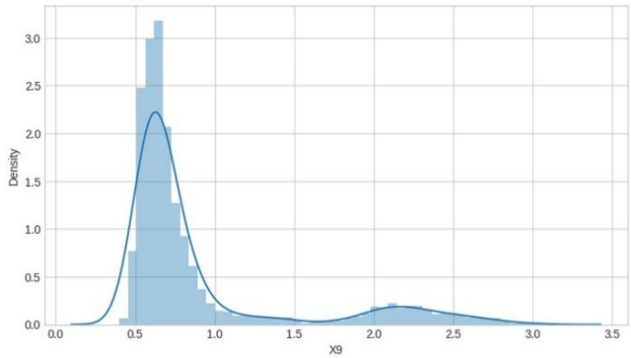


Source. Data analysis result of the research

(9) Void ratio (X9): The volume of voids to the volume of solids in the soil, indicating porosity and permeability.

Figure 9

Histogram of Void Ratio (X9)

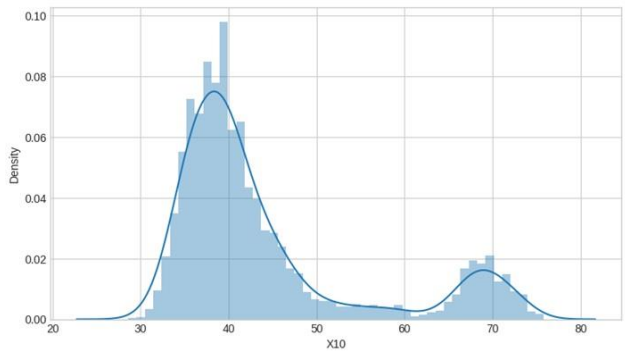


Source. Data analysis result of the research

(10) Porosity (X10): The percentage of the soil volume that is voids, affecting water and air movement through the soil.

Figure 10

Histogram of Porosity (X10)

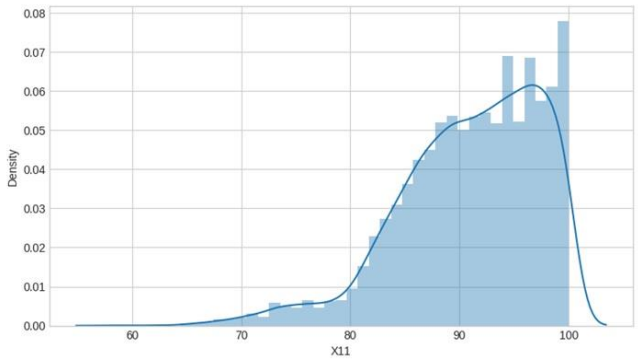


Source. Data analysis result of the research

(11) Degree of saturation (X11): The ratio of the volume of water to the volume of voids in the soil, indicating how saturated the soil is with water.

Figure 11

Histogram of Degree of Saturation (X11)

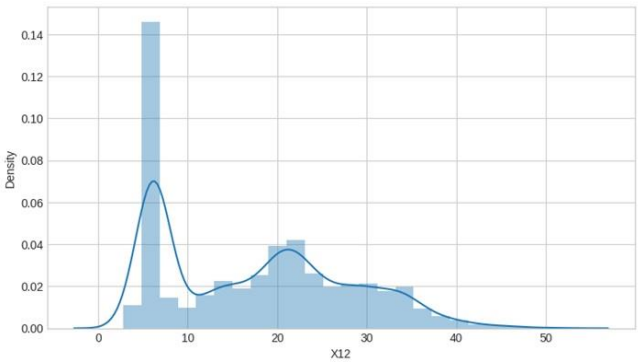


Source. Data analysis result of the research

(12) Plasticity index (X12): Calculated from the liquid and plastic limits, it measures the moisture content range over which the soil remains plastic.

Figure 12

Histogram of Plasticity Index (X12)

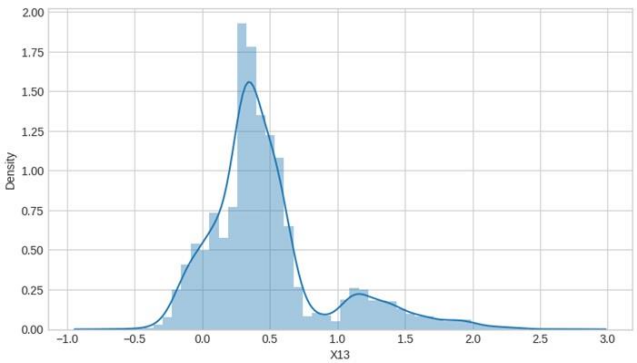


Source. Data analysis result of the research

(13) Liquidity index (X13): This index provides insight into the soil's current state relative to its liquid and plastic limits.

