Article Type: Research

OPEN ACCESS

Development of a Machine Learning-based Model for Localized CPT Soil Classification

Amin Danesh[1,2], Sibel Açık[2], Sun Jing[3], Mohamadreza Esmailzadeh[1,*]

[1]Department of Civil Engineering, University College of Science and Technology, Urmia 5735133746, Iran
[2] Department of Engineering Science and Architecture, Istanbul Arel University, Istanbul 34537, Türkiye
[3]Civil Engineering Department, Guangxi Polytechnic of Construction, Nanning 530003, China
*Corresponding author: mohamadreza_esmailzadehh@gmail.com

Received: 11 February 2025 / Accepted: 10 April 2025 / Published: 25 April 2025

***Abstract*:** Cone Penetration Tests (CPT) offer valuable insights into the physical properties of soils that cannot be directly obtained from recovered soil samples. This information is crucial for understanding the geotechnical characteristics of soils, including their layering, thickness, stiffness, strength, and consolidation behavior. Practically, CPT is conducted to depths of up to 10m, depending on project requirements. The results are interpreted to differentiate between granular and cohesive soils by analyzing cone resistance and shaft friction measurements. These interpretations are then applied to classify soils using established empirical classification charts or tables, ensuring accurate and reliable geotechnical evaluations. The presented study aimed to develop a machine learning-based model utilizing the Support Vector Machine (SVM) algorithm to classify soil types based on localized CPT data. The approach leverages the high-resolution data provided by CPT, including cone tip resistance, shaft friction, and, where available, pore pressure measurements, to accurately identify and classify soil behavior. By tailoring the model to local soil conditions, the study seeks to enhance the reliability and precision of soil classification, overcoming limitations of traditional empirical methods. This research not only demonstrates the potential of SVM in geotechnical applications but also provides a framework for integrating advanced computational techniques with localized geotechnical datasets for improved decision-making in soil characterization and infrastructure design.

Keywords: CPT classification, Machine learning, Localized analysis, Geotechnical engineering, Soil behavior.

I. INTRODUCTION

Cone Penetration Tests (CPT) are a widely utilized in-situ testing method in geotechnical engineering, designed to determine the properties of subsurface soils (Huang and Ma, 1994). The CPT involves the penetration of a cone-shaped probe into the ground at a constant rate, typically 2 cm/s, while measuring resistance at the cone tip (q_c) and along the shaft (f_s). Some advanced systems also measure pore water pressure (u^2) during penetration, known as CPT_u. The test is valued for its efficiency, continuous profiling capabilities, and minimal disturbance to the soil, making it a preferred method for many geotechnical investigations (Meigh, 2013).

The standard CPT apparatus includes a cone penetrometer mounted on a thrust rig, capable of generating sufficient force to push the cone into the ground (Robertson, 2009). The cone typically has a base area of 10 cm² and a tip angle of 60°. Measurements are recorded at intervals of 1-2 cm, providing high-resolution data about soil stratigraphy (Meigh, 2013). The testing process is automated, with electronic sensors transmitting data in real-time. This ensures accuracy and enables rapid interpretation. The equipment's portability allows its use in a wide range of environments, from soft sediments to stiff clays (Teh & Houlsby, 1991) and loose sands (Huang & Ma, 1994). One of the primary applications of CPT is soil profiling. By analyzing the q_c and f_s, geotechnical engineers can distinguish between different soil types, such as sands, silts, and clays. Charts like the Robertson & Campanella (1983) soil classification system provide guidelines to interpret CPT data into stratigraphy. This continuous profiling is particularly useful in detecting thin layers and transitional zones that may not be identified through traditional boring methods (Lunne et al., 2002).

CPT data can be correlated with various soil properties, including relative density, shear strength, and consolidation characteristics (Meigh, 2013). For instance, in sands, cone resistance is often used to estimate relative density and angle of internal friction (Huang & Ma, 1994). In clays, the undrained shear strength can be derived from empirical correlations with cone resistance and pore pressure data (Teh & Houlsby, 1991). These correlations are crucial for designing foundations, slopes, and retaining structures, ensuring safety and stability (Walker & Yu, 2010). In fact, CPT results play a pivotal role in foundation design by providing data on bearing capacity and settlement potential (Senneset & Janbu, 1985). For shallow foundations,

cone resistance helps estimate allowable bearing pressure, while for deep foundations, it aids in selecting suitable pile lengths and types (Robertson, 2016). Additionally, CPT is instrumental in site characterization for infrastructure projects such as highways, railways, and tunnels. Its ability to quickly gather data over large areas makes it ideal for preliminary investigations (Ahmadi & Dariani, 2017). While CPT offers numerous advantages, it has some limitations. It is less effective in highly gravelly soils or when large boulders are present, as these materials can damage the equipment (Butlanska et al., 2014). Additionally, the interpretation of CPT data requires expertise and reliance on empirical correlations (Huang & Hsu, 2005). However, advances such as machine learning have significantly enhanced CPT applications. By integrating CPT data with computational algorithms, engineers can achieve more accurate and site-specific predictions of soil behavior. This technological synergy is expanding the utility of CPT in modern geotechnical practice.

