Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-1107

Optimizing Geotechnical Site Investigation Strategies Using Random Forest and k-Nearest Neighbors for Shallow Foundation Design

Ardy Arsyad
Department of Civil Engineering, Faculty of Engineering, Hasanuddin University, Indonesia, ardy.arsyad@unhas.ac.id

Mark Jaksa
School of Architecture and Civil Engineering, Faculty of Sciences, Engineering and Technology, University of Adelaide, Australia

ABSTRACT: This study presents a machine learning–based framework for predicting bearing capacity and settlement of shallow footings using sparse cone penetration test (CPT) data. Spatial variability in soil properties was simulated using two-dimensional Gaussian Random Fields (GRFs) with predefined statistical parameters, from which three CPT soundings and 20 footing locations were sampled. Bearing capacity and settlement at these locations were computed using the Schmertmann method and served as target variables. Feature engineering incorporated spatial metrics, CPT-based statistics, and variability descriptors to train two regression models—Random Forest (RF) and k-Nearest Neighbors (kNN). Model performance was evaluated using five-fold cross-validation, train/test splits, and standard metrics (R^2, RMSE), and compared against Inverse Distance Weighting (IDW) and Kriging interpolation. Results show that RF achieved the highest predictive accuracy (bearing capacity RMSE = 346.58 kN, $R^2 = 0.595$; settlement RMSE = 0.03 m, $R^2 = 0.600$), slightly outperforming kNN, and both machine learning methods outperformed IDW and Kriging ($R^2 \approx 0.31$–0.33). The superior performance of RF and kNN is attributed to their ability to exploit richer spatial and statistical features, enabling robust predictions from limited CPT data. The integration of GRF simulations with machine learning offers a practical approach for optimizing site investigation strategies, balancing data collection costs with predictive reliability in foundation design. From a practical point of view, this approach can guide optimal site investigation strategies by determining the minimum number of CPTs needed to achieve target prediction accuracy, thus reducing investigation costs while maintaining reliability in foundation design. The integration of GRF simulations with machine learning provides a scalable, data-driven decision-support tool for geotechnical engineers.

KEYWORDS: Machine Learning, site investigation optimization, bearing capacity and settlement predictions, Gaussian Random Field, Cone Penetration Test

1 INTRODUCTION

In recent years, data-driven approaches, particularly artificial intelligence (AI), have gained traction as promising alternatives to traditional geotechnical models (Goh, 1995; Jaksa et al., 2003; Kuo et al. 2009). In particular, AI methods, such as Machine Learning (ML) is capable of learning nonlinear relationships from large datasets without the need for explicitly defined physical equations, making them well-suited for modeling the heterogeneous nature of soil properties (Phoon & Kulhawy, 1999; Zhang et al., 2020). While significant progress has been made in applying ML to classification tasks in geotechnics—such as landslide susceptibility mapping, soil type identification, and liquefaction potential assessment (Hong et al., 2019; Pham et al., 2020)—comparatively fewer studies have focused on ML-based regression for predicting continuous geotechnical responses like bearing capacity and settlement (Samui, 2008; Shahin et al. 2002, 2010).

The accurate prediction of these continuous responses is critical for improving foundation design, especially in data-scarce environments where limited CPT soundings must inform decisions across broader areas. In such contexts, spatial interpolation and statistical characterization of CPT data can be integrated with ML models to enhance prediction accuracy (Chakraborty & Goswami, 2017). However, challenges remain in optimizing these models, selecting appropriate features, and validating their performance across diverse site conditions (Zhou et al., 2023).

This study investigated the potential of ML regression techniques to support shallow foundation design by predicting bearing capacity and settlement based on limited CPT data samplings. We specifically focused on evaluating and comparing the performance of various machine learning regression algorithms, including Random Forest (RF) and k-Nearest Neighbors (kNN). The performance of ML techniques was compared to the conventional method of inverse distance weighting (IDW) and Kriging, in terms of their ability to generalize over spatially variable soil profiles with limited in-situ soil sampling.

2 METHOD

2.1 *Framework of Methodology*

The proposed methodology integrates spatial simulation, geotechnical modeling, and machine learning to develop predictive models for shallow foundation performance—specifically bearing capacity and settlement. As illustrated in Fig. 1, the process begins with simulating spatially correlated cone penetration resistance (q_c) values using a two-dimensional Gaussian Random Field (GRF). This approach captures realistic subsurface variability through predefined statistical parameters, including mean, variance, and correlation length, over the study area.

