Geodata and AI 3 (2025) 100019

Contents lists available at ScienceDirect

Geodata and AI

journal homepage: www.elsevier.com/locate/geoai

Influence of feature selection on machine learning prediction of pile foundation – The role of soil-pile interaction knowledge and application to base resistance

Thanh T. Nguyen [a,*], Khuong Le Nguyen [b,c], Thien Q. Huynh [a], Quangdung Tran [d]

[a] Transport Research Centre (TRC) & School of Civil and Environmental Engineering, University of Technology Sydney, Australia
[b] University of Transport Technology, Vietnam
[c] University of Canberra, Australia
[d] Faculty of Civil and Industrial Construction, Hanoi University of Civil Engineering, Giai Phong Street, Hanoi, Vietnam

ARTICLE INFO

Keywords:
Base resistance
Pile foundation
Machine learning
Static load tests
Soil-pile interaction
Feature selection

ABSTRACT

The use of machine learning (ML) to predict and classify various aspects of pile foundation has rapidly gained paramount attention over the past decade, resulting in various sets of input features and algorithms that have been employed. This study classifies those input features into 2 groups, i.e., with and without understanding of soil-pile interaction, and examines their impacts on the development and performance of ML models. Six different ML algorithms covering from the most fundamental decision trees to advanced bagging and boosting techniques are adopted. An extensive database gathering 86 cases (1129 datapoints) of pile load tests in complex soft soil region is developed and employed to investigate the influence that 2 different sets of input features can have on predicting base resistance of piles. Through a comprehensive training-to-validation process, the results prove the importance of having soil-pile interaction features such as the load-displacement curves in optimizing the ML prediction of base resistance. On the other hand, the use of only soil and pile features for model inputs can help reduce the cost for pile tests, though it can lead to poorer prediction performance. Not only does the current study establish novel models based on machine learning to predict base resistance of piles, but it also employs quantitative feature analysis to gain insight into the load-transfer process, where the interaction between soil and pile develops downward from side to base resistances. The study proposes an adaptive training strategy that effectively improves model application to new contexts, reducing the need for extensive data collection and field survey costs, thereby enhancing the cost-effectiveness and scalability of ML deployments in geotechnical engineering.

List of acronyms

A	cross-sectional area of pile
AE	axial stiffness of pile
D	equivalent diameter of pile
d_p	displacement of loading point
E	Young's modulus
L	embedded length of pile
L_1	the first 1/3 of embedded pile length from the pile head
L_2	the second 1/3 of embedded pile length from the pile head
L_3	the third 1/3 of embedded pile length from the pile head
L_p	distance from the loading point to the pile toe
N_{SPT}	SPT value of soil
P	applied axial load
Q_b	Base resistance of pile
SPT	standard penetration test
S_{u1}	average undrained shear strength of soil within the first 1/3 pile length from the pile head (side soil)
S_{u2}	average undrained shear strength of soil within the second 1/3 pile length from the pile head (side soil)
S_{u3}	average undrained shear strength of soil within the third 1/3 pile length from the pile head (side soil)
S_{ub}	undrained shear strength of soil underneath the pile toe (base soil)

Introduction

There are approximately 40 % of population around the world living in near-coastal regions where soft soil often causes immense challenges to the design and construction of buildings and infrastructure (Reimann et al., 2023). In some countries like Australia, this propor-

* Corresponding author.
E-mail addresses: thanh.nguyen-4@uts.edu.au (T.T. Nguyen), khuongln@utt.edu.vn, khuong.lenguyen@canberra.edu.au (K.L. Nguyen).

https://doi.org/10.1016/j.geoai.2025.100019
Received 1 December 2024; Received in revised form 18 March 2025; Accepted 28 April 2025
3050-483X/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

tion can even exceed 85 % (DCCEEW, 2021). The demand for high-rise buildings and heavy transport infrastructure is usually massive in these regions due to the intensive concentration of population and dynamic industries (Truong et al., 2021), making pile foundation become the most preferable option. Many cities such as Ho Chi Minh, Jakarta, Bangkok, etc. lie on extremely soft ground with the depth can reaches 20–40 m, requiring excessive long and large piles to be used. For example, the length of bored piles used in various high-rise buildings in Ho Chi Minh City in the past 2 decades often ranges from 50 m to 90 m, some can even exceed 100 m. Traditional theories and practices of pile foundations were often limited to a depth of less than 40 m (Dai et al., 2012; El Gendy et al., 2014; Bohn et al., 2017; Al-Atroush et al., 2022), whereas understanding long pile foundation is still ambiguous due to a lack of validation data. In fact, field tests for long piles are usually costly and resource demanding as they require complex loading system with numerous sensors to reach the failure stage. It is important to note that most long piles have very large ultimate bearing capacity, e.g., > 5000 to exceed 10,000 tonnes, requiring the loading system and test procedure to be designed properly (Huynh et al., 2023).

Subjected to increasing axial load, skin resistance usually takes place first and this develops over the depth until the base resistance becomes larger with mobilised skin friction, i.e., deboning with plastic yielding downward (Sharo et al., 2022). When the pile becomes longer, the load propagation will occur to a larger depth with complex behaviours of skin friction and base resistance, resulting in considerable challenges in the practice of prediction and design. For example, the base resistance can reach a critical level with significant displacement without a full mobilisation of the skin resistance, i.e., partial plastic deformation along the depth (Li et al., 2023; Xiao et al., 2023). Despite this complex and important matter, there is still a limited number of studies on the development of base resistance especially in long piles. Some studies (Nguyen & Fellenius, 2015; Xiao et al., 2023) investigated long piles, but the main focus often goes to estimate general bearing capacity rather than the concurrent behaviour of individual components such as skin and base resistances. In fact, while there exist many practice codes and models to estimate bearing capacity of soil (AASHTO Specifications, 2010; TCVN 10304, 2014; Li et al., 2023; Xiao et al., 2023), the role of pile settlement is often neglected. This context indicates greater effort to establish an effective method to overcome this long-lasting issue of pile foundation.

Various machine learning (ML) algorithms have been used to predict behaviour of pile foundation with great success in the past 2 decades, nevertheless, there is a lack of attention to the load transfer as well as the development of shaft and base resistances under loading. Table 1 summarises different ML algorithms that have been employed for various aspects and data types of pile foundation since early efforts in 1995 and 2005 (Chan et al., 1995; Goh et al., 2005). The artificial neural network (ANN) was probably the earliest and also the most impactful algorithm that has been used to predict bearing capacity and settlement of piles, as shown in Table 1. Later, many other algorithms such as support vector machine (SVM), recurrent ANN, random forest (RF), XGBoost, etc. have successfully been applied with considerable improvement of model performance. Design and construction assisting software have been developed and adopted in practice of pile driving based on machine learning (Hazewinkel, 2022; VIKTOR, 2022). Despite various efforts, using ML to predict load transfer and base resistance, especially for long piles in soft soil is very limited due to the lack of appropriate data.

Momeni et al. (2015) was among few studies that established ANN model to predict the base resistance of piles, but their base and shaft resistance data were computed from dynamic test (PDA) results rather than direct measurement (i.e., static pile load tests). It is important to note that the results based on PDA are often found less reliable due to the use of multiple calibration parameters and the difference between dynamic (short-term) and static (long-term) pile behaviour (Chan et al., 1995; Budi et al., 2015). Similarly, Millán et al. (2023) made considerable effort to estimate tip resistance of piles embedded in rock using ANN, their data are obtained from numerical simulations that usually lack influencing factors from real construction condition. Further, past ML models could only address pre-casted and short piles with small data size, resulting in limited implications. Recent attempt (Huynh et al., 2023) was made to overcome these limitations, however, the model exhibits significant low quality of prediction using the basic ANN algorithm. More effort to apply advanced ML methods to predict complex contexts (e.g., long and large bored piles in soft soil) with bigger and more realistic database hence becomes crucial.