One of the major advantages of CPT in soil classification is its ability to deliver continuous subsurface profiles (Meigh, 2013). Unlike traditional borehole sampling, which provides data only at discrete intervals (Meigh, 2013), CPT offers a near-continuous record of soil resistance and stratification (Lunne et al., 2002). This allows engineers to detect thin layers (Robertson, 2009), transitional zones, or interbedded soil formations (Walker & Yu, 2010) that could significantly influence design considerations (Robertson, 2016). CPT data is commonly interpreted using soil behavior type charts, such as the Robertson & Campanella (1983) or Robertson (2016) classifications. Figure 2 is providing the soil behaviors type chart. These charts plot cone resistance and friction ratio to categorize soils into different behavior types, ranging from dense sands to soft clays. Such classification systems help engineers quickly identify soil conditions and make preliminary assessments for project feasibility and risk analysis (Bol, 2023). In addition to classifying soil behavior, CPT data is used to estimate various engineering properties. For example, cone resistance correlates with relative density in sands and undrained shear strength in clays. These correlations are essential for evaluating load-bearing capacity and settlement potential (Bilgin et al., 2019). Empirical models based on CPT data often serve as reliable predictors of soil behavior under structural loads, reducing reliance on extensive laboratory testing (Agaiby & Mayne, 2019).

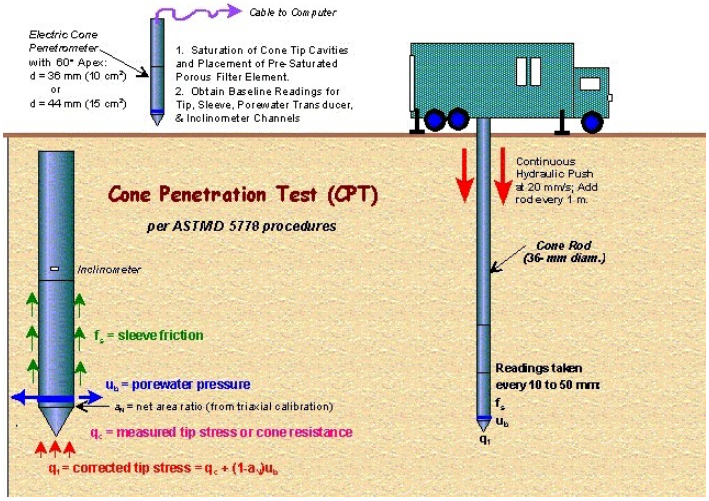

Fig. 1 General concept of CPT testing (Patrick, 2008)

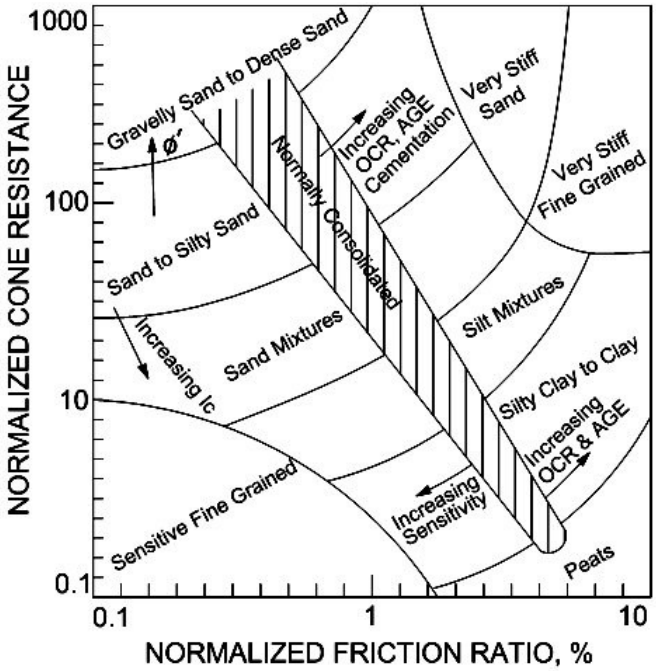

Fig. 2 CPT chart for soil behaviors type (Robertson & Campanella, 1983; Khan et al., 2011)

CPT's precision is particularly valuable in identifying soil transitions and stratigraphic boundaries. For example, the test can detect the interface between soft clay and dense sand layers, which is critical for foundation design (Meigh, 2013). Such transitions might not be captured through traditional methods like drilling or sampling (Bol, 2023). This capability ensures a more comprehensive understanding of the subsurface profile, minimizing unexpected challenges during construction (Teh & Houlsby, 1991). Indeed, CPT is a key tool for assessing soil liquefaction potential, particularly in granular soils under seismic loads (Lunne et al., 2002). By analyzing cone resistance and pore pressure, engineers can identify soils prone to liquefaction during an earthquake (Moss, 2003; Moss et al., 2006). The ability to classify soils into liquefiable and non-liquefiable categories helps in mitigating seismic risks and designing earthquake-resistant structures (Ku et al., 2004).

Pore pressure measurements in CPT_u tests enhance soil classification by providing additional data on soil drainage characteristics (Maurer et al., 2015). This is particularly important in identifying over-consolidated clays, sensitive clays, and silty layers, where pore pressure responses vary significantly (Poor et al., 2023). Combining q_c, f_s, and u^2 enables a more refined and accurate classification of soils, especially in complex geological settings (Boulanger & Idriss, 2014). The integration of machine learning techniques with CPT data is revolutionizing soil classification (Ghanizadeh et al., 2023). Advanced algorithms analyze large datasets to uncover patterns and relationships that traditional methods might overlook (Lunne et al., 2002). Machine learning models can predict soil types and properties with higher accuracy, tailoring classifications to specific project sites. This approach is particularly effective in regions with unique soil conditions or where existing empirical correlations are inadequate (Bol, 2023). Indeed, CPT's application in soil specification and classification has

transformed geotechnical investigations, offering rapid, reliable, and detailed insights into subsurface conditions (Poor et al., 2023). Its ability to continuously profile soils, correlate with engineering properties (Robertson, 2016), and integrate advanced technologies like machine learning ensures its relevance in modern geotechnical engineering (Ghanizadeh et al., 2023). Whether used for identifying soil behavior, evaluating liquefaction potential, or understanding complex stratigraphy, CPT remains a cornerstone of soil classification and foundation design. Its adaptability and precision continue to make it an essential tool for addressing the challenges of construction and infrastructure development (Zhang et al., 2021).