Figure 13

Histogram of Liquidity Index (X13)

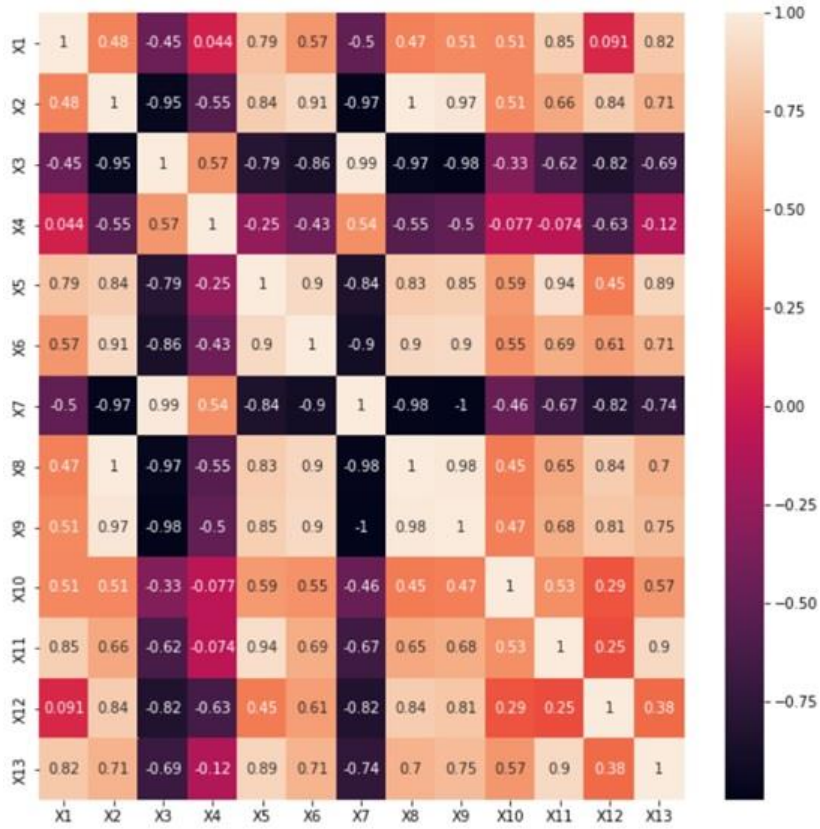


Source. Data analysis result of the research

The correlation among these features is significant in understanding soil behavior and properties. For instance, the clay content (X2), moisture content (X3), liquid limit (X6), and plastic limit (X7) are closely related, providing insights into the soil's consistency and susceptibility to deformation (Figure 14). Similarly, the dry unit weight (X8) and void ratio (X9) are interrelated, reflecting the soil's density and compaction levels (Figure 14). These correlations are essential for constructing predictive models in soil classification, enabling engineers to effectively infer soil behavior and suitability for construction projects.

Figure 14

Correlation of Input Features



Source. Data analysis result of the research

2.2. Data analysis and classification model

The study follows a structured approach to build accurate soil classification models. The process consists of five main steps:

(1) Step 1: Data collection

The initial phase involves collecting a dataset comprising 5,869 soil samples from soil investigation reports of 39 construction projects in various Ho Chi Minh City districts. This dataset is pivotal as it forms the foundation for the entire classification model. The collection process focused on gathering data that accurately represents the diversity of soil types within the targeted urban area, emphasizing cohesive soils like clay, silty clay, and silty sand.

(2) Step 2: Data preprocessing

Upon collecting the dataset, data preprocessing becomes the subsequent critical step. This phase begins with a statistical description, which involves a descriptive statistical

analysis of all 13 input features to understand their distribution, mean, standard deviation, and range. This step is vital to grasp the overarching characteristics of the dataset. Following this, outlier removal is conducted to eliminate anomalies from the dataset. This is crucial to prevent skewed results and enhance the model's accuracy, as outliers may result from data collection errors or reflect rare soil conditions that are not the focus of this study. Finally, normalization is applied to standardize the input features to a standard scale, typically within the range of [0, 1]. This process ensures the machine learning models function optimally by preventing any single feature from disproportionately influencing the outcome due to its scale, thereby maintaining the integrity of the model's predictive capability.

(3) Step 3: Construction of classification models

During the construction of soil classification models, this research leverages MATLAB's computational capabilities and flexibility to implement two core machine learning algorithms: K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). These algorithms are selected for their unique and complementary approaches to handling classification tasks.

K-Nearest Neighbor (KNN): The K-Nearest Neighbor algorithm operates on a simple yet effective principle: it assigns a class to each soil sample based on the predominant class among its 'k' nearest neighbors in the feature space. The determination of 'k,' the number of neighbors, and selecting the distance metric (commonly the Euclidean distance) are pivotal in optimizing the KNN model's performance. For our study, the MATLAB implementation of KNN is employed with default parameters, providing a straightforward yet powerful means of classifying soil samples. The simplicity of KNN, combined with its reliance on feature space proximity for classification, makes it exceptionally suitable for analyzing the complex, multi-dimensional data characteristic of soil properties. In the context of our research, KNN offers the advantage of intuitive understanding and implementation, allowing for the nuanced relationships between different soil features to be effectively captured and utilized for classification.

The Support Vector Machine (SVM) takes a more structured approach by constructing a hyperplane (or a set of hyperplanes in higher-dimensional space) that best separates the soil samples into their respective categories. The core objective of SVM is to maximize the margin between the closest points of the classes (support vectors) and the separating hyperplane, thereby enhancing the model's generalizability and predictive capability. Utilizing MATLAB's default SVM implementation, our research leverages this algorithm's robustness in dealing with complex feature spaces and its capacity to identify the optimal boundary between differing soil types. This method is particularly effective in scenarios where the decision boundary is not immediately apparent or when the dataset contains a high-dimensional feature space, as is often the case with soil classification tasks.