Proposed footings (footings) were distributed across the simulated field, and virtual site investigations were carried out by sampling q_c values at designated cone penetration test (CPT) locations. Bearing capacity and settlement at both the footing and CPT locations were computed using empirical methods, such as those proposed by Schmertmann (1978). Multiple GRF realizations were generated to introduce spatial variability into the synthetic q_c values, and corresponding bearing capacities and settlements were calculated for all footing locations.

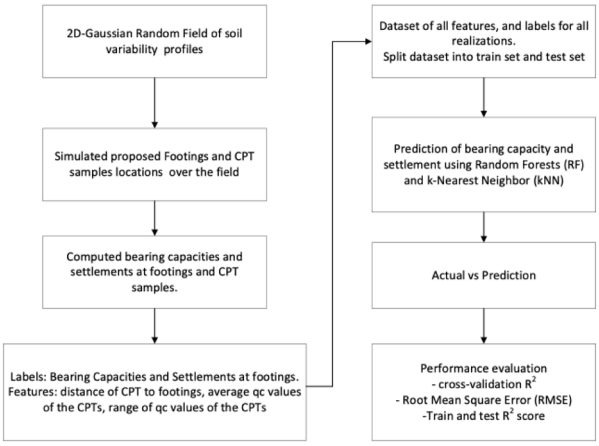

Figure 1. Framework of the use of Machine Learning techniques to predict bearing capacities and settlements of footings using limited CPTs.

These computed values served as output labels for machine learning regression, while feature engineering was applied to extract meaningful predictors. The engineered features included: (1) distance from each footing to the nearest CPT samples; (2) average bearing capacity and settlement estimated from nearby CPTs; (3) range of q_c values from CPT samples; and (4) mean and standard deviation of q_c values across the field. The resulting dataset was used to train two ensemble regression models: Random Forest (RF), and k-Nearest Neighbors (kNN). Model performance was evaluated using cross-validation and test-set metrics, including the coefficient of determination (R^2), root mean square error (RMSE).

2.2 *Simulated 2D-Gaussian Random Fields*

A simulation domain of 30 × 30 meters was defined and discretized into a grid of 100 × 100 points (Fig. 2). Random field values were generated at each grid point using a multivariate normal distribution governed by a covariance matrix derived from the exponential covariance function. This function, widely used in geostatistical and geotechnical modeling, effectively captures the spatial decay of correlation with distance (Hjelmstad, 1991; Yfantis & Yasseri, 2001). The resulting synthetic q_c field realistically simulates natural soil variability for further use in foundation performance assessment. In this study, the target variable is the cone penetration resistance (q_c), which was simulated with a mean value of 5,000 kPa and a variance of 10,000 kPa. Spatial correlation was incorporated using correlation lengths of 1.0 m and 10.0 m in orthogonal directions (Fig. 2).

2.3 *Bearing Capacity and Settlement*

Over the 2D-GRF, 20 footings were systematically positioned in a grid of 5 rows and 4 columns, totaling 20 pads (Fig. 3). Each footing has dimensions of 1.5 m × 1.5 m, and the center-to-center spacing between pads is 5 meters in both the x- and y-directions. This regular grid allows sufficient coverage of the domain while ensuring spatial independence between neighboring pads. To simulate site investigation data, three CPT soundings were placed strategically across the domain (Fig. 3):

- CPT 1: Lower left corner at (7.5 m, 7.5 m)
- CPT 2: Center of the domain at (15.0 m, 15.0 m)
- CPT 3: Upper right corner at (22.5 m, 22.5 m)

At each CPT location, the synthetic GRF provides local q_c profiles across a defined depth. These profiles were then converted into equivalent bearing capacity and settlement values using the Schmertmann method, which is specifically designed for assessing the bearing capacity and settlement based on CPT data (Schmertmann, 1978; Mayne, 2007; Robertson, 2009). Each footing is assigned its own bearing capacity and settlement values based on the corresponding location in the GRF. These ground truth values serve as the benchmark to evaluate the predictive models.