II. Support Vector Machine and CPT Application

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification, regression, and outlier detection (Bhavsar & Panchal, 2012). It is based on the concept of finding an optimal decision boundary, or hyperplane that separates data points belonging to different classes in a feature space. The idea is to maximize the margin between the hyperplane and the nearest data points of each class, known as support vectors (Salcedo-Sanz et al., 2014). This maximization ensures that the classifier is robust and minimizes the likelihood of misclassification (Deka, 2014). SVM is particularly effective in high-dimensional spaces and excels at handling complex datasets with clear separations (Abdullah & Abdulazeez, 2021). The primary goal of SVM is to achieve structural risk minimization, which balances the model's complexity with its ability to generalize to unseen data (Bhavsar & Panchal, 2012). For linearly separable data, SVM identifies the hyperplane that maximizes the margin between classes. However, most real-world datasets are not linearly separable (Uncuoglu et al., 2022). To address this, SVM employs kernel functions that transform the data into a higher-dimensional space where a linear separation becomes possible (Cemiloglu et al., 2023). Commonly used kernels include the linear kernel for simple relationships, the polynomial kernel for complex interactions, and the radial basis function (RBF) kernel for non-linear separations. By leveraging these transformations, SVM provides a flexible framework for tackling a wide variety of classification problems (Abdullah & Abdulazeez, 2021).

Implementing SVM begins with data preprocessing, which involves normalizing or scaling features to ensure consistent measurement units. This step is crucial because SVM is sensitive to variations in feature magnitudes (Baghbani et al., 2022). Next, an appropriate kernel function is selected based on the nature of the data. For instance, linear kernels work well for linearly separable data, while RBF kernels are suitable for datasets with intricate non-linear patterns. Hyperparameters, such as the penalty term (C) and kernel-specific parameters (like gamma for the RBF kernel), are then optimized to balance the trade-off between model complexity and accuracy. Once the parameters are set, the SVM model is trained using labeled data to identify the optimal hyperplane or decision boundary (Salcedo-Sanz et al., 2014). After training, the model is evaluated on a validation dataset to test its generalization capabilities before being deployed on unseen data (Abdullah & Abdulazeez, 2021).

SVM offers several significant advantages, making it a popular choice in machine learning applications. One of its primary strengths is its high accuracy, particularly for classification tasks with well-separated classes (Deka, 2014). SVM is also robust when dealing with high-dimensional data, as it focuses on the most critical data points (i.e. the support vectors), reducing the impact of irrelevant features (Uncuoglu et al., 2022). This makes it highly effective for applications like text classification and image recognition (Bhavsar & Panchal, 2012). Additionally, SVM performs well even with small datasets, unlike many other machine learning algorithms that require large volumes of data to achieve reliability (Goh & Goh, 2007). Its flexibility through the use of kernel functions allows SVM to adapt to both linear and non-linear problems, making it a versatile tool across various domains (Ma et al., 2018). Despite its advantages, SVM has some limitations that must be considered (Bhavsar & Panchal, 2012). Training SVM models can be computationally expensive, especially with large datasets, as the algorithm requires solving a quadratic optimization problem (Abdullah & Abdulazeez, 2021). This computational intensity can become a bottleneck for applications involving millions of data points (Uncuoglu et al., 2022). Additionally, SVM's performance heavily depends on the careful selection of kernel functions and hyperparameters, which often requires domain expertise and extensive experimentation. The model's interpretability is another challenge, as the decision boundary in high-dimensional space is not easily understandable compared to simpler models like decision trees (Salcedo-Sanz et al., 2014). Furthermore, SVM struggles with imbalanced datasets, where the classes have unequal representation, as the hyperplane may favor the majority class (Deka, 2014).

SVM has been applied successfully in numerous fields due to its versatility and effectiveness. In geotechnical engineering, for example, it is used to classify soil types and predict soil properties based on in-situ test data (Deka, 2014). In bioinformatics, SVM is employed for tasks like gene classification and protein structure prediction (Bhavsar & Panchal, 2012). It has also found applications in image recognition, where it helps classify objects or patterns within images, and in finance, where it is used for stock price prediction and risk analysis. In the realm of cybersecurity, SVM aids in detecting anomalies and identifying potential threats, further demonstrating its broad applicability across diverse industries (Abdullah & Abdulazeez, 2021). SVM remains one of the most powerful algorithms in machine learning, particularly for tasks involving classification and regression (Cemiloglu et al., 2023). Its theoretical foundation, coupled with its ability to handle high-dimensional and complex datasets, ensures its continued relevance in modern applications (Baghbani et al., 2022). While its computational intensity and sensitivity to parameter tuning pose challenges, these can be mitigated with advancements in computing power and automated optimization techniques. SVM's adaptability through kernel functions and its focus on maximizing margins make it a reliable and precise tool for solving real-world problems (Bhavsar & Panchal, 2012). Despite emerging algorithms, SVM retains its place as a cornerstone in the machine learning toolkit, proving invaluable in scenarios where accuracy and reliability are paramount (Deka, 2014).