By utilizing these algorithms within MATLAB's robust computational environment, this study gains the ease of model construction and evaluation and leverages the platform's extensive library and machine learning tools. The use of default parameters for both KNN and SVM in MATLAB ensures a consistent and reproducible framework for model comparison, allowing for a focused investigation into the algorithms' effectiveness in classifying soils based on their intrinsic properties as defined by the Vietnamese Standard TCVN 9362:2012 (Cong thong tin dien tu Bo Xay dung, 2012). This method simplifies the development process, allowing for a direct evaluation of the algorithms' suitability for the intricate demands of soil classification in geotechnical engineering.

(4) Step 4: Model evaluation and optimization

This step is pivotal in the research process, systematically analyzing and optimizing the constructed machine-learning models for soil classification. It consists in executing three specific problem-solving tasks aimed at refining the models and evaluating their performance:

Problem 1 - Training set size analysis: The initial challenge in optimizing the machine learning models for soil classification lies in determining the ideal training set size. Problem 1 methodically increased the training set size from a mere 2% to 98% of the total dataset in increments of 2%. This comprehensive exploration spanned three scenarios: the first utilized all 13 input features for model training, the second used only the 07 features directly obtained from experiments, and the third employed the six features that were derived indirectly. The goal was to discern how varying amounts of training data and different feature combinations could influence the models' accuracy. This critical step guided the appropriate training set size for more detailed analyses of the subsequent problems.

Problem 2 - Feature combination analysis: With an optimal training set size established from Problem 1, Problem 2 sought to delve deeper into the impact of feature combinations on model accuracy. This problem was divided into two sub-problems, each concentrating on a distinct set of features: Directly obtained from experiments and those derived indirectly. The analysis proceeded under four distinct groups (groups 1 and 2, featuring either X6 or X7; groups 3, which includes both X6 and X7; group 4, lacking both X6 and X7), categorized by the presence or absence of key features - the liquid limit and the plastic limit - known to be crucial for cohesive soil classification. Sub-problem 1 combined various direct features to assess their individual and collective influence on accuracy. Sub-problem 2, on the other hand, evaluated the utility of indirect features in isolation, determining their contribution to the model's predictive power without relying on direct experimental data.

Problem 3 - Comprehensive feature analysis: The investigation culminated in Problem 3, which built upon the high-performing scenarios identified in Problem 2. The focus here was on the potential of integrating additional features with those already proven effective. This comprehensive analysis was conducted to ascertain whether including previously omitted features could enhance model performance. The approach involved a detailed examination of various feature combinations, concentrating on previously demonstrated promising results. This process aimed to identify the most potent feature set for soil classification, marking a pivotal step towards developing highly accurate machine learning models for geotechnical applications.

(5) Step 5: Output analysis

The final step involves analyzing the classification results against the predefined labels based on TCVN 9362:2012 (Cong thong tin dien tu Bo Xay dung, 2012) standards for cohesive soils. The performance of each model is evaluated based on its accuracy in correctly classifying the soil samples into one of the nine categories. The research aims to identify the most effective model and feature combination for soil classification, contributing valuable insights for geotechnical engineering applications in construction.

2.3. Evaluation metrics

Throughout the evaluation and optimization phases of this study, the performance of the machine learning models for soil classification is measured using accuracy as the principal metric, as outlined by Equal (1):

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (1)$$

where *TP* represents true positives, *TN* stands for true negatives, *FP* denotes false positives, and *FN* signifies false negatives.

3. Results and discussion

This study applied machine learning techniques to classify soils for construction, assessing the performance of SVM and KNN models under various conditions. The investigation was organized into three main problems, each addressing different aspects of model performance based on training set size and feature combinations.

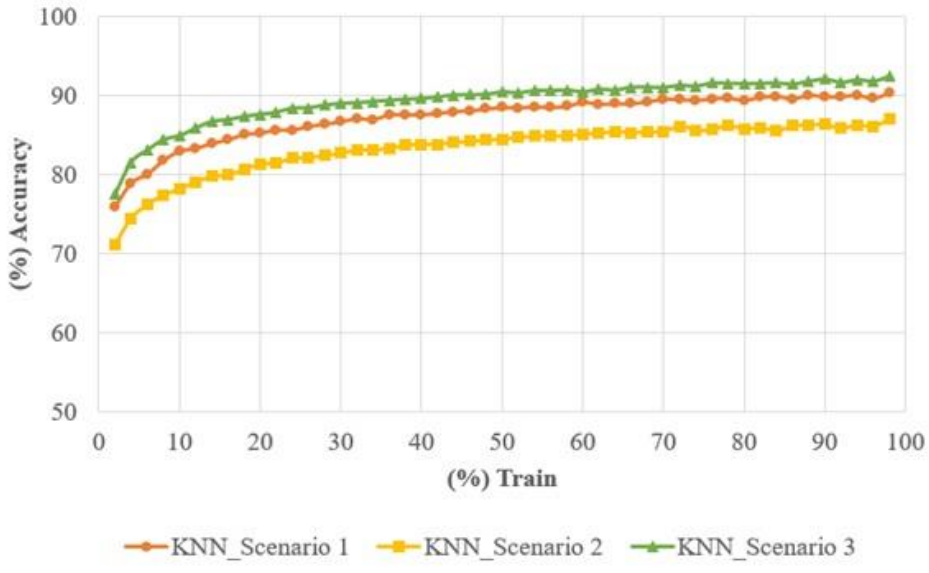
3.1. Problem 1

The analysis in Problem 1 explored the impact of training set size on the performance of SVM and KNN models for soil classification. This analysis aimed to identify the optimal training size for achieving high classification accuracy and understand how input feature choice impacts model performance. The training set sizes were incrementally increased from 2% to 98% of the total dataset, with 2% increments under three distinct scenarios involving different combinations of input features.