Past studies have used various sets of input features to develop ML model for predicting behaviour of pile foundation. As Fig. 1 shows, the features which have commonly been used as inputs for model training can be categorised into 2 major distinct groups, i.e.,

Group (i) incorporates response characteristics of soil-pile system captured through pile tests such as the PDA, the ultrasonic and static pile load tests. This can be the settlement and stress distribution behaviours of pile under static loading; the dynamic response of pile under impact (e.g., hammering) loads; the wave characteristics from non-destructive tests, among others. Of these, the use of loading-settlement/displacement (*P-S* or *t-z*) curves obtained from static pile load tests is the most preferable approach across various predictions due to its simplicity and reliability.

Group (ii) includes individual features of soil and pile without knowing soil-pile interaction detail. This can be divided into 2 different types:

- geometric features of piles (i.e., size and shape characteristics such as the diameter and length) and soils (e.g., the thickness of different soil layers);
- engineering features of piles and soils, specifically pile properties such as the axial stiffness, compression and tensile moduli; and soil properties such as shear strength, effective stress, cohesion, friction angle, characteristic data from field tests, e.g., SPT and CPT;

In the above, Group (*i*) are intrinsic features of soils and piles that are often known earlier in design stage, whereas Group (*ii*) employs test data which can only be obtained during and/or after construction phase through in-situ tests. Therefore, Group (*i*) reflects the true soil-pile interaction and real work condition of pile foundation, they are thus more reliable and provides more direct impact. On the other hand, Group (*ii*) is very crucial to understand pile foundation at early stage of design. The use of input variables from Group (*ii*) alone has been used widely to predict pile behaviour such as the ultimate bearing capacity and settlement (Kardani et al., 2020; Amjad et al., 2022; Cao et al., 2022). On the other hand, combining variables from Groups (*i*) and (*ii*) to form input dataset has also been used effectively to promote prediction outcomes (Momeni et al., 2014; Shahin, 2014; Pooya Nejad & Jaksa, 2017). Despite this distinct difference, the influence that using different sets of input features can have on ML-based prediction has not been addressed in past studies.

In view of the above, this paper aims to assess the role of using soil-pile interaction knowledge in prediction of pile foundation based on machine learning techniques. Specifically, the study compares prediction performance of ML models which are developed using different sets of input features, i.e., with and without soil-pile interaction detail to predict base resistance of piles. Different advancements of ensemble learning, i.e., from the most basic form Decision Trees to the highly rated XGBoost and CatBoost models are applied to a database (1129 datapoints) collected from 37 different real-life construction projects. Two distinct sets of input features serving different purposes of design and construction stages are investigated, while post-analysis is carried out to enhance our understanding of load-transfer mechanism in pile foundation. Furthermore, the developed models are examined by applying them to an independent dataset collected from literature, followed by performance evaluation and proposed solutions to improve prediction.

Table 1
Summary of past machine learning (ML) models used to predict behaviour of pile foundations, indicating the lack of attention to tip/base resistance which is one of two key components of pile's bearing capacity. Most of data to develop ML models for tip/base resistance are obtained from either numerical simulations or indirect measurements in the field.

No	Major categories of application/ prediction	Input variables/features and Classification	Measured data for shaft (skin) and base/tip resistance	Algorithm and Approach	References
1	Axial bearing capacity	*Group (ii)*: Pile features (L 4.7–33.6 m, D 0.38–1.2 m); *Group (i)*: Hammering parameters, PDA test results	Not available	ANN optimised by genetic algorithm (GA);	(Momeni et al., 2014)
		Group (ii): Pile features (L 5.5–45.1 m, D 0.25–0.9 m); CPT data of side and base soils	Not available	Basic ANN without optimisation	(Shahin, 2010)
		Group (ii): Pile features (L 5–10 m, D 0.3–0.4 m); SPT of side and base soils;	Not available	Heterogeneous Ensemble method combining multivariable adaptive regression splines (MARS) and radial basis neural network (RBNN) and equilibrium optimiser (EO)	(Cao et al., 2022)
		Group (ii): Pile features (L 3–47 m, D 0.1–0.9 m); shear resistance angle of side and base soils; effective stress of base soil;	Not available	6 different algorithms: XGBoost, Random Forest (RF); Decision Tree (DT); K-nearest Neighbour (KNN); Multilayer Perception Artificial Neural Network (MLPANN); support vector machine (SVM);	(Kardani et al., 2020)
		Group (ii): Pile features (cross-sectional area, L = 14–98 m, material properties); cohesion and friction angle of soils, *Group (ii)*: pile-soil friction;	Not available	10 different algorithms: Deep Neural Network (DNN); Extreme Learning Machine (ELM); Random Forest (RF); LASSO; KRidge; Stepwise Regression; Generic Programming (GP)	(Benbouras et al., 2021)
		Group (ii): Pile features (L 5–25 m, D < 0.4 m); SPT of side and base soils	Not available	ANN and RF (other investigations used Deep NN for the same data and objective are not presented in this stable)	(Pham et al., 2020)
		Group (ii): PHC nodular piles: pile feature (L < 50 m) and SPT of soil	Not available	Bayesian Regularized ANN	(Nguyen et al., 2022)
		Group (ii): Corroded piles: pile feature (L 30–60 m, D 1.2–2 m), corrosion depth, spalling thickness	Not available	Kernel ridge regression; Multilayer perceptron (validation with FEM results)	(Yousheng et al., 2024)
2	Side/skin friction/resistance factor α	*Group (i)*: Using PDA test data	Not available	Basic ANN	(Chan et al., 1995)
		Group (ii): Drilled shafts (L 1.62–77 m, D 0.18–1.8 m), only 2 input variables: undrained shear strength and effective vertical stress of side soil.	Not available	Bayesian Regularized ANN	(Goh et al., 2005)
3	Base/tip resistances Base/tip resistances	*Group (ii)*: Pile features (L5–35 m, D 0.38–1.2 m); and *Group (i)*: PDA data, driving piles.	Not available (indirect estimate from PDA data)	Basic ANN	(Momeni et al., 2015)
		Group (ii): Pile features (pile width, embedded lengths in rock and soil), and rock properties (self-weight, compressive strength, rock mass quality index)	Not available (numerical data, not physical measurement)	Normal ANN: 1 hidden layer with 8 neurons (applied to predict tip/base resistance of piles in rock)	(Millán et al., 2023)
		Group (ii): Pile features (L40–99 m, D 1–2.5 m), SPT of soil, and *Group (i)*: pile head displacement from static load tests	Yes (physical measurement)	Basic (normal) ANN without optimisation; the average prediction accuracy R2 = 0.85; no consideration of input features	(Huynh et al., 2023)
4	Lateral bearing capacity	*Group (ii)*: Pile features (L and D, not detailed); eccentricity of load; shear strength of soil	Not available	Least square support vector machine (LSSVM) and multivariate adaptive regression spline (MARS)	(Samui & Kim, 2013)
			Not available	Extreme learning machine (ELM) compared to optimised ANN	(Muduli et al., 2013)
			Not available	Multivariate adaptive regression splines (MARS) and Functional Networks (FN)	(Das & Suman, 2015)
5	Ultimate vertical displacement	*Group (ii)*: Pile features (L 8–81 m); SPT of soil and the applied load;	Not available	ANN; 2–4 hidden layers; R2 (validation) > 0.93;	(Pooya Nejad et al., 2009);
6	Load-settlement curves/ behaviour	*Group (ii)*: Pile features (L 5.5–56 m, D 0.3–1.29 m); and CPT data of soil (tip and shaft resistances)	Not available	Recurrent neural network (RNN)	(Shahin, 2014);
			Not available	Feed-forward neural network (FFNN)	(Moayedi & Hayati, 2018)
			Not available	Normal ANN;	(Pooya Nejad & Jaksa, 2017)
7	Pile-soil interaction: spatial displacement	*Group (ii)*: Pile features; *Group (i)*: pile-soil interaction characteristics; load transfer governing equations	Not available	Physics-informed neural network (PINN) to solve partial differential equation for soil-pile interaction: vertical and lateral displacements	Vahab et al. (2023); Ouyang et al. (2024)
8	Drive time of pile driving	*Group (ii):* Pile features (geometric measurements) and CPT data of soil	Not available	Basis ANN	(VIKTOR, 2022)
9	Driveability of vibratory piles	*Group (i):* Using existing pile driving data from industry practice	Not available	multilayer perceptron neural network (MLPNN); radial basis function neural network (RBFNN)	(Hazewinkel, 2022)
10	Pile damage in construction	*Group (i):* Wave signal from the in-hole multipoint traveling wave decomposition (MPTWD)	Not available	logistic regression (LR), XGBoost and multilayer perceptron (MLP)	(Wu et al., 2023)