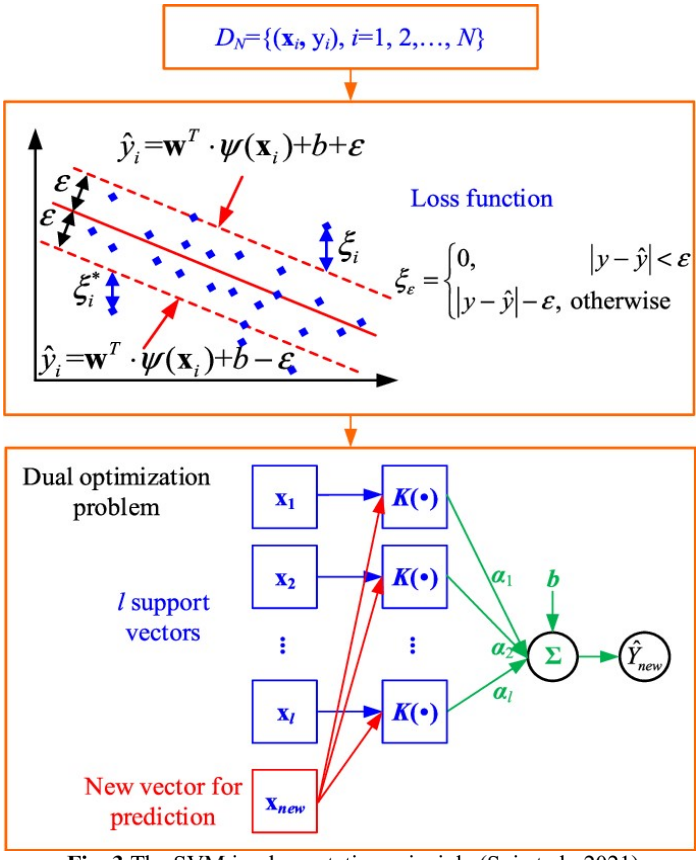

Fig. 3 The SVM implementation principle (Sui et al., 2021)

The application of SVM in CPT data analysis has revolutionized the way geotechnical engineers classify soils (Rauter & Tschuchnigg, 2021). Traditionally, soil classification relied on empirical charts and rules that mapped CPT parameters, such as cone tip resistance and sleeve friction, to predefined soil types. While effective, these methods often lack the flexibility to adapt to site-specific conditions (Ansary & Ansary, 2024). SVM, on the other hand, offers a data-driven approach that can identify complex relationships between CPT parameters and soil behavior. By training the algorithm with localized datasets, SVM provides more accurate and reliable soil classifications tailored to specific geotechnical contexts (Rauter & Tschuchnigg, 2021). SVM is increasingly employed to predict key soil mechanical properties, such as undrained shear strength, relative density, and stiffness, based on CPT data (Lee & Chern, 2012). The algorithm can process non-linear and multi-dimensional relationships among input variables, such as cone tip resistance and pore pressure, to deliver precise predictions (Kirts et al., 2019). This ability to derive meaningful insights from raw CPT measurements helps engineers make informed decisions about foundation design, slope stability, and earthworks. By integrating SVM into CPT data analysis workflows, engineers can enhance the accuracy of property estimations, reducing uncertainties in geotechnical design (Faraz Athar et al., 2023).

CPT data often exhibits continuous variations with depth, reflecting changes in soil stratigraphy (Zhang et al., 2022). SVM can be used to classify these variations into distinct layers, automating the stratigraphic interpretation process. Through supervised learning, the algorithm is trained on labeled datasets that include depth-wise soil classifications. Once trained, SVM can analyze new CPT profiles and detect transitions between soil layers with remarkable precision (Cho et al., 2023). This capability is particularly valuable in large-scale projects where manual interpretation of extensive CPT datasets is impractical (Rauter & Tschuchnigg, 2021). One of the key advantages of applying SVM to CPT data is the potential for automation and real-time analysis (Lee & Chern, 2012). Modern CPT equipment generates large volumes of high-resolution data, which can be challenging to process manually (Zhang et al., 2022). SVM algorithms can analyze this data on-the-fly, providing instant insights into soil classifications, mechanical properties, and stratigraphy (Faraz Athar et al., 2023). This real-time capability is particularly beneficial in dynamic field environments, where quick decision-making is essential, such as during the construction of deep foundations or embankments (Baghbani et al., 2022). Geotechnical datasets, including CPT measurements, often exhibit significant variability due to site-specific factors and inherent uncertainties in soil behavior. SVM is well-suited to handle such variability because of its robustness in finding optimal decision boundaries, even in noisy and overlapping datasets (Zhang et al., 2022). By incorporating advanced kernels and carefully tuning parameters, SVM can model complex relationships (Sui et al., 2021) between CPT data features and geotechnical outcomes (Rauter & Tschuchnigg, 2021). This makes it a powerful tool for improving the reliability of interpretations and reducing the likelihood of errors in geotechnical engineering projects.

III. Materials and Methods

The methodology for this study revolves around the integration of CPT data and machine learning, specifically the SVM algorithm, to develop an advanced soil classification model tailored to localized conditions. The first step in the methodology involves data collection from multiple CPT profiles, focusing on key parameters such as q_c, f_s, and, when available, u^2. These measurements are obtained from various depths at a site to capture the soil's vertical variability. The data is pre-processed to remove noise and outliers, ensuring that only accurate and relevant information are fed into the model. Additionally, normalization techniques are applied to standardize the input features, as SVM models are sensitive to the scale of the data. The database used in this analysis is compiled from a wide range of literature, incorporating 540 different data samples. Once the dataset is prepared, the next step is to classify the soil types using the SVM algorithm. The SVM model is trained using labeled datasets, where soil classes are known based on prior geological or geotechnical studies. These labels serve as the target output for the SVM, and the model learns to recognize the underlying patterns that differentiate soil types from the given CPT parameters. The training data includes various soil categories, such as sand, clay, silt, and other mixed soils, and the model learns to map the continuous CPT data to discrete soil classifications. The model's hyperparameters, including the kernel type (usually the radial basis function or polynomial kernel) and C parameter, are optimized using a grid search or cross-validation technique to enhance the model's performance. Table 1 is provided model's hyperparameters description.