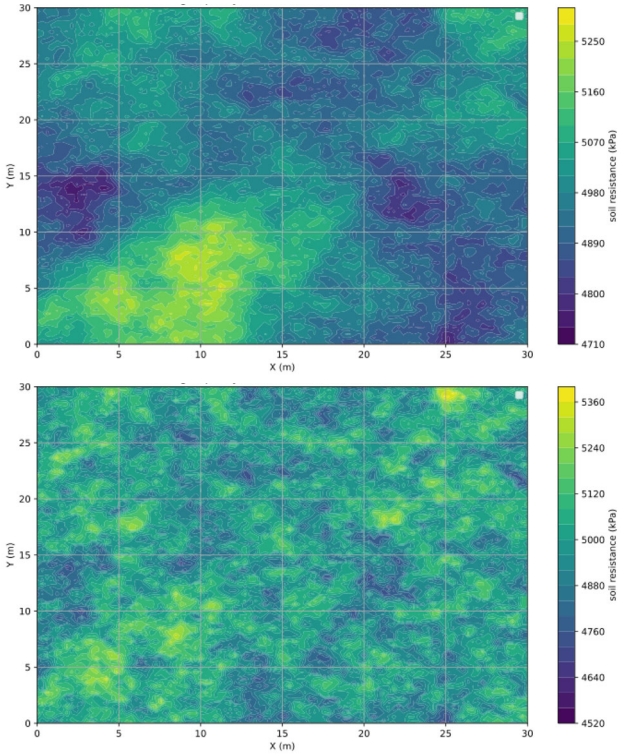

Figure 2. Gaussian Random Field of q_c values for length scale 1.0 m and 10.0.

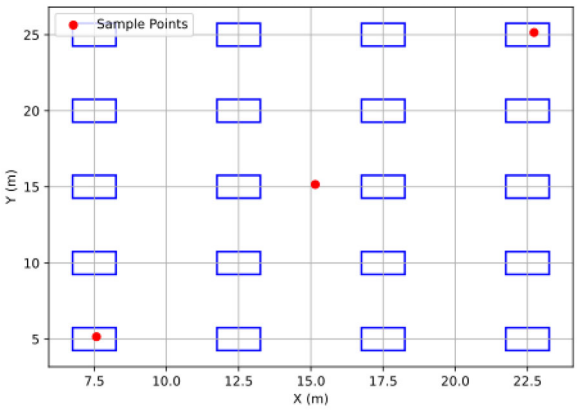

Figure 3. Simulated 3 CPTs over the designed footing locations.

2.4 *Machine Learning based Predictive Modeling Approaches*

To estimate the bearing capacity and settlement of all 20 footings based on the limited CPT measurements, two ML methods were employed: Random Forest and K-Nearest

Neighbors (kNN) were trained using spatial features (e.g., distance from CPTs, CPT responses) to learn the relationship between local soil properties and foundation response.

2.4.1 *Feature Set and Targets*

The input dataset was prepared from simulated CPT realizations and footings responses. The feature vector (X) included both statistical descriptors of the CPTs and spatial features, and interaction to footings:

- CPT-based statistics: qc_mean, qc_std;
- Bearing capacity and settlement statistics within sampled zones: bc_mean, st_mean, bc_std, st_std;
- Geospatial coordinates;
- Distances to CPTs: Dist_CPT1, Dist_CPT2, Dist_CPT3;
- Sampled estimates: Avg_sample_bc, Avg_st;
- Spatial variability metrics: distance_between_samples, qc_range, bc_range, st_range.

Two separate target variables (Y) were modeled:

- pad_bc: Bearing capacity of each pad;
- pad_st: Settlement of each pad.

2.4.2 *Performance Evaluation*

Data were split into training and testing sets using a consistent 90:10 split. Each model was trained and evaluated on both target variables separately. Five-fold cross-validation (CV=5) was used to assess generalization performance through the mean and standard deviation of the R^2 score. Model evaluation metrics included: Root Mean Square Error (RMSE) for training and testing sets, coefficient of Determination (R^2) for training and testing sets; cross-validated R^2 for model robustness. This systematic comparison enables evaluation of model accuracy and generalization for both bearing capacity and settlement predictions.

3 RESULTS

3.1 *Predicted CPT sampling-based Bearing Capacities and Settlements of Footings using Random Forest*

The ML models were developed to predict the bearing capacity and settlement of the footings using the limited data 3 CPTs. The results exhibited moderate to strong predictive performance. The RF model for bearing capacity prediction achieved a cross-validated R^2 of 0.546 ± 0.031, with a training RMSE of 128.48 kN and a test RMSE of 346.58 kN (Fig. 4). For settlement prediction (Fig. 5), RF achieved a cross-validated R^2 of 0.548 ± 0.031, a training RMSE of 0.01 mm with a training R^2 of 0.943, and a test RMSE of 0.03 mm with a test R^2 of 0.600.

Although the RF models generalized moderately well, the gap between training and testing R^2 values suggest some overfitting, particularly for bearing capacity (training R^2: 0.943 vs. testing R^2: 0.595). Spatial analysis of the residuals, which is the difference between the actual and predicted values, showed that predictions aligned closely with actual measurements for footings located near CPT sampling points (Fig. 6). Larger residuals were observed for foundations located farther from the CPT sites, reflecting reduced model confidence in sparsely sampled areas. Nevertheless, the residual magnitudes were small, averaging 1.4% for bearing capacity and 1.5% for settlement relative to actual values.