Note: L is the embedded length of pile; D is the equivalent diameter of pile; PDA: Pile Dynamic Analysis; SPT: Standard Penetration Test; CPT: Cone Penetration Test; only studies with distinctive objectives, data, algorithms across literature are selected for this table; *Group (i)* features soil-pile interaction; while *Group* (*ii*) includes individual parameters of soil, pile and load.

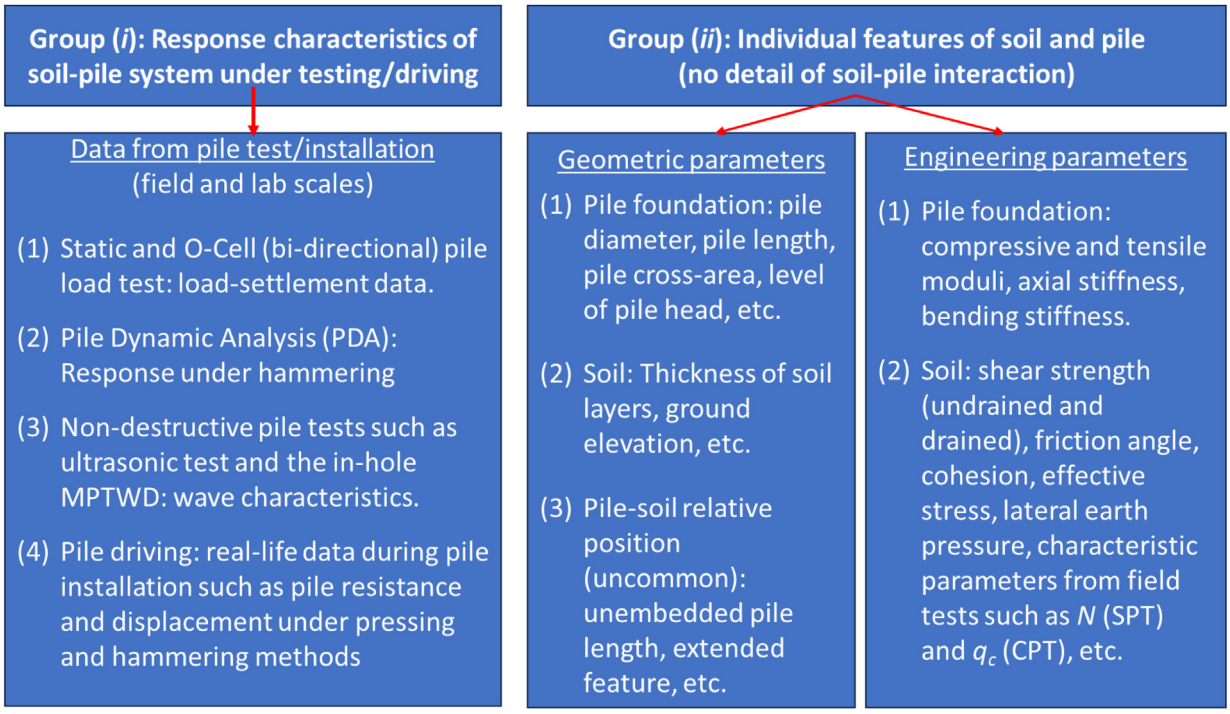


Fig. 1. Classification of features used for training machine learning models based on past ML studies and existing knowledge and practice of pile foundation.

The results bring considerable implications to the practice of using ML models for real-life prediction of pile foundation.

Data characteristics and processing

Site investigation and data collection

The current study adopted the data of in-situ static pile load tests collected through an extensive site investigation over the past 10 years in Mekong Delta (South of Vietnam) (Huynh et al., 2023). The investigation went through 37 different high-rise building projects located at very soft ground of Ho Chi Minh City, resulting in a database of pile load tests on 86 different long and large piles. These pile load tests were carried out with strain gauges installed along the piles to determine stress-strain response and load transfer over the depth. The base resistance of piles was calculated based on the applied load and shaft resistance obtained from the strain gauges through the pile tests (Nguyen & Fellenius, 2015; Huynh et al., 2023). The piles with an equivalent diameter ranging from 1.0 to 2.5 m and a length up to nearly 100 m were installed in complex geological strata including considerable layers of very soft soil (e.g., 35 m thick layer of soil with SPT < 10). The SPT of soil only exceeds 50 at a depth > 70 m (dense sand) which is also the common required depth of pile embedment, making very large and long piles often used for high-rise buildings and infrastructure systems in this region. Undoubtedly, these piles are considerably larger and longer than those used in many previous machine learning studies (Shahin, 2014; Alkroosh et al., 2015; Baghbani et al., 2022; Cao et al., 2022), where piles were shorter than 40 m, resulting in very different behaviour of load transfer and base resistance.

Evaluation of influencing features and formation of input datasets

Previous investigation (Huynh et al., 2023) have identified 5 key input parameters that can be used to predict base resistance of pile foundation. They are: (*i*) the applied load *P* (tonnes), (*ii*) the displacement of loading point d_p, (mm), (*iii*) the axial stiffness *AE* (kNm/m) where *A* is the cross-sectional area of pile (m^2), and *E* is the equivalent Young's modulus (i.e., *E* = 36 GPa), (*iv*) the average SPT values *N* of the soil within 3D beneath the pile toe and, (*v*) the distance from the loading point to the pile toe L_p (m). It is noteworthy that for O–Cell tests which are often used when piles are very long (*L* > 60 m), the loading point usually positions close to the pile's toe (e.g., 10–15 m away), making L_p shorter than L. Further details for this load test with long piles can be found in past site studies (Fellenius & Nguyen, 2013; Nguyen & Fellenius, 2014; Huynh et al., 2023). Among these features, the correspondence between the displacement of loading point d_p, and the applied load *P* can only be obtained from pile load tests, it is thus well representative for soil-pile interaction knowledge, i.e., Group *(ii)*. Compared to other studies (Pooya Nejad & Jaksa, 2017; Cao et al., 2022) where the number of input features can even exceed 10 parameters, this set of 5 input features is relatively simpler and easy to obtain from pile test results, it was hence adopted for machine learning development in this current study. For convenient comparison with other sets of input features discussed later in this paper, this feature set is termed as **Set I** (see Fig. 2), which is a typical case of data Group (*i*) as represented earlier in Fig. 1.