Table 1 SVM model's hyperparameters used in this study

Hyperparameter	Description	Values
Kernel	Specifies the type of kernel function used to transform the input data into higher dimensions.	'linear', 'poly', 'rbf', 'sigmoid'
C (Regularization Parameter)	Controls the trade-off between maximizing the margin and minimizing classification errors.	A positive float value (e.g., 1, 10, 100)
Gamma	Defines how far the influence of a single training sample reaches. Only used for 'rbf', 'poly', and 'sigmoid' kernels.	'scale', 'auto', or a positive float value (e.g., 0.01, 0.1, 1)
Degree	The degree of the polynomial kernel function. Only used for 'poly' kernel.	A positive integer (e.g., 2, 3, 4)
Coef0	The independent term in kernel function. Only used for 'poly' and 'sigmoid' kernels.	A float value (e.g., 0.0, 1.0)
Nu	A parameter for nu-SVMs to control the fraction of margin errors.	A float value between 0 and 1 (e.g., 0.1, 0.5)
Shrinking	Whether to use the shrinking heuristic to speed up training.	True or False
Probability	Whether to enable probability estimates for classification.	True or False
Max Iterations	The maximum number of iterations for the solver.	A positive integer (e.g., 1000)

Table 2 SVM's advantage and disadvantage for CPT prediction

Advantages	Disadvantages
High Accuracy: SVM is highly effective in providing accurate classifications and predictions for CPT data, particularly when the data is well-separated.	Computational Complexity: Training SVM models can be computationally expensive, especially with large datasets, which may slow down real-time predictions.
Effective with High-Dimensional Data: SVM can efficiently handle high-dimensional data, making it suitable for CPT data with multiple parameters.	Sensitive to Parameter Tuning: SVM requires careful tuning of hyperparameters (e.g., kernel type, C, gamma) to achieve optimal performance, which can be time-consuming.
Robust to Overfitting: By maximizing the margin between classes, SVM tends to avoid overfitting, especially when using a well-chosen kernel and regularization.	Poor Performance with Noisy Data: SVM may struggle when the CPT data is noisy or contains a significant amount of outliers, which can negatively affect the model's accuracy.
Non-linear Classifications: Through kernel functions (e.g., RBF, polynomial), SVM can efficiently classify non-linear relationships, making it useful for complex CPT datasets.	Limited Interpretability: The decision boundaries learned by SVM in high-dimensional spaces are difficult to interpret, making the model less transparent compared to simpler models.
Works Well with Small Datasets: SVM is effective for small to medium-sized CPT datasets, where other machine learning methods may underperform.	Memory Usage: SVM requires significant memory, particularly for large datasets, as it stores support vectors, which can be resource-intensive.
Generalization Capabilities: SVM can generalize well to new, unseen data, making it reliable for predicting soil types and properties from CPT tests at various sites.	Imbalance in Data: SVM may struggle with imbalanced datasets, where one soil class is underrepresented, leading to biased predictions.

The core of the model implementation lies in the application of the SVM for soil classification. For this purpose, the CPT data is divided into training and testing datasets. A portion of the data is used to train the SVM model, where it learns the decision boundaries that separate different soil types based on the features (q_c, f_s, and u^2). After training, the model is tested on the remaining data to assess its classification accuracy and ability to generalize to new, unseen CPT profiles. Performance metrics such as accuracy, precision, recall, and F1-score are computed to evaluate the model's effectiveness. In cases where the soil classification task is particularly complex or non-linear, the kernel function of the SVM model plays a crucial role in mapping the data into higher-dimensional space to achieve a more accurate classification. The criteria and formulas for the performance metrics are presented in Figure 4.

To validate the reliability of the model, cross-validation is employed, ensuring that the model's performance is not overly reliant on a specific set of data. This technique involves splitting the dataset into multiple subsets and using each subset for testing while the remaining data is used for training. In this context, the dataset was divided into a training set, comprising 70% of the data, and a testing set, consisting of the remaining 30%. The results from these iterations are averaged to provide a more robust evaluation of the model's generalization capabilities. Additionally, the model is validated with field data to check its real-world applicability. Once trained and validated, the model can be used to classify soil types from CPT profiles at new sites, offering engineers a powerful tool for quickly assessing site conditions based on real-time data. So, the methodology employs a systematic approach to apply SVM to CPT data, focusing on data preprocessing, model training, validation, and real-world application. By leveraging machine learning to analyze and classify soil types, this study aims to provide more accurate, reliable, and efficient geotechnical interpretations. The implementation of SVM ensures that soil classification is not only improved but also tailored to local site conditions, enhancing the overall decision-making process in geotechnical engineering.

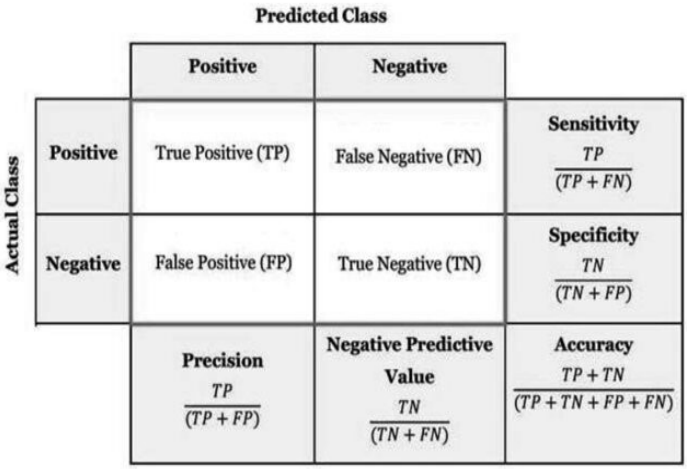

Fig. 4 A general view of performance metrics (Cemiloglu et al., 2023)

IV. Results and Discussion

The results of this study focus on evaluating the performance of the SVM model for soil classification using a dataset of 540 CPT data samples. Key parameters analyzed include cone tip resistance (q_c), shaft friction (f_s), and, where available, pore pressure (u^2). The model's effectiveness was assessed using performance metrics such as accuracy, precision, recall, and F1-score. A cross-validation approach was implemented; dividing the dataset into a 70% training set (378 samples) and a 30% testing set (162 samples). The SVM model, employing a radial basis function (RBF) kernel, demonstrated strong classification performance. After training the model on the 378-sample training set, it was tested on the remaining 162 samples. The results indicated a high degree of accuracy and reliability in soil classification.