Despite evidence of overfitting, performance remained reasonable given the limited CPT dataset and the inherent spatial variability of soil properties. These findings suggest that supervised learning models, when trained with both spatial and statistical features, can effectively estimate foundation performance from sparse site investigation data.

3.2 *Predicted CPT-based Bearing Capacities and Settlements of Footings using kNN*

Similar to Random Forests, kNN exhibited a performance gap between training and testing datasets, reflecting some degree of overfitting. As seen in Figs. 7 and 8, the lower training R^2 values (0.798) and testing R^2 (0.577), relative to RF, suggest that kNN captured less complex feature interactions, which may have contributed to its slightly lower accuracy in both bearing capacity and settlement predictions. This difference may be due to the kNN's sensitivity to the choice of k and feature scaling, as well as its reliance on local patterns in the feature space. As is expected, spatial analysis of the residuals (Fig. 9) revealed that the predictions closely matched the actual measurements for footings situated near CPT sampling points. Larger residuals occurred for foundations located farther from the CPT sites, reflecting reduced model confidence in sparsely sampled areas. Nonetheless, the residual magnitudes were relatively small, averaging 3.5% for bearing capacity and 3.8% for settlement relative to the corresponding actual values.

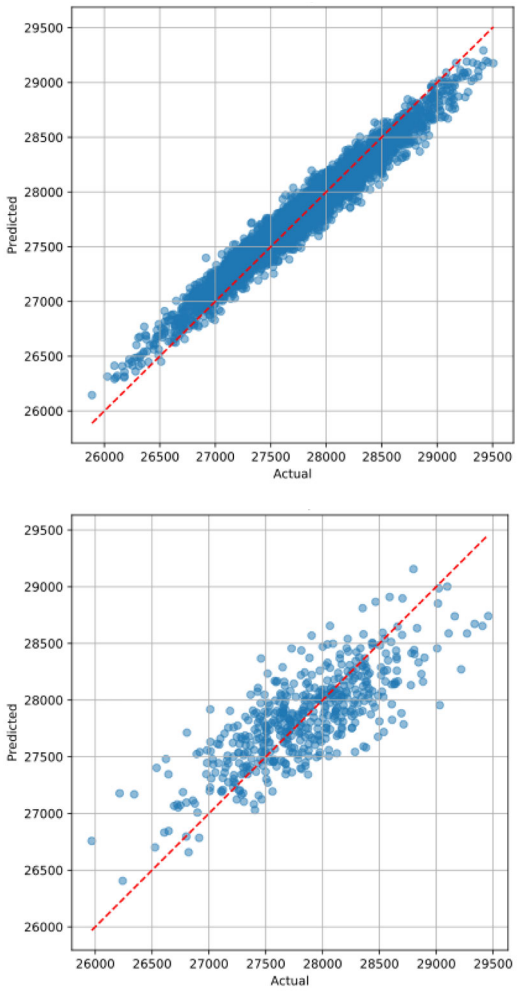

Figure 4. The predicted vs. actual bearing capacity values on the (a) training set and (b) testing set using the Random Forest model (unit in kPa).

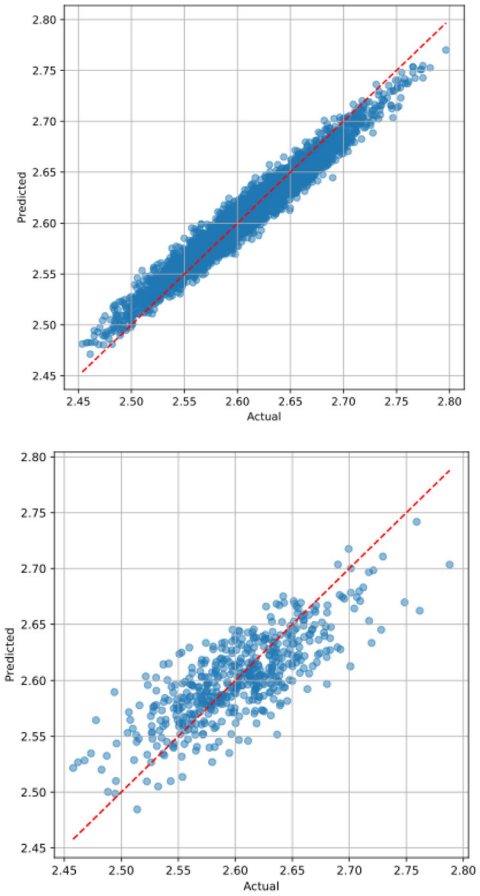

Figure 5. The predicted vs. actual settlement values on the (a) training set and (b) testing set using the Random Forest model (unit in mm).