While **Set I** above includes the displacement of pile head as one of the primary features for model input, its major limitation is that in many real-life situations, the prediction is required without knowing the settlement of piles. For example, it is common in practice that the base resistance is estimated when only the soil properties and pile features are known. Therefore, it is important to establish a machine learning model to predict base resistance for these contexts. For this purpose, another set of input features, termed as **Set II** (corresponding to data Group (*ii*) defined earlier in Fig. 1), is proposed. This includes: (*i*) the applied load *P* (tonnes), (*ii*) the distance from the loading point to the pile toe L_p (m), (*iii*) the axial stiffness *AE* (kNm/m), (*iv*) the shear strength S_{ub} (kPa) of the soil beneath the pile toe; and (*v*) to (*vi*) the averaged shear strengths, i.e., S_{u1}, S_{u2} and S_{u3} (kPa) of the soils positioning within the first, second and third 1/3 of the pile's length from the head (see Fig. 3).

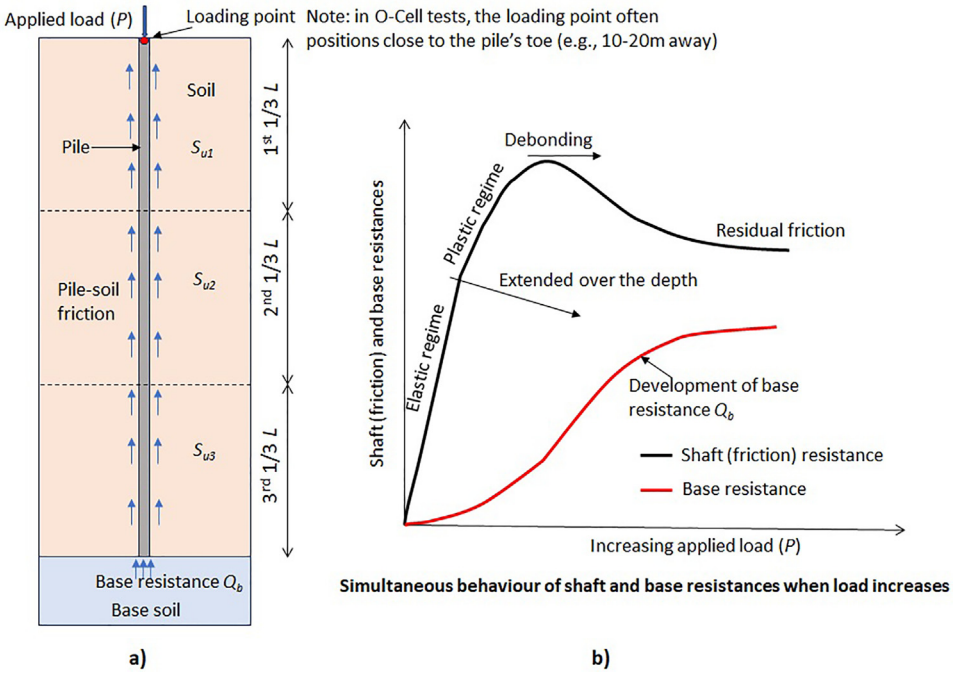


Fig. 2. Overall understanding of pile load test: a) defined input features; and b) simultaneous development of shaft and base resistances of pile under increasing applied load.

Considering SPT and shear strength of soils as possible representative indices for shaft resistance, the undrained shear strength was found as the most relevant option because it is more reliable and can be determined based on various laboratory and filed tests. Data Set I and Set II are provided in Supplementary Material S1 and S2 included in this paper. It is also worth noting that data Set I and Set II are specific representatives for the aforementioned Group (*i*) and Group (*ii*) (see Fig. 1), which are broader categories and can include different data sets depending on prediction contexts and soil-pile interaction tests

Violin (Seaborn) plots, which can effectively demonstrate the distribution characteristics of data, are used to represent the two datasets as shown in Fig. 3. The results show 3 distinct different types of data distribution among the investigated features.

Type 1: End-skewed distribution: The applied load (P), the displacement of loading point and the base resistance have relatively the same distribution form in which the majority of datapoints are located at one end of the data range. Specifically, most of the applied load are under 3500 tonnes, the displacement values are less than 50 mm, while the base resistance is lower than 4000 tonnes. This indicates highly non-uniform distributions of data over their ranges.

Tye 2: Normal distribution: On the other hand, the axial stiffness, the SPT value and shear strength of soil at deep layers (i.e., 2/3 of the piles downward) are distributed mainly at the middle of the data range, despite some fluctuations along the range. This is understandable as the soils at deeper layers are often more homogeneous with thicker layer and consistent properties.

Type 3: Bone-shaped distribution: The distance from the loading point to the pile toe (L_p) is a typical example of this type. The datapoints concentrate equally at both ends of the range, whereas there are minimal datapoints at the middle, i.e., 40–50 mm of the range. Similar type of distribution is also observed with shear strength of the soil within the first 1/3 depth of piles, where there are complex variations of different soil types.

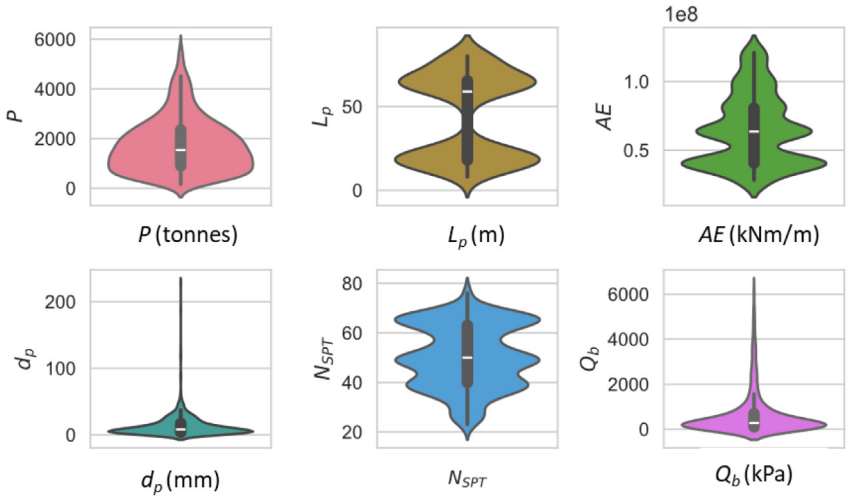
Fig. 4 represents the correlation matrix among different features of piles in Set I and Set II. In this presentation, the closer the value to zero, the less the connection between two variables. On the other hand, when the correlation coefficient of two variables is nearer to unity, this means a stronger interaction between the 2 variables. Fig. 4a (Set I) shows that the displacement of loading point d_p and the applied load are the two key features giving prominent impacts on the base resistance of piles. On the other hand, the SPT of base soil and the distance from the loading point to the pile tip induce the less influence on the base resistance of piles. In short, this figure confirms that the 5 selected features in Set I can be used to build a ML model to predict base resistance as they can all influence the behaviour of base resistance of piles. Similarly, Fig. 4b represents impact scores among different features used in Set II. In this case, the applied load P becomes the primary impact on the predicted Q_b as it correlation score is much larger than those induced by other variables.