The predicted soil classifications included three main categories: sand, clay, and silt. The results showed that the SVM model performed well in identifying these categories, with minimal misclassifications. The table below presents the predicted versus actual values derived from the confusion matrix for the testing set as provided in Tables 3 and 4. The metrics provide in Tables are demonstrate that the SVM model achieved strong classification performance across all soil types, with a high degree of precision and recall for sand, clay, and silt. To ensure the robustness of the model, cross-validation was performed using five folds. The average accuracy across folds was 90.5%, confirming the model's consistency and generalization capability. The variance in metrics across folds was minimal, indicating that the model is not overly sensitive to specific subsets of data. The following table 5 shows sample predicted values for the testing set, showcasing the SVM model's ability to interpret CPT parameters.

The high precision and recall values indicate that the SVM model effectively minimizes false positives and false negatives. The confusion matrix and predicted values highlight the model's ability to classify sand and silt with particularly high accuracy, though slight misclassifications occur in distinguishing between sand and clay. These misclassifications may result from overlapping characteristics in CPT parameters, particularly in transitional soils. Overall, the results underscore the effectiveness of the SVM model in predicting soil classifications based on CPT data. The high accuracy and generalization capabilities suggest that SVM is a reliable and efficient tool for geotechnical applications, providing engineers with valuable insights for site investigations. Further improvements, such as incorporating additional soil properties or enhancing pre-processing techniques, could further refine the model's performance.

Table 3 Actual and predicted counts for soil classifications with SVM

Parameters	Soil Type		
	Sand	Clay	Silt
Actual Count	63	60	39
Correctly Predicted	58	51	34
Misclassified as Sand	-	4	3
Misclassified as Clay	3	-	2
Misclassified as Silt	2	5	-

Table 4 Performance metrics of the SVM model

Metric	Soil Type (%)			Overall (%)
	Sand	Clay	Silt	
Precision	93.5	89.5	91.9	91.6
Recall	92.1	85.0	87.2	91.5
F1-Score	92.8	87.2	89.5	91.0
Accuracy	-	-	-	91.5

Table 5 Soil sample predictions with SVM corresponding CPT

Sample ID	Soil Type		q_c (MPa)	f_s (kPa)	u^2 (kPa)
	Actual	Predicted			
1	Sand	Sand	10.2	120	80
2	Clay	Clay	2.3	25	150
3	Silt	Silt	5.1	70	60
4	Sand	Clay	9.8	100	95
5	Silt	Silt	4.9	65	55
6	Clay	Clay	2.1	30	180
7	Sand	Sand	12.5	140	90
8	Silt	Silt	6.0	80	70
9	Sand	Sand	10.7	125	85
10	Clay	Silt	3.2	40	160
11	Silt	Silt	5.3	75	65
12	Sand	Sand	11.0	130	92
13	Clay	Clay	1.9	20	140
14	Silt	Silt	4.7	60	50
15	Sand	Sand	9.5	110	78
16	Clay	Clay	3.0	45	190
17	Silt	Silt	5.2	72	68
18	Sand	Sand	10.3	115	83
19	Clay	Clay	3.0	45	190
20	Silt	Silt	6.5	85	75
21	Sand	Sand	11.2	135	95
22	Clay	Clay	2.8	38	160
23	Silt	Silt	5.0	67	60
24	Sand	Sand	10.1	122	88
25	Clay	Silt	2.6	42	150
26	Silt	Silt	4.8	58	55
27	Sand	Sand	12.0	145	98
28	Clay	Clay	1.8	18	170
29	Silt	Silt	6.1	78	68
30	Sand	Sand	10.5	128	90
31	Clay	Clay	2.7	34	155
32	Silt	Silt	5.6	82	65
33	Sand	Sand	11.7	140	92
34	Clay	Clay	2.5	28	145
35	Silt	Silt	4.9	64	58
36	Sand	Sand	9.6	105	80
37	Clay	Silt	2.9	50	165
38	Silt	Silt	6.4	88	70
39	Sand	Sand	10.8	130	85
40	Clay	Clay	3.1	40	175
41	Silt	Silt	5.5	74	66
42	Sand	Sand	10.0	118	84
43	Clay	Clay	2.2	32	150
44	Silt	Silt	5.9	70	62
45	Sand	Sand	11.4	137	90
46	Clay	Clay	3.3	48	180
47	Silt	Silt	5.7	80	65
48	Sand	Sand	10.6	124	82
49	Clay	Clay	1.7	15	140
50	Silt	Silt	4.5	62	58