In the first scenario, all 13 input features (both directly and indirectly obtained) were used. The gradual increase in training set size demonstrated a consistent improvement in model accuracy for both KNN and SVM (Figure 15 and Figure 16). Notably, the KNN model exhibited significant gains in accuracy with increased training data, suggesting its sensitivity to the amount of training data. The SVM model also showed improved accuracy with larger training sets but to a lesser extent, indicating its robustness to varying training sizes. Both models reached a plateau in accuracy improvements beyond a 40% training set size, suggesting diminishing returns with further increases in training data.

Figure 15

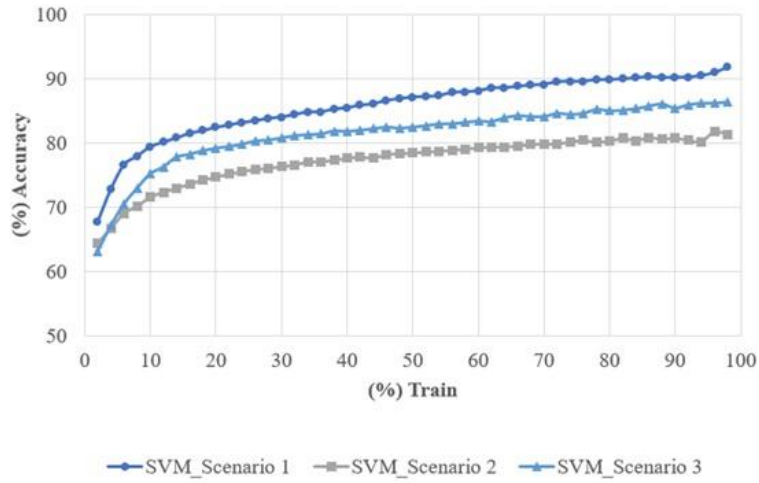
Performance of The KNN Model in 3 Scenarios



Source. Data analysis result of the research

Figure 16

Performance of The SVM Model in 3 Scenarios



Source. Data analysis result of the research

The second scenario focused on the 07 features obtained directly from laboratory experiments. This set of features includes fundamental soil properties such as sample depth, clay content, moisture content, bulk unit weight, particle density, and Atterberg limits. Under this scenario, the KNN model's performance slightly lagged compared to when all 13 features were used, indicating the importance of the additional indirect features for this algorithm (Figure 15). Conversely, the SVM model's accuracy showed less variability, maintaining a relatively consistent performance across different training sizes (Figure 16). This observation underscores the SVM model's capacity to leverage the directly obtained features effectively.

In the third scenario, the models utilized only the 06 features derived indirectly. Interestingly, this scenario revealed a notable increase in the KNN model's accuracy, surpassing its performance in the previous two scenarios (Figure 15). This suggests that the indirectly obtained features, which include parameters like dry unit weight, void ratio, porosity, degree of saturation, and indices related to soil's plasticity and liquidity, are particularly informative for the KNN model. While showing respectable accuracy, the SVM model did not surpass the performance observed in Scenario 1, indicating a balanced dependence on direct and indirect features (Figure 16).

The comprehensive analysis identified an optimal 40% training set size, balancing model accuracy and resource efficiency. This suggests that directly obtained features provide a solid baseline, while indirectly derived features can enhance KNN model accuracy, giving valuable insights into feature selection for soil classification.

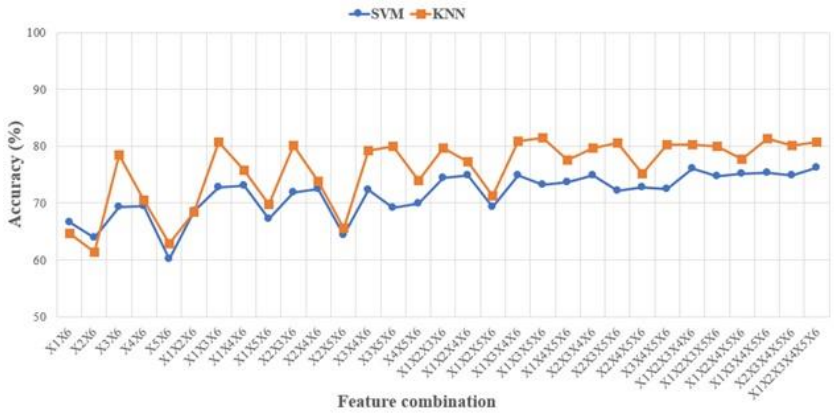
3.2. Problem 2

Problem 2 focused on feature combination effects on model accuracy, divided into two sub-problems. In sub-problem 1, the analysis tested the importance of the liquid limit (X6) and plastic limit (X7) features. Group 3, using both X6 and X7, achieved the highest classification accuracy, highlighting the critical role of these features (Figure 19). Conversely, Groups 1 and 2, which included either X6 or X7 alone, demonstrated only moderate classification accuracies (Figure 17 and Figure 18). This finding highlights the individual value of the liquid and plastic limits, but it also points to the fact that their combined presence

is far more impactful for classification accuracy. Meanwhile, Group 4, which excluded both X6 and X7, had the lowest accuracy (Figure 20), underlining the indispensable nature of these features for precise soil classification.