Model building: selection of algorithms, training-validation process, and optimisation

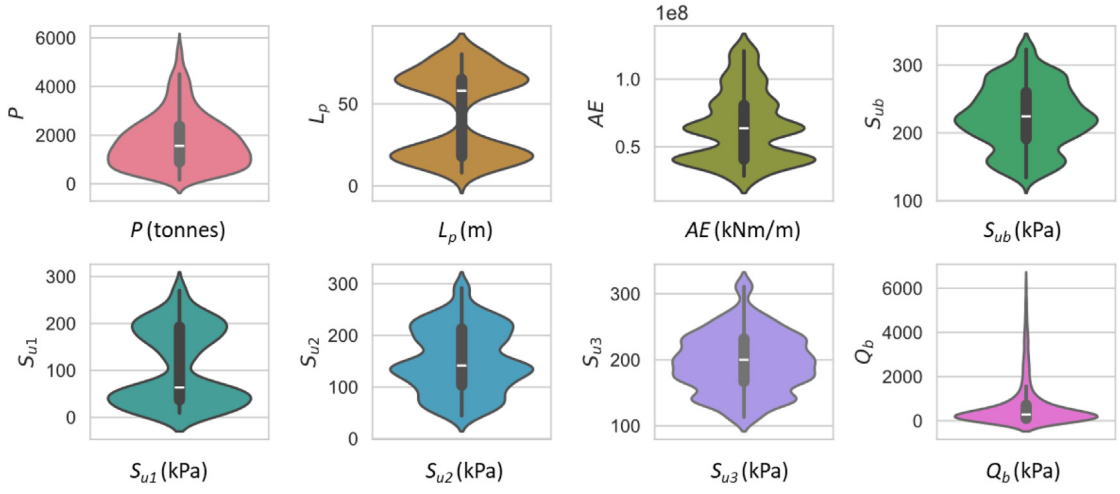
Selection of machine learning algorithms for model training

In this study, we employed a variety of ensemble learning methods to train the models, followed by a comprehensive assessment of their performance. Given their inherent low computational cost, ease of implementation, and robust performance, algorithms based on decision trees were the primary focus in our analysis. Specifically, six distinct machine learning techniques were adopted: (*i*) Decision Trees (DT), (*ii*) Random Forest (RF), (*iii*) Gradient Boosting Regressor (GBR), (*iv*) Light Gradient Boosting Machine (LGBM), (*v*) CatBoost, and (*vi*) XGBoost.

The Decision Trees (DT) method serves as the fundamental building blocks for more advanced ensemble methods to grow. Decision trees are constructed using a hierarchical structure, which splits data based on feature values to make predictions (Quinlan, 1986). On the other hand, Random Forest (RF), an extension of decision trees DT, leverages the bagging (bootstrap aggregating) technique to enhance model stability and accuracy. The current study also adopted 4 representatives, i.e., GBR, LGBM, CatBoost, and XGBoost of boosting algorithms, which iteratively improve model performance by focusing on errors from previous iterations. These methods are designed to correct the mistakes of weak learners by combining them into a strong ensemble model. Detailed algorithms and distinct features of these models have been well described in past studies (Hancock & Khoshgoftaar, 2020; Nguyen et al., 2023),



a)



b)

Fig. 3. Violin (Seaborn) plots for all input features used in 2 data sets: a) **Set I**; and b) **Set II**.

they are thus not given in this paper. By selecting these hierarchy and diverse models, our study aims to provide a thorough evaluation of the decision trees family, from the basic form DT to advanced versions like CatBoost and XGBoost.

Training, testing and optimisation procedures

The database was partitioned into training and testing sets, adhering to an 80–20 % distribution. This segregation was executed to uphold a robust training and evaluation framework, ensuring that the model's performance was assessed on unseen data. The training dataset was subject to 10-fold cross-validation to ascertain model stability and predictive performance. Throughout this process, the training dataset was subdivided into 10 folds, where the model used nine folds for training and the rest for cross-validation. This process was cycled through each fold to guarantee each subset was utilized for validation once, thus equal treatment of datapoints. The outcomes from cross-validation were assessed in parallel with the prediction of testing data, further reinforcing the rigor of assessment method.

For optimizing the efficiency of model training, the current study adopted Bayesian Optimization (BO) associated with Gaussian Process (GP). The BO was used to adeptly tune the hyperparameters of the machine learning models, providing a systematic and principled approach to model optimization. This method leverages a probabilistic model to map the relationship between hyperparameters and the objective function, aspiring to find the minimal (or maximal) value of the objective function (the mean squared error MSE) with computational efficiency. Further, GP was utilized as a surrogate model to represent the objective function, approximating the distribution over functions to identify the potential regions that are likely to yield enhanced model performance. As demonstrated in Fig. 5, the true objective function, $f(x)$ (red dotted line), represents the MSE across different hyperparameter configurations. Initial sample points (red dots) are utilized by a GP to predict the MSE across the hyperparameter space, illustrated by the blue line and shaded region representing the mean $\mu(x)$ and uncertainty, respectively. The Expected Improvement (EI) acquisition function (green dashed line) is derived from the GP and determines the subsequent sampling point (blue cross) by identifying the location that minimizes the EI. This point will be evaluated and incorporated into the next iteration, refining the GP and progressively guiding the search towards the optimum value.

In this study, performance of models was assessed through 4 different indices, including the coefficient of determination (R^2), root mean-

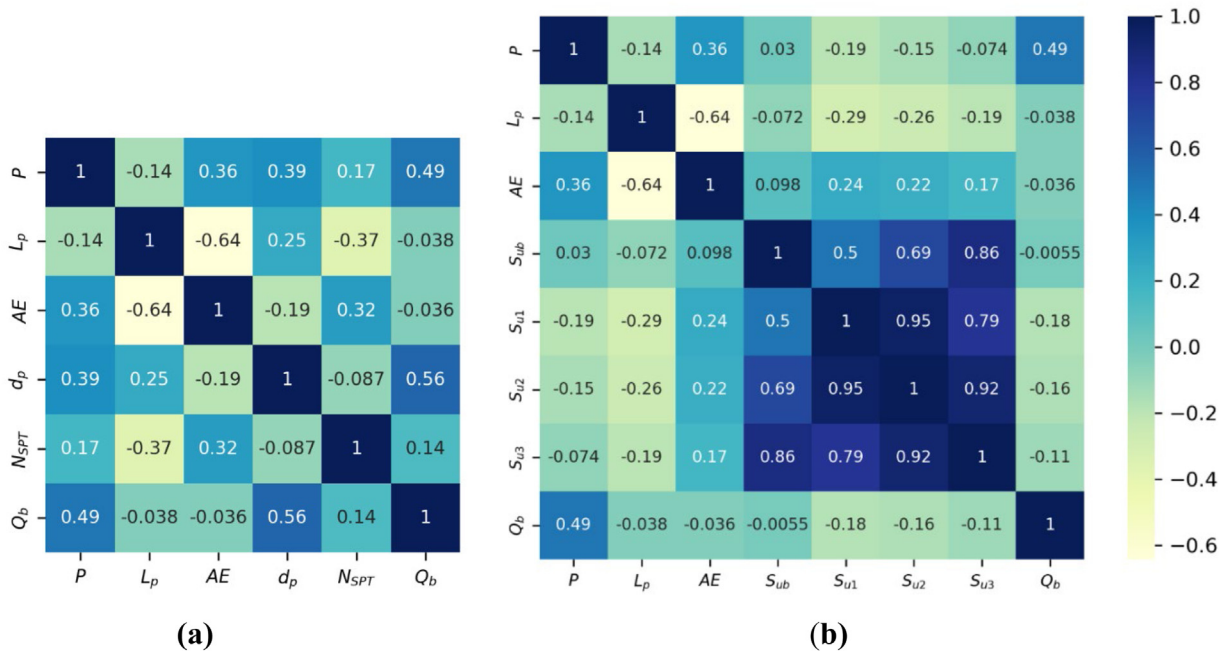


Fig. 4. Correlation matrix of input features in: (a) Set I; and (b) Set II.