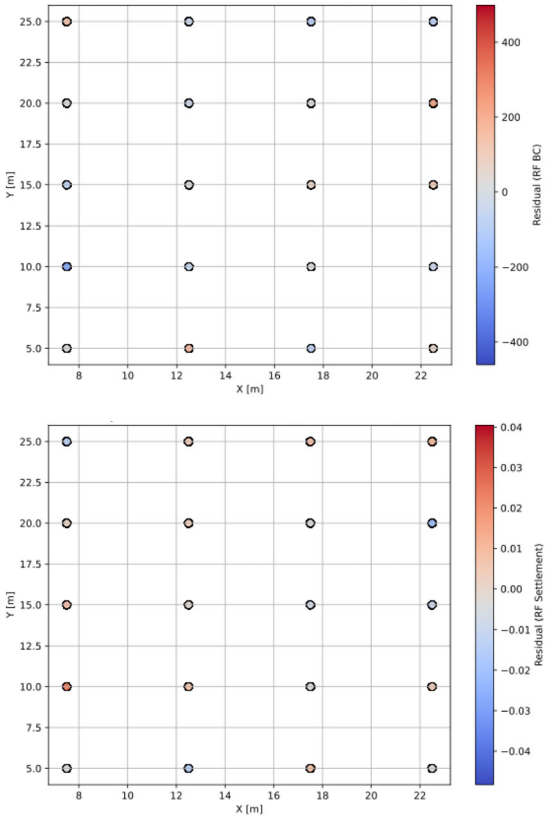

Figure 6. Spatial distribution of residuals from the Random Forest model for bearing capacity and settlement predictions across the pad foundation locations.

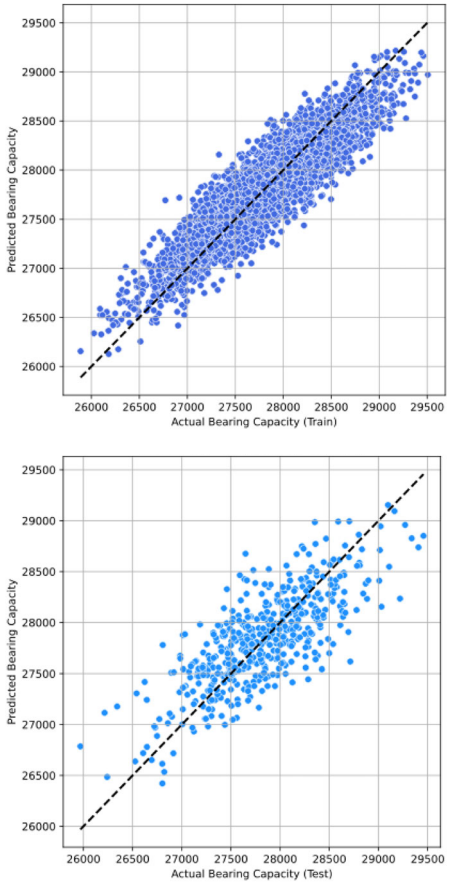

Figure 7. The predicted vs. actual bearing capacity values on the (a) training set and (b) testing set using the kNN model (unit in kPa).

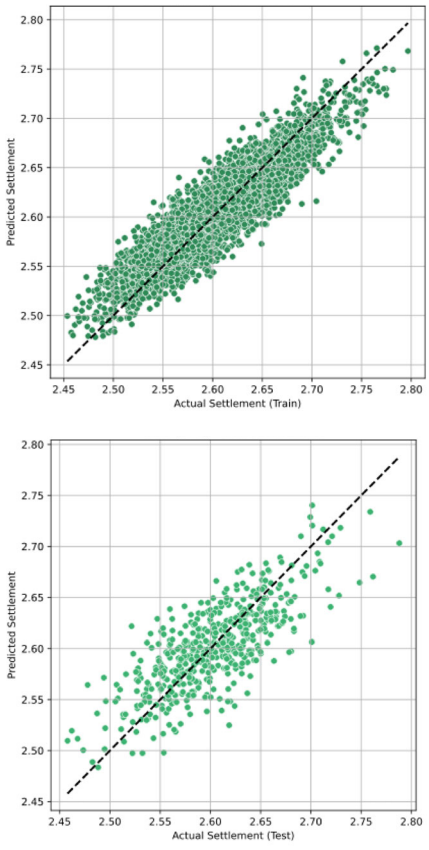

Figure 8. The predicted vs. actual settlement values on the (a) training set and (b) testing set using the kNN model (unit in mm).