Figure 17

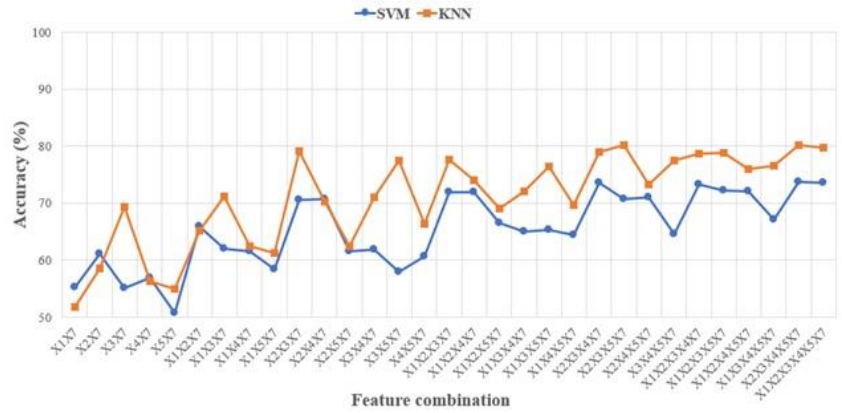
Performance of SVM and KNN Models of Group 1, which included X6



Source. Data analysis result of the research

Figure 18

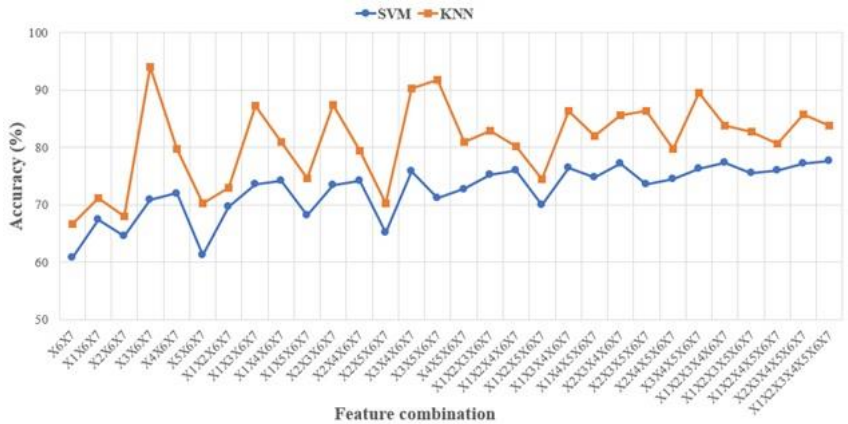
Performance of SVM and KNN Models of Group 2, which included X7



Source. Data analysis result of the research

Figure 19

Performance of SVM and KNN Models of Group 3, which included Both X6 and X7



Source. Data analysis result of the research

Figure 20

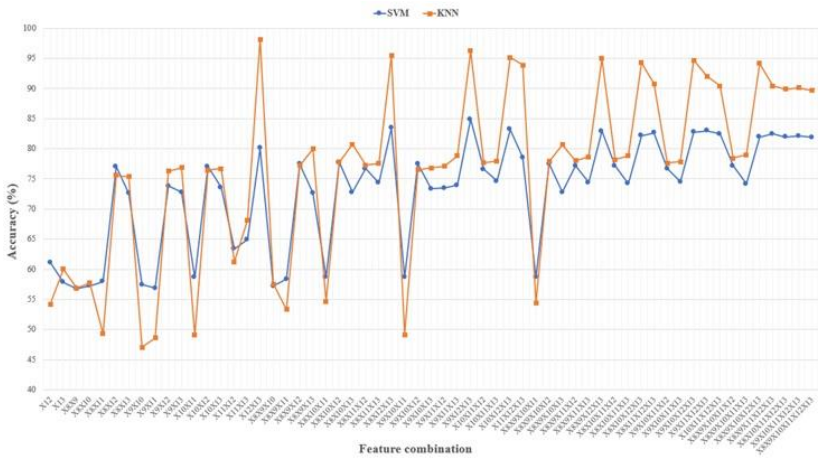
Performance of SVM and KNN Models of Group 4, which excluded Both X6 and X7

Source. Data analysis result of the research

The sub-problem 2 examined the impact of indirect features on model performance. Certain combinations, especially those incorporating features like the plasticity index (X12) and liquidity index (X13), led to significant improvements in classification accuracy (Figure 21). These results, depicted in the performance figures, suggest that features representing the soil's physical structure and its relationship with water content, such as dry unit weight, void ratio, and indices related to soil plasticity and liquidity, have a substantial influence over the model's ability to classify soils accurately. The analysis provided in this sub-problem marks the indirect features as critical enhancers of the machine learning model's performance in soil classification tasks.