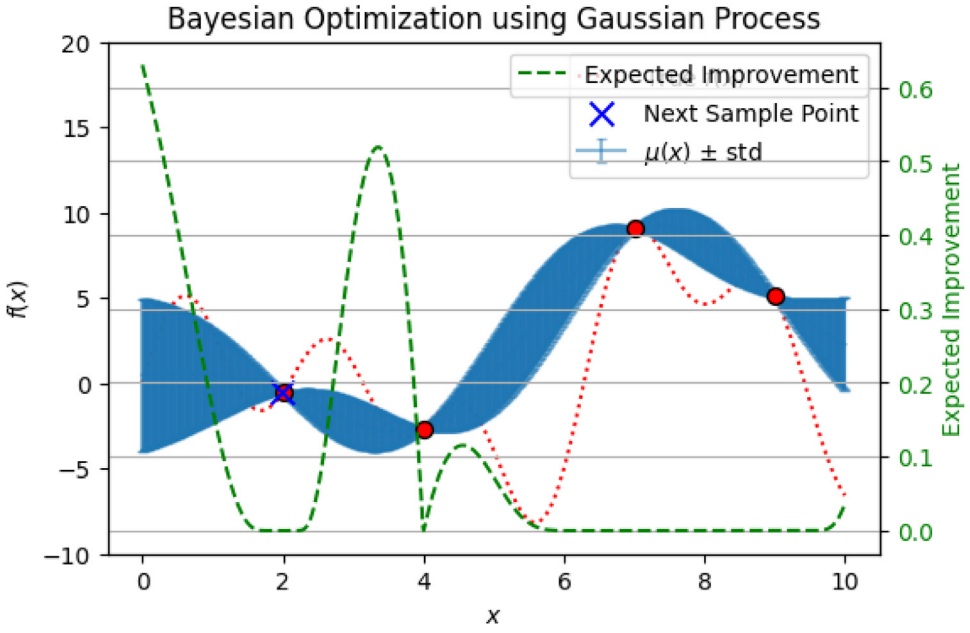


Fig. 5. Performance of Bayesian Optimization assisted by Gaussian Process in hyperparameter and objective function (f(x)) space where μ(x) is the mean value.
Note: $\mu(x)$ is the mean value.

squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). As these indices have been used widely in various machine learning studies, their calculations and purposes are not explained in detail.

Effectiveness of Bayesian-Gaussian coupling optimization

The efficiency of Bayesian-Gaussian (BG) coupling optimization technique is examined using the results based on data Set I. It is important to note that 3 different cases were examined, i.e.,

(i) Training with 10-fold cross validation (notated by "train CV"): In the training subset (80 % of the entire dataset), the data continued to be divided into 10 different sets in which 9 were used for training and the other for testing (cross-validation). This training-testing cycles over 10 subsets of data resulted in the ranges of prediction outcomes, including min, max and average (mean) values that are considered.

(ii) Testing the trained models by applying them to predict the remaining 20 % of dataset which was independently separated from the training dataset (notated by "test").

(iii) Testing the trained model for the entire data, i.e., the combined training and testing sets (notated by "all").

The results from the optimized models (Fig. 6) show that performances of all the boosting models have been improved significantly. Specifically, the values of R^2 in the testing phase by all boosting models become greater than 0.88 when the optimization was applied. For example, the value of R^2 by XGBoost algorithm increases significantly from 0.85 for the default configuration to 0.9 for the optimised models. The

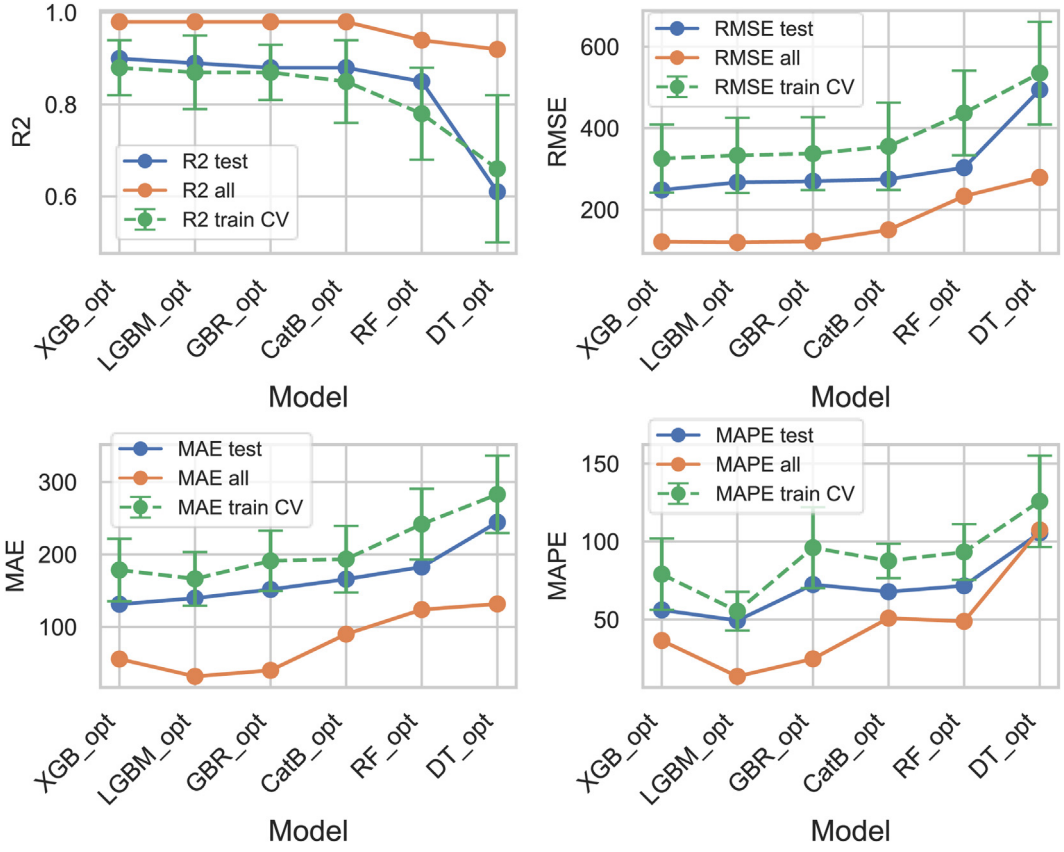


Fig. 6. Performances of the six ML models using optimized (tuned) hyperparameters.

prediction errors also decrease, for example, RMSE reduced from 305 to approximately 250 thanks to the optimization. On the other hand, Decision Trees (DT) and Random Forest (RF) do not show considerable improvements in their prediction as their R^2 values remain relatively the same after optimization. It is also interesting to note that the optimization has resolved the over-training issue with DT when the values of error indices such as the RMSE and MAE become more relevant to the corresponding poor level of R^2 (only 0.6). Compared to the boosting algorithms, the DT and RF resulted in considerably lower R^2. This indicated that the current optimization technique more suitable for boosting algorithms, whereas it made insignificant impact on the performance of RF and DT. Therefore, the 4 boosting algorithms, i.e., XGBoost, CatBoost, GBR and LGBM were selected for further investigations on the influence that using different sets of input features can have on the prediction performance.