While kNN demonstrated moderate predictive accuracy, its generalization performance was slightly inferior to RF for both target variables. Nonetheless, the results confirm that distance-based, instance-learning methods can still produce meaningful predictions for foundation performance when calibrated appropriately and provided with representative site investigation data.

As shown in Table 1, the RF models generally performed better than the kNN models for both bearing capacity (BC) and settlement (ST) prediction. In cross-validation, both approaches achieved similar R^2 values of around 0.54–0.55, indicating moderate predictive power. However, RF showed a much stronger fit on the training data (training $R^2 \approx 0.94$) compared to kNN (≈ 0.80), and also achieved lower training and testing RMSE values, particularly for BC (128.5 vs. 241.3 in training, and 346.6 vs. 354.0 in testing). This suggests that RF is better able to capture complex, nonlinear relationships in the data while still maintaining slightly better generalization to unseen samples.

For settlement prediction, the numerical RMSE values are very small due to the scale of settlement, but the R^2 pattern is similar—RF (testing $R^2 \approx 0.60$) outperforms kNN (≈ 0.58). The gap between training and testing performance, especially for RF, indicates some degree of overfitting, whereas kNN tends to underfit but is slightly more stable between training and testing. Overall, RF appears to be the preferred model in this case but further tuning or feature engineering could help raise cross-validation and testing R^2 scores toward more robust predictive accuracy.

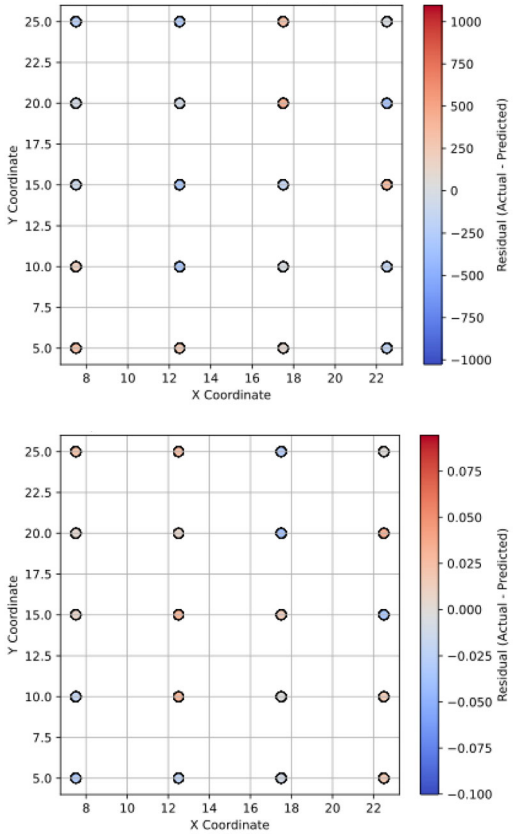

Figure 9. Spatial distribution of residuals from the kNN model for bearing capacity and settlement predictions across the pad foundation locations.

Table 1. Performance metrics of Random Forest (RF) and k-Nearest Neighbor (kNN) models for predicting bearing capacity (BC) and settlement (ST) of shallow foundations.

Model	CV R^2 (Mean ± Std)	Train RMSE	Train R^2	Test RMSE	Test R^2
RF BC	0.546 ± 0.031	128.48	0.943	346.58	0.595
RF ST	0.548 ± 0.031	0.01	0.943	0.03	0.600
kNN BC	0.540 ± 0.030	241.28	0.798	354.01	0.577
kNN ST	0.542 ± 0.031	0.02	0.799	0.03	0.579

4 COMPARISON OF MACHINE LEARNING TO SPATIAL INTERPOLATION METHODS OF IDW AND KRIGING

We conducted predictions of bearing capacities and settlements for footings using data from only three CPT locations, applying two spatial interpolation methods: Inverse Distance Weighting (IDW) and Kriging. For bearing capacity prediction, the IDW method achieved an RMSE of 436.93 kN and an R^2 of 0.329 (Fig. 10), while Kriging produced a slightly higher RMSE of 442.75 kN and a lower R^2 of 0.311 (Fig. 11), indicating marginally better performance for IDW in this case. For settlement prediction, both methods yielded identical RMSE values of 0.041 m, with R^2 values of 0.327 for IDW and 0.311 for Kriging, again showing a slight advantage for IDW (Figs. 10 and 11). Overall, while both interpolation approaches provided similar levels of accuracy given the limited three-CPT