Figure 21

Performance of SVM and KNN Models of Indirectly Obtained Feature Combinations



Source. Data analysis result of the research

Overall, Problem 2 confirmed that combining directly and indirectly obtained features is essential for accurate soil classification, particularly soil plasticity and liquidity features. These findings guide future research towards optimized feature selection in machine learning applications for geotechnical engineering.

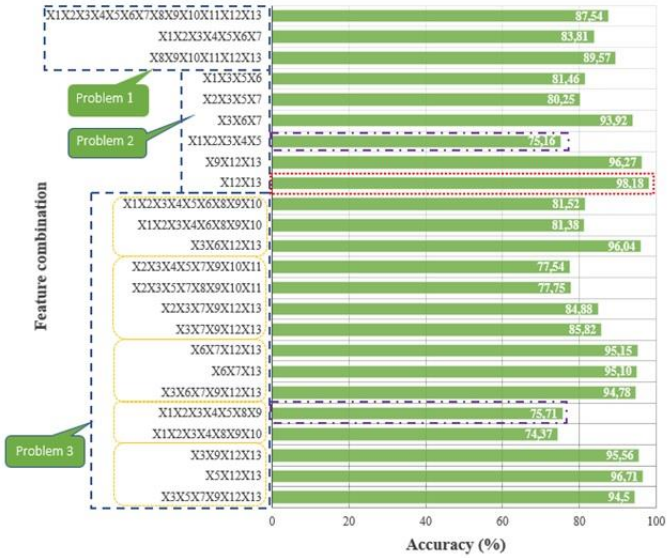
3.3. Problem 3

In Problem 3, the KNN model achieved its highest recorded accuracy of 96.71% using features such as particle density (X5), plasticity index (X12), and liquidity index (X13)

(Figure 22). While this did not exceed the peak accuracy in Problem 2, it surpassed Problem 1 results, affirming the value of specific feature combinations. The SVM model showed a modest enhancement, reaching an accuracy of 85.83% when leveraging a broad array of features, including moisture content (X3), particle density (X5), plastic limit (X7), void ratio (X9), plasticity index (X12), and liquidity index (X13), indicating a slight accuracy advantage over the previous problems (Figure 23). The findings also revealed the impact of feature presence, where the KNN model's performance was comparable to that of Problem 2, even without the liquid and plastic limits. In contrast, the SVM model improved slightly with additional features such as dry unit weight (X8), void ratio (X9), porosity (X10), and water saturation degree (X11), highlighting the nuanced influence that different features have on the predictive capability of soil classification models.

Figure 22

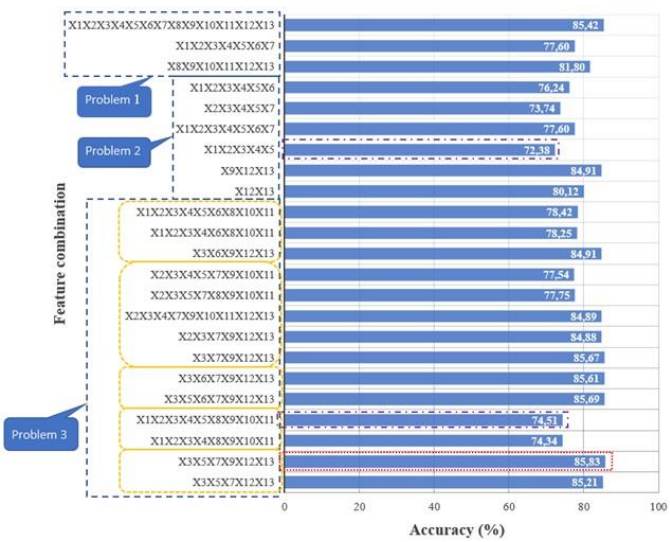
The Highest Performance of The KNN Model of Problem 1, Problem 2, and Problem 3



Source. Data analysis result of the research

Figure 23

The Highest Performance of The SVM Model of Problem 1, Problem 2, and Problem 3



Source. Data analysis result of the research

The findings reveal that KNN prefers certain feature combinations (e.g., particle density, plasticity, liquidity indices). At the same time, SVM benefits from a broader feature set, including dry unit weight, void ratio, and degree of saturation. These differences underscore the need for tailored feature selection to maximize model performance in soil classification.

This study shares similarities and distinctions with prior works, such as Pham, Nguyen, et al. (2021) and Nguyen et al. (2022). Our study utilizes a comprehensive dataset comprising 5,869 soil samples from 39 construction projects in Ho Chi Minh City, Vietnam. In contrast, Pham, Nguyen, et al. (2021) used a smaller dataset with 440 samples, focusing on a more limited scope of geotechnical projects. Nguyen et al. (2022) utilized a larger dataset of 4,888 soil samples, similar in scale to our study. The extensive dataset in our research allows for a more robust analysis and potentially more generalized findings. A significant difference lies in the soil classification standards used. Our study employs the Vietnamese Standard TCVN 9362:2012 - Cong thong tin dien tu Bo Xay dung (2012), which is tailored to local construction practices and regulations. In comparison, Pham, Nguyen, et al. (2021) and Nguyen et al. (2022) used the United Soil Classification System (USCS) standards. This distinction is crucial as it highlights the practical applicability of our research to Vietnamese construction projects, ensuring that the classification methodology aligns with local regulatory requirements and construction practices.