Model performance considering data randomness aided by Monte Carlo

The influence of data homogeneity and randomness is investigated in this section based on Monte Carlo technique. The division of dataset into training and testing subsets was randomly repeated 300 times. Their results (for testing only) are represented in Violin (Seaborn) plots as shown in Fig. 7 using default and optimized hyperparameters. The major advantage of this type of presentation compared to the boxplot is that the concentration of datapoints can be quantified over the size and shape of Violin. The wider and shorter the Violin, the more concentration the datapoints, thus greater confidence the prediction. Fig. 7 shows that the optimized GBR and XGBoost models are on the top of others because their Violin shapes are shorter and wider. The performance indices changed in a wide range over different divisions, for example, R^2 of GBR model without optimisation can vary significantly from around 0.69 to 0.96, while its optimized version result in significantly narrower range, i.e., from 0.8 to 0.96. In fact, advanced boosting algorithms such as CatBoost and XGBoost are less influenced by data randomness as their variations in prediction (R^2 and error indices) were considerably smaller. For example, R^2 of the optimized XGBoost varies from 0.84 to 0.95 approximately, while for CatBoost model, the values of R^2 for both default and optimized configurations consistently change in the range from 0.8 to 0.97. The results also indicated that the optimization has brought significant positive effect on the model performance as it did not only enhance the magnitude of performance metrics, but also narrow their variations due to data randomness, thus more reliable predictions.

Model performance using different input features: set I vs set II

By applying the BG technique for optimizing model training while adopting different data sets of input features, i.e., Set I and Set II, the prediction outcomes are compared to explore the influence of input features on prediction performance. Note that only boosting algorithms, which have been marked in topmost performing models in earlier investigation, were employed for this purpose.

Prediction outcomes from the best optimized models using input set I

Fig. 8 represents predictions of pile's base resistance given by the best trained models based on the 4 boosting algorithms, i.e., XGBoost, CatBoost, GBR and LGBM. The boundaries of 10 % and 20 % prediction errors are set to facilitate the assessment. The results show that XGBoost resulted in a larger concentration of predicted points around the Best-Line (i.e., the equal line between the actual and predicted data, represented by the black dash line) for both training and testing compared to 3 other methods. On the other hand, there are more predicted points that are outside the 20 % Error Line, representing less reliable performance in CatBoost, LGBM and GBR models. Most of datapoints are

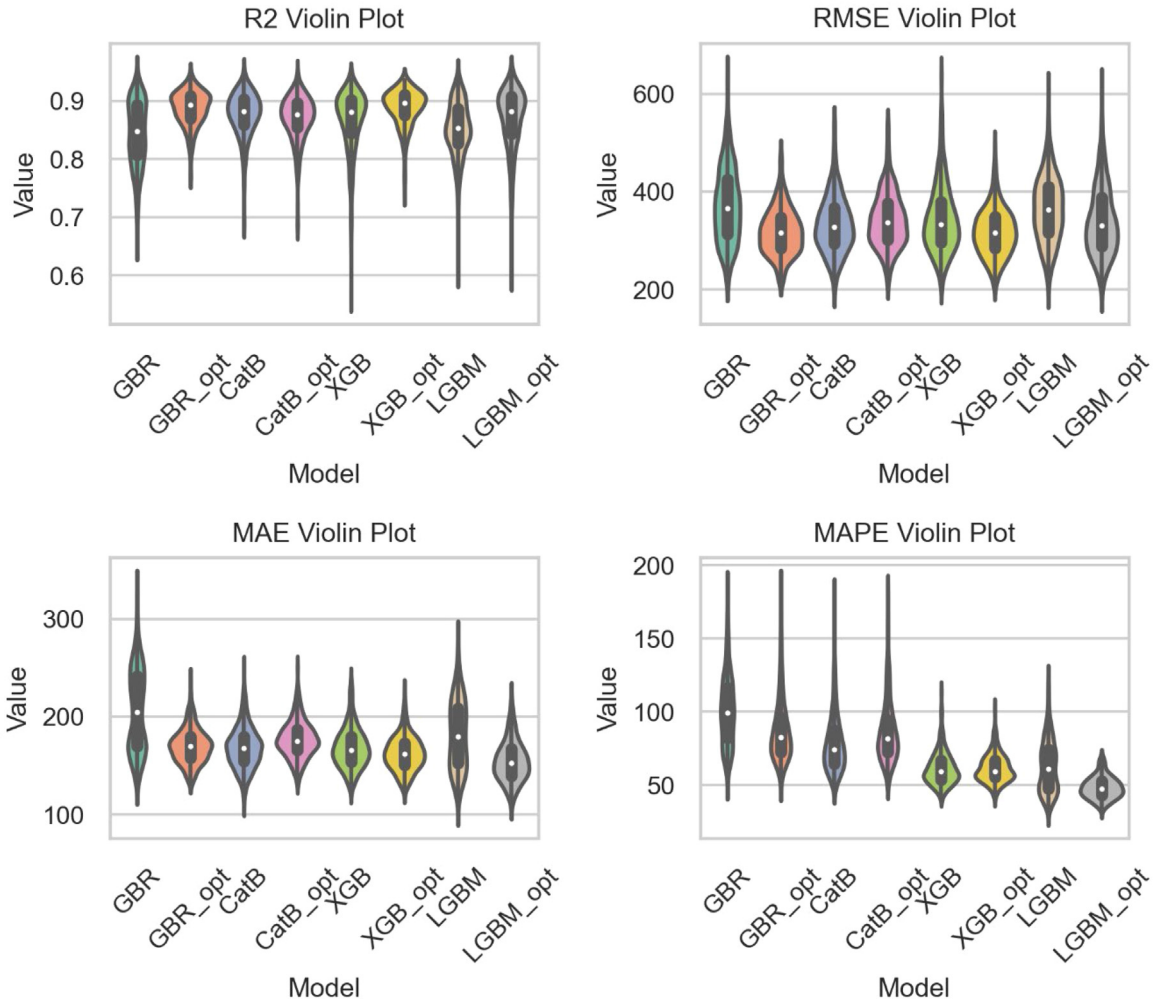


Fig. 7. Violin plots to compare performance of models with and without optimized hyperparameters (300 simulations).

located within the range of less than 4000 kPa, especially < 2000 kPa. This means that there is a limitation in the examination and developed models when the value of base resistance becomes larger than 4000 kPa. In practical design, the base resistance of pile foundations often accounts for up to 30 % of the total bearing capacity with its magnitude that is normally under 4000 kPa in many cases (Nguyen et al., 2024).