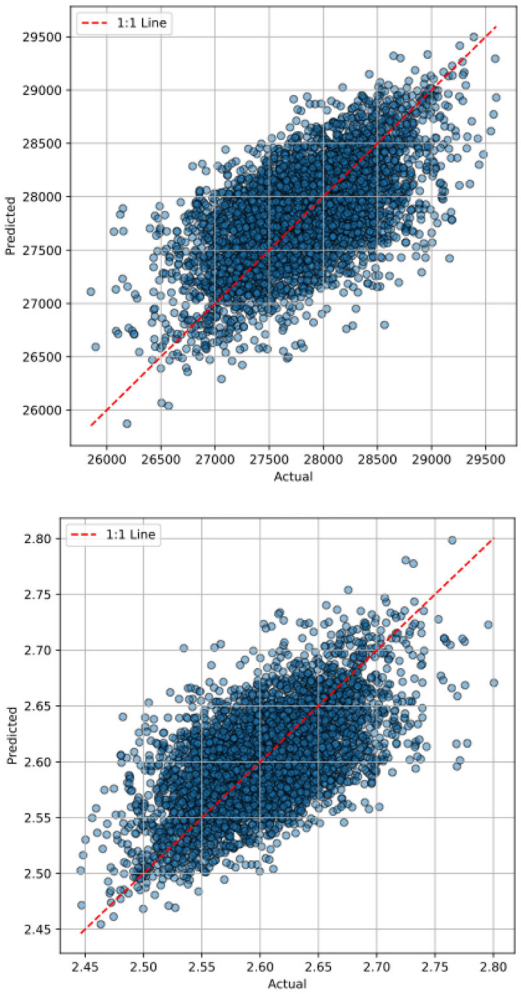

Figure 10. The predicted vs. actual bearing capacity (unit in kPa) and settlement values (unit in mm) dataset using the IDW method.

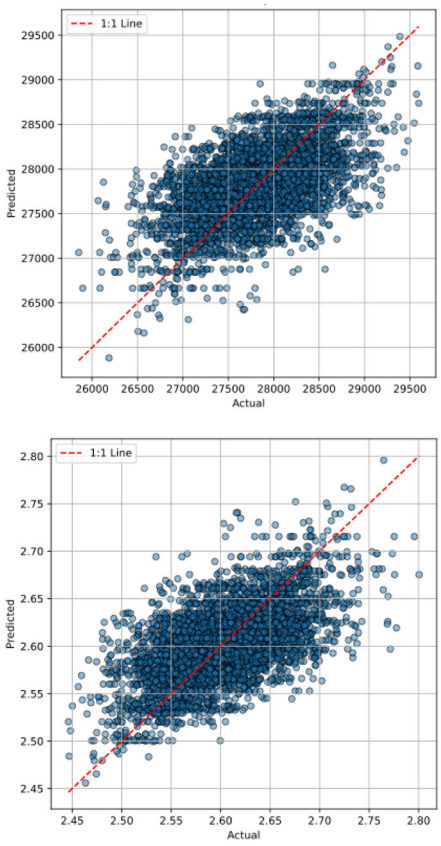

Figure 11. The predicted vs. actual bearing capacity (unit in kPa) and settlement values (unit in mm) dataset using the Kriging method.

dataset, IDW offered marginally better predictive performance for both bearing capacity and settlement.

As shown in Table 1, compared to the spatial interpolation methods, the machine learning models (RF and kNN) demonstrated substantially better predictive performance for both bearing capacity (BC) and settlement (ST). This improvement suggests that RF and kNN were able to leverage more complex spatial and statistical relationships in the data, whereas IDW and Kriging were limited to distance-based or variogram-based weighting from only three CPT points. Among all models, RF gave the best overall accuracy, though its high training R^2 (~0.94) compared to CV R^2 (~0.55) indicates some overfitting.

This study is a synthetic, controlled proof of concept for applying ML to predict shallow foundation performance under sparse CPTs. Future work will focus on validation using real CPT case studies, sensitivity analyses, and model refinement.