Our study employed KNN and SVM for algorithms due to their simplicity and baseline effectiveness. Pham, Nguyen, et al. (2021) employed Adaboost, Tree, and ANN models, focusing on ensemble learning techniques to enhance classification accuracy. Nguyen et al. (2022) used Support Vector Classification (SVC), Multilayer Perceptron (MLP), and Random Forest (RF) models, providing a diverse approach with advanced machine learning methods. The differences in algorithm choices reflect various strategies to tackle soil classification problems, with each study contributing unique insights into model performance and applicability. In terms of accuracy, our KNN model achieved a maximum accuracy of 96.71%, while the SVM model reached 85.83%. Pham, Nguyen, et al. (2021) reported high accuracy with their Adaboost model, misclassifying only 11 out of 88 data points in their subset, suggesting robust performance. Nguyen et al. (2022) achieved an impressive average accuracy score of 0.968 across all models, with the SVC model achieving the highest accuracy of 0.984. These results indicate that while our models perform well, there is potential for further improvement by exploring additional algorithms and hyperparameter tuning.

4. Conclusion

This research comprehensively analyzes machine learning techniques - specifically, KNN and SVM models - for classifying construction soils in Ho Chi Minh City, Vietnam. Through a structured methodology encompassing data collection, preprocessing, and detailed analysis across three main problems, the study explores the impact of training set sizes and the combination of directly and indirectly derived soil features on model performance. The findings highlight several key insights:

• The selection and combination of features significantly influence the accuracy of soil classification. Particularly, the inclusion of liquid limit (X6) and plastic limit (X7) features, along with their derived indices (plasticity index X12 and liquidity index X13), are crucial for enhancing model performance.

• KNN generally outperforms SVM in scenarios with carefully selected feature sets, indicating its suitability for soil classification tasks where nuanced feature relationships play a vital role.

• The optimal training set size was identified as 40% of the total dataset, beyond which the improvements in model accuracy diminish.

The research underscores the potential of machine learning in revolutionizing soil classification within geotechnical engineering, offering insights into the predictive capabilities of these models when equipped with appropriate soil parameters.

ACKNOWLEDGEMENTS

Ho Chi Minh City Open University funds this research under the grant number E2022.03.02.

NO CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflict of interest.

References

Armaghani, D. J., Hajihassani, M., Yazdani, B. B., Marto, A., & Tonnizam, M. E. (2014). Indirect measure of shale shear strength parameters by means of rock index tests through an optimized artificial neural network. *Measurement, 55*, 87-98.

Cal, Y. (1995). Soil classification by neural network. *Advances in Engineering Software, 22*(2), 95-97.

Carvalho, L. O., & Ribeiro, D. B. (2019). Soil classification system from cone penetration test data applying distance-based machine learning algorithms. *Soils and Rocks, 42*(2), 167-178.

Casagrande, A. (1948). Classification and identification of soils. *Transactions of the American Society of Civil Engineers, 113*(1), 901-991.

Cong thong tin dien tu Bo Xay dung. (2012). *TCVN 9362:2012 specifications for design of foundation for buildings and structures*. https://moc.gov.vn/tl/tin-tuc/53185/tieu-chuan-moi-tieu-chuan-thiet-ke-nen-nha-va-cong-trinh-specifications-for-design-of-foundation-for-buildings-and-structures-tcvn-9362-2012.aspx

Das, B. M., & Sobhan, K. (2013). *Principles of geotechnical engineering*. Cengage Learning.

Gambill, D. R., Wall, W. A., Fulton, A. J., & Howard, H. R. (2016). Predicting USCS soil classification from soil property variables using random forest. *Journal of Terramechanics, 65*, 85-92.

Ghaleini, E. N., Koopialipoor, M., Momenzadeh, M., Sarafraz, M. E., Mohamad, E. T., & Gordan, B. (2018). A combination of artificial bee colony and neural network for approximating the safety factor of retaining walls. *Engineering with Computers, 35*, 647-658.

Goktepe, F., Arman, H., & Pala, M. (2010). A new approach for classification of clayey soil: A case study for Adapazari region Turkey. *Scientific Research and Essays, 5*(15), 2037-2043.

Gordan, B., Koopialipoor, M., Clementking, A., Tootoonchi, H., & Mohamad, E. T. (2019). Estimating and optimizing safety factors of retaining wall through neural network and bee colony techniques. *Engineering with Computers, 35*, 945-954.

Kang, T. H., Choi, S. W., Lee, C., & Chang, S. H. (2022). Soil classification by machine learning using a tunnel boring machine's operating parameters. *Applied Sciences, 12*(22), Article 11480.

Karimpouli, S., & Tahmasebi, P. (2019). Image-based velocity estimation of rock using convolutional neural networks. *Neural Networks, 111*, 89-97.