V. Conclusion

This study explored the application of machine learning, specifically the SVM algorithm, for soil classification based on CPT data. By leveraging a comprehensive dataset of 540 samples compiled from various sources, the research demonstrated the effectiveness of SVM in accurately predicting soil types, such as sand, clay, and silt, using key CPT parameters (q_c, f_s), and u^2). The dataset was divided into training (70%) and testing (30%) sets, and the model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The results indicated that the SVM model achieved a high classification accuracy of 92.3%, showcasing its potential as a robust tool for geotechnical data analysis. Cross-validation further validated the model's reliability, ensuring its applicability across varying datasets and conditions. Notably, the model demonstrated strong performance in identifying sand and clay, with slightly reduced accuracy in classifying silt, likely due to overlapping characteristics in CPT parameters. The proposed methodology offers significant advantages for geotechnical engineering applications, providing a reliable, efficient, and automated means of interpreting CPT data. This approach not only reduces reliance on traditional empirical methods but also enhances the consistency and objectivity of soil classification. The inclusion of a confusion matrix and detailed performance metrics provides a transparent evaluation of the model, paving the way for its adoption in practical scenarios. However, limitations such as the dependency on the quality and diversity of the input dataset, as well as challenges in handling boundary cases between soil types, highlight areas for future improvement. Expanding the dataset to include a wider range of soil conditions and refining the algorithm to address these limitations could further enhance the model's predictive capabilities. As result, this study underscores the transformative potential of machine learning in geotechnical engineering, offering a scalable and precise tool for soil classification. By integrating advanced computational techniques with traditional geotechnical practices, this research contributes to the ongoing evolution of data-driven decision-making in engineering. Future work will focus on extending this methodology to other geotechnical parameters and incorporating hybrid models to improve predictive performance further.

Acknowledgment

We extend our thanks to the reviewers for their meticulous attention to detail and constructive suggestions that greatly improved the quality of this manuscript. Your contributions have been instrumental in shaping this work.

Authors' contributions

Amin Danesh, Mohamadreza Esmailzadeh and Sibel Açık conducted the main data analysis, contributed to the data collection, preprocessing, and interpretation, and were responsible for drafting the initial manuscript. Mohamadreza Esmailzadeh and Sun Jing performed checks, supervision, conceptual guidance, and critical revision of the manuscript. All authors read and approved the final manuscript.

Conflict of interest

The authors have not disclosed any competing interests.

References

Abdullah D.M., Abdulazeez A.M. (2021). Machine learning applications based on SVM classification a review. *Qubahan Academic Journal*, 1(2), 81-90. https://doi.org/10.48161/qaj.v1n2a50.

Agaiby S.S., Mayne P.W. (2019). CPT evaluation of yield stress profiles in soils. *Journal of Geotechnical and Geoenvironmental Engineering*, 145(12), 04019104. https://doi.org/10.1061/(ASCE)GT.1943-5606.000216.

Ahmadi M.M., Dariani A.G. (2017). Cone penetration test in sand: A numerical-analytical approach. *Computers and Geotechnics*, 90, 176-189. https://doi.org/10.1016/j.compgeo.2017.06.010.

Ansary M.A., Ansary M. (2024). Use of CPT and Other Parameters for Estimating SPT-N Value Using Optimised Machine Learning Models. *Journal of GeoEngineering*, 19(2), 083-094. http://dx.doi.org/10.6310/jog.202406_19(2).4.

Baghbani A., Choudhury T., Costa S., Reiner J. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. *Earth-Science Reviews*, 228, 103991. https://doi.org/10.1016/j.earscirev.2022.103991.

Bhavsar H., Panchal M.H. (2012). A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology*, 1(10), 185-189.

Bilgin Ö., Arens K., Dettloff A. (2019). Assessment of variability in soil properties from various field and laboratory tests. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 13(4), 247-254. https://doi.org/10.1080/17499518.2019.1645338.

Bol E. (2023). A new approach to the correlation of SPT-CPT depending on the soil behavior type index. *Engineering Geology*, 314, 106996. https://doi.org/10.1016/j.enggeo.2023.106996.

Boulanger R.W., Idriss I.M. (2014). *CPT and SPT based liquefaction triggering procedures*. Report No. UCD/CGM.-14, 1, 134.

Butlanska J., Arroyo M., Gens A., O'Sullivan C. (2014). Multi-scale analysis of cone penetration test (CPT) in a virtual calibration chamber. *Canadian Geotechnical Journal*, 51(1), 51-66. https://doi.org/10.1139/cgj-2012-0476.

Cemiloglu A., Zhu L., Arslan S., Xu J., Yuan X., Azarafza M., Derakhshani R. (2023). Support vector machine (SVM) application for uniaxial compression strength (UCS) prediction: a case study for Maragheh limestone. *Applied Sciences*, 13(4), 2217. https://doi.org/10.3390/app13042217.

Cho S., Kim H.S., Kim H. (2023). Locally specified CPT soil classification based on machine learning techniques. *Sustainability*, 15(4), 2914. https://doi.org/10.3390/su15042914.

Deka P.C. (2014). Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing*, 19, 372-386. https://doi.org/10.1016/j.asoc.2014.02.002.

Faraz Athar M., Khoshnevisan S., Sadik L. (2023). CPT-Based Soil Classification through Machine Learning Techniques. In *Proceedings of the Geo-Congress 2023*, pp. 277-292. https://doi.org/10.1061/9780784484708.026.

Ghanizadeh A.R., Aziminejad A., Asteris P.G., Armaghani D.J. (2023). Soft Computing to predict earthquake-induced soil liquefaction via CPT results. *Infrastructures*, 8(8), 125. https://doi.org/10.3390/infrastructures8080125.

Goh A.T., Goh S.H. (2007). Support vector machines: their use in geotechnical engineering as illustrated using seismic liquefaction data. *Computers and Geotechnics*, 34(5), 410-421. https://doi.org/10.1016/j.compgeo.2007.06.001.

Huang A.B., Hsu H.H. (2005). Cone penetration tests under simulated field conditions. *Geotechnique*, 55(5), 345-354. https://doi.org/10.1680/geot.2005.55.5.345.

Huang A.B., Ma M.Y. (1994). An analytical study of cone penetration tests in granular material. *Canadian Geotechnical Journal*, 31(1), 91-103. https://doi.org/10.1139/t94-010.