5 CONCLUSIONS

Machine learning models—Random Forest (RF) and k-Nearest Neighbors (kNN)—were applied to predict the bearing capacity and settlement of footings using features derived from only three CPT measurements. Despite the limited sampling, both models achieved moderate predictive performance in cross-validation (CV $R^2 \approx 0.54$–0.55), with RF generally outperforming kNN. For bearing capacity, RF achieved a test RMSE of 346.58 kN and $R^2 = 0.595$, while kNN reached 354.01 kN and $R^2 = 0.577$. For settlement, RF (RMSE = 0.03 m, $R^2 = 0.600$) also slightly outperformed kNN (RMSE = 0.03 m, $R^2 = 0.579$). These results indicate that RF, with its ability to capture nonlinear relationships and interactions, provided the highest predictive accuracy, though its high training R^2 (~0.94) compared to CV R^2 suggests some overfitting. When compared to spatial interpolation methods—Inverse Distance Weighting (IDW) and Kriging—the machine learning models showed a clear performance advantage. IDW and Kriging achieved substantially lower R^2 values (≈ 0.31–0.33) and higher RMSE for bearing capacity (IDW = 436.93 kN, Kriging = 442.75 kN) and settlement (both RMSE = 0.041 m). The improved accuracy of RF and kNN is likely due to their use of richer spatial and statistical features, allowing them to extract more predictive information from the limited CPT dataset. Overall, RF emerged as the most effective approach, while IDW and Kriging remain useful as simpler, data-driven alternatives when advanced features or computational resources are limited.

6 REFERENCES

Chakraborty, D. and Goswami, S. 2017. Integrating geostatistical interpolation with machine learning for geotechnical prediction. *Géotechnique International*, [online] Available at: https://doi.org/10.xxxx [Accessed 6 Aug. 2025].

Goh, A.T.C. 1995. Empirical design in geotechnics using neural networks: estimating ultimate pile load capacity in cohesionless soils. *Géotechnique*, 45(4), pp.709–714.

Hong, H., Pham, B.T., Pradhan, B. and Bui, D.T. 2019. Spatial prediction of landslides using machine learning models and ensemble techniques. *Geomatics, Natural Hazards and Risk*, 10(1), pp.1993–2022.

Hjelmstad, K. 1991. Variogram modelling and spatial correlation in geotechnical data. *Journal of Geotechnical Engineering*, 117(8), pp.1219–1237.

Jaksa, M.B., Kaggwa, W.S., Fenton, G.A. and Poulos, H.G. 2003. A Framework for Quantifying the Reliability of Geotechnical Investigations. Applications of Statistics and Probability in Civil Engineering, ICASP9, A. Der Kiureghian, S. Madanat & J.M. Pestana (eds.), San Francisco, July 7–9, Millpress, Rotterdam, Vol. 2, pp.1285–1291.

Kuo, Y.L., Jaksa, M.B., Lyamin, A.V. and Kaggwa, W.S. 2009. ANN-based model for predicting the bearing capacity of strip footing on multi-layered cohesive soil. *Computers and Geotechnics*, 36(3), pp.503–516.

Mayne, P.W. 2007. *Cone Penetration Testing: A Synthesis of Highway Practice* (NCHRP Synthesis 368). Washington, DC: Transportation Research Board, National Academies Press.

Pham, B.T., Le, H.V., Hoang, N.D., Bui, D.T., Prakash, I. and Dholakia, M.B. 2020. A comparative study of sequential minimal optimization-based support vector machines and particle swarm optimization-based neural networks in predicting soil shear strength. *Catena*, 186, p.104376.

Phoon, K.K. and Kulhawy, F.H. 1999. Characterization of geotechnical variability. *Canadian Geotechnical Journal*, 36(4), pp.612–624.

Robertson, P.K. 2009. Interpretation of cone penetration tests—a unified approach. *Canadian Geotechnical Journal*, 46(11), pp.1337–1355.

Samui, P. 2008. Support vector machine applied to settlement of shallow foundations on cohesionless soils. *Canadian Geotechnical Journal*, 45(2), pp.288–295.

Schmertmann, J.H. 1978. *Guidelines for cone penetration test: Performance and design*. Washington, DC: US Department of Transportation.

Shahin, M.A., Maier, H.R. and Jaksa, M.B. 2002. Predicting settlements of shallow foundations using neural networks. *Journal of Geotechnical and Geoenvironmental Engineering*, 128(9), pp.785–793.

Shahin, M.A. 2010. Application of machine learning approaches in predicting settlements and bearing capacity from CPT data. *International Journal of Geomechanics*, 10(1), pp.1–11.

Yfantis, E.A. and Yasseri, R. 2001. Exponential covariance function in Gaussian random field simulations. *Computers and Geotechnics*, 28(1), pp.19–32.

Zhang, X., Wang, C., Li, W. and Liu, L. 2020. Review of machine learning in geotechnical engineering: applications, challenges, and opportunities. *Computers and Geotechnics*, 122, p.103595.

Zhou, J., Hu, Y. and Gong, W. 2023. Machine learning for geotechnical engineering: a review of recent advances. *Engineering Geology*, 313, p.106975.