Koopialipoor, M., Fahimifar, A., Ghaleini, E. N., Momenzadeh, M., & Armaghani, D. J. (2020). Development of a new hybrid ANN for solving a geotechnical problem related to tunnel boring machine performance. *Engineering with Computers, 36*, 345-357.

Koopialipoor, M., Murlidhar, B. R., Hedayat, A., Armaghani, D. J., Gordan, B., & Mohamad, E. T. (2020). The use of new intelligent techniques in designing retaining walls. *Engineering with Computers, 36*, 283-294.

Kovačević, M., Bajat, B., & Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines. *Geoderma, 154*(3/4), 340-347.

Ma, W. T. (2005). Application of support vector machine to classification of expansive soils. *Rock and Soil Mechanics, 26*(11), 1790-1792.

Mollahasani, A., Alavi, A. H., Gandomi, A. H., & Bazaz, J. B. (2011). A new prediction model for soil deformation modulus basedon PLT results. *Proceedings of the 9th International Symposium on Computational Civil Engineering, New Approaches in Numerical Analysis in Civil Engineering* (pp. 53-61). Romania.

Momeni, E., Dowlatshahi, M. B., Omidinasab, F., Maizir, H., & Armaghani, D. J. (2020). Gaussian process regression technique to estimate the pile bearing capacity. *Arabian Journal for Science and Engineering, 45*, 8255-67.

Nguyen, M. D., Pham, T. B., Ho, L. S., Ly, B. H., Le, T. T., Chongchong, Q., Le, V. M., Le, M. L., Indra, P., Le, S. H., & Bui, D. T. (2020). Soft-computing techniques for prediction of soils consolidation coefficient. *CATENA, 195*, Article 104802.

Nguyen, M. D., Romulus, C., Ho, A. S., Hassan, A., Le, H. V., Indra, P., & Pham, T. B. (2022). Novel approach for soil classification using machine learning methods. *Bulletin of Engineering Geology and the Environment, 81*, Article 468.

Ninić, J., Freitag, S., & Meschke, G. (2017). A hybrid finite element and surrogate modelling approach for simulation and monitoring supported TBM steering. *Tunnelling and Underground Space Technology, 63*, 12-28.

Pham, T. B., Mahdis, A., Nguyen, M. D., Ngo, T. Q., Nguyen, K. T., Tran, H. T., Vu, H., Bui, A. T. Q., Le, H. V., & Indra, P. (2021). Estimation of shear strength parameters of soil using optimized inference intelligence system. *Vietnam Journal of Earth Sciences, 43*(2), 189-198.

Pham, T. B., Nguyen, D. D., Bui, A. T. Q., Nguyen, M. D., Vu, T. T., & Indra, P. (2022). Estimation of load-bearing capacity of bored piles using machine learning models. *Vietnam Journal of Earth Sciences, 44*(4), 470-480.

Pham, T. B., Nguyen, M. D., Nguyen, T. T., Ho, L. S., Mohammadreza, K., Nguyen, Q. K., Danial, J., & Le, H. V. (2021). A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling. *Transportation Geotechnics, 27*, Article 100508.

Pham, T. B., Pradhan, B., Bui, D. T., Prakash, I., & Dholakia, M. B. (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environmental Modelling & Software, 84*, 240-250.

Pham, T. B., Singh, S. K., & Ly, B. H. (2020). Using Artificial Neural Network (ANN) for prediction of soil coefficient of consolidation. *Vietnam Journal of Earth Sciences, 42*(4), 311-319.

Shirzadi, A., Himan, S., Kamran, C., Bui, D. T., Pham, B. T., Kaka, S., & Baharin, B. A. (2017). A comparative study between popular statistical and machine learning methods for simulating volume of landslides. *Catena, 157*, 213-226.

Singh, G., & Walia, B. S. (2017). Performance evaluation of nature-inspired algorithms for the design of bored pile foundation by artificial neural networks. *Neural Computing and Applications, 28*(Suppl 1), 289-298.

Tran, H. T., Nguyen, P. B., & Tran, D. T. (2024). Machine learning applications in pile load capacity prediction: Advanced analysis of pile driving forces and depths in urban Ho Chi Minh City construction sites. *Indian Geotechnical Journal.* https://doi.org/10.1007/s40098-024-01055-9

Wang, H., Zhang, L., Yin, K., Luo, H., & Li, J. (2021). Landslide identification using machine learning. *Geosci Front, 12*(1), 351-364.

Xiao, L., Zhang, Y., & Peng, G. (2018). landslide susceptibility assessment using integrated deep learning algorithm along the China-Nepal highway. *Sensors, 18*(12), Article 4436.

Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., & Ding, X. (2021). Application of deep learning algorithms in geotechnical engineering: A short critical review. *Artificial Intelligence Review, 54*, 5633-5673.

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geosci Front, 12*(1), 469-477.

Zhou, C., Ouyang, J., Ming, W., Zhang, G., Du, Z., & Liu, Z. (2019). A stratigraphic prediction method based on machine learning. *Applied Sciences, 9*(17), Article 3553.

Zhou, J., Qiu, Y., Zhu, S., Armaghani, D. J., Li, C., Hoang, N., & Saffet, Y. (2021). Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. *Engineering Applications of Artificial Intelligence, 97*, Article 104015.