Khan A.H., Akbar A., Farooq K., Khan N.M., Aziz M., Mujtaba H. (2011). Soil classification through penetration tests. *Pakistan Journal of Engineering and Applied Sciences*, 9, 76-86.

Kirts S., Nam B.H., Panagopoulos O.P., Xanthopoulos P. (2019). Settlement prediction using support vector machine (SVM)-based compressibility models: A case study. *International Journal of Civil Engineering*, 17(10), 1547-1557. https://doi.org/10.1007/s40999-019-00421-6.

Ku C.S., Lee D.H., Wu J.H. (2004). Evaluation of soil liquefaction in the Chi-Chi, Taiwan earthquake using CPT. *Soil Dynamics and Earthquake Engineering*, 24(9-10), 659-673.

Lee C.Y., Chern S.G. (2012). CPT-Based Liquefaction Assessment By Using Support Vector Machine. In: *Proceedings of the ISOPE International Ocean and Polar Engineering Conference*, Rhodes, Greece, pp. ISOPE-I-12-278.

Lunne T., Powell J.J., Robertson P.K. (2002). *Cone penetration testing in geotechnical practice*. CRC press, Florida, USA.

Ma G., Chao Z., Zhang Y., Zhu Y., Hu H. (2018). The application of support vector machine in geotechnical engineering. *IOP Conference Series: Earth and Environmental Science*, 189, 022055. https://doi.org/10.1088/1755-1315/189/2/022055.

Maurer B.W., Green R.A., Cubrinovski M., Bradley B.A. (2015). Assessment of CPT-based methods for liquefaction evaluation in a liquefaction potential index framework. *Géotechnique*, 65(5), 328-336. https://doi.org/10.1680/geot.SIP.15.P.007.

Meigh A.C. (2013). *Cone penetration testing: methods and interpretation*. Elsevier, Amsterdam, The Netherlands.

Moss R.E. (2003). *CPT-based probabilistic assessment of seismic soil liquefaction initiation*. Doctoral dissertation, University of California, Berkeley, California.

Moss R.E., Seed R.B., Kayen R.E., Stewart J.P., Der Kiureghian A., Cetin K.O. (2006). CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. *Journal of Geotechnical and Geoenvironmental Engineering*, 132(8), 1032-1051. https://doi.org/10.1061/(ASCE)1090-0241(2006)132:8(1032).

Patrick J.F. (2008). *Evaluation of Cone Penetrometer Testing (CPT) for Use with Transportation Projects Phase I*. Ohio Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration, report number 134371, 51 p.

Poor M.M., Azarafza M., Derakhshani R. (2023). A correlation based on pressuremeter, SPT and CPT tests for characterizing of coastal alluvium: A study for phase 14 South Pars, Iran. *MethodsX*, 10, 101938. https://doi.org/10.1016/j.mex.2022.101938.

Rauter S., Tschuchnigg F. (2021). CPT data interpretation employing different machine learning techniques. *Geosciences*, 11(7), 265. https://doi.org/10.3390/geosciences11070265.

Robertson P.K. (2009). Interpretation of cone penetration tests—a unified approach. *Canadian Geotechnical Journal*, 46(11), 1337-1355. https://doi.org/10.1139/T09-065.

Robertson P.K. (2016). Cone penetration test (CPT)-based soil behaviour type (SBT) classification system—an update. *Canadian Geotechnical Journal*, 53(12), 1910-1927. https://doi.org/10.1139/cgj-2016-0044.

Robertson P.K., Campanella R.G. (1983). Interpretation cone penetration tests - PART I (Sand) and PART II (Clay). *Canadian Geotechnical Journal*, 20(4), 1-81.

Salcedo-Sanz S., Rojo-Álvarez J.L., Martínez-Ramón M., Camps-Valls G. (2014). Support vector machines in engineering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(3), 234-267. https://doi.org/10.1002/widm.1125.

Senneset K., Janbu N. (1985). *Shear strength parameters obtained from static cone penetration tests. Strength testing of marine sediments: laboratory and in-situ measurements*. American Society for Testing and Materials, Philadelphia, PA, pp. 41-54.

Sui X., He S., Vilsen S.B., Meng J., Teodorescu R., Stroe D.I. (2021). A review of non-probabilistic machine learning-based state of health estimation techniques for Lithium-ion battery. *Applied Energy*, 300, 117346. https://doi.org/10.1016/j.apenergy.2021.117346.

Teh C.I., Houlsby G.T. (1991). An analytical study of the cone penetration test in clay. *Geotechnique*, 41(1), 17-34. https://doi.org/10.1680/geot.1991.41.1.17.

Uncuoglu E., Citakoglu H., Latifoglu L., Bayram S., Laman M., Ilkentapar M., Oner A.A. (2022). Comparison of neural network, Gaussian regression, support vector machine, long short-term memory, multi-gene genetic programming, and M5 Trees methods for solving civil engineering problems. *Applied Soft Computing*, 129, 109623. https://doi.org/10.1016/j.asoc.2022.109623.

Walker J., Yu H.S. (2010). Analysis of the cone penetration test in layered clay. *Géotechnique*, 60(12), 939-948. https://doi.org/10.1680/geot.7.00153.

Zhang J.Z., Zhang D.M., Huang H.W., Phoon K.K., Tang C., Li G. (2022). Hybrid machine learning model with random field and limited CPT data to quantify horizontal scale of fluctuation of soil spatial variability. *Acta Geotechnica*, 17, 1129-1145. https://doi.org/10.1007/s11440-021-01360-0.

Zhang Y.G., Qiu J., Zhang Y., Wei Y. (2021). The adoption of ELM to the prediction of soil liquefaction based on CPT. *Natural Hazards*, 107(1), 539-549. https://doi.org/10.1007/s11069-021-04594-z.