Prediction outcomes from optimised models using input set II

Interestingly, the results from the optimized models developed based on data Set II agree with the earlier findings that the 4 boosting algorithms outperform the bagging methods, i.e., Random Forest (RF). Fig. 9 show that the average R^2 by the RF model is only around 0.8 with a wide distribution (min = 0.6 and max = 0.9), while other boosting algorithms lead to an average $R^2 > 0.85$. In fact, the values of mean R^2 by the GBR and CatBoost are the largest, i.e., 0.9 with a narrow range from 0.8 to 0.95, proving them as the 2 best predicting models among the 4 investigated boosting algorithms. It is also noted that despite having the average R^2 of around 0.85, the LGBM model resulted in an extremely large variation of R^2 (from approx. 0.5 to 0.9), thus very low reliability. On the other hand, the Random Forest produced the largest errors across different forms, i.e., RMSE, MAE and MAPE, while GBR and CatBoost consistently resulted in the smallest errors. In short, all the performance indices, including R^2 and different error matrices proved the more accurate and reliable predictions by boosting algorithms, especially the GBR and CatBoost models using the input feature Set II.

Fig. 10 shows prediction outcomes by the best models (mean $R^2 = 0.95$ for testing) which were trained using the most accurate and reliable algorithm, i.e., CatBoost. For comparison, the results by XGBoost algorithm (mean $R^2 = 0.93$ for testing), which has shown the best prediction in feature data Set I (see Figs. 6 and 7) is also presented. The figures show that show majority of the predicted values are within the 20 % deviation from the actual base resistance.

Comparison of model performance by input set I and set II

Using prediction results in the testing phase for 300 random data divisions across 5 different models (i.e., 4 boosting and 1 bagging RF algorithms), Taylor' diagram is plotted to compare performances of different ML models based on input feature Set I and Set II (Fig. 11). In this presentation, the correlation coefficient R^2 and standard deviation by different models are collated together into the same circular diagram where the size of the radius represents the scale of standard deviations while the sector changes with the correlation value. In this diagram, the standard deviations of predictions by different models are normalized with the largest value, i.e., std_i/std_{max} where std denotes the standard deviation, so the biggest radius is equivalent to unity as shown in Fig. 11. According to this diagram, a point which is closer to the centre (i.e., the smaller the standard deviation) and the horizontal line (the larger the R^2) indicates a better prediction (more accurate and reliable). The results show that the models trained based on data Set I have an average better prediction compared to those based on data Set II as they posi-

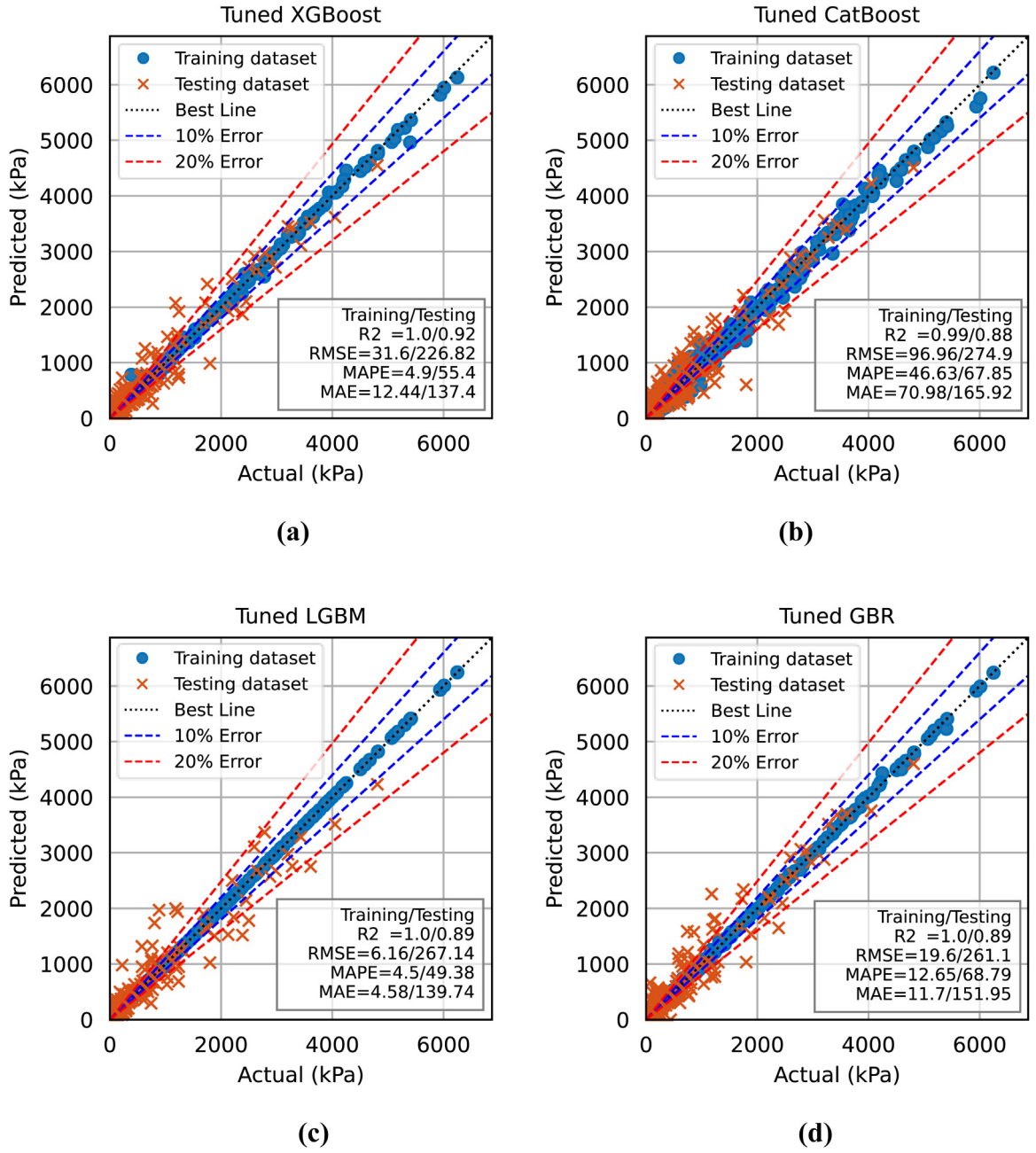


Fig. 8. Predicted vs actual values of base resistance using best models built on data Set I by (a) XGBoost,(b) CatBoost, (c) LGBM and (d) GBR.

tion closer to the centre point with an average larger value of R^2. The 2 best models that can be identified are GBR and XGB based on data Set I, followed by GBR and CatBoost models using data Set II.

Understanding of feature impact

Influencing factors of features in data set I

SHAPley analysis was implemented to interpret the model outcomes based the best performing model XGBoost as shown in Fig. 12. The results indicated the predominant role of the displacement of loading point d_p on the behaviour of base resistance. The SHAP value by d_p exceeds 350 which is greater than the second-ranked feature, i.e., the applied load P by a factor of approximately 1.5. On the other hand, the axial stiffness (AE) and the SPT of base soil caused lesser impact on the base resistance. The results also show that while increasing d_p, P and the N_{SPT} can induce positive impact on the base resistance Q_b, (i.e., the larger the d_p, the larger the Q_b), an opposite response is obtained when rising L_p and the axial stiffness AE. These behaviours are understandable with reference to conventional knowledge of pile foundation (Poulos, 1989; Hirayama, 1990; Bohn et al., 2017). For example, it is often agreed that the base resistance Q_b increases in a hyperbolic form with greater displacement of the pile tip which is in fact proportional to the displacement of loading point used in the current study.

Influencing factors of features in data set II

Without having the displacement of loading point, the feature analysis based on Set II results show predominant impact of the applied load P on the base resistance of piles. The results based on XGBoost model are shown here for convenient comparison with the feature analysis of data Set I presented earlier. The XGBoost model's results show that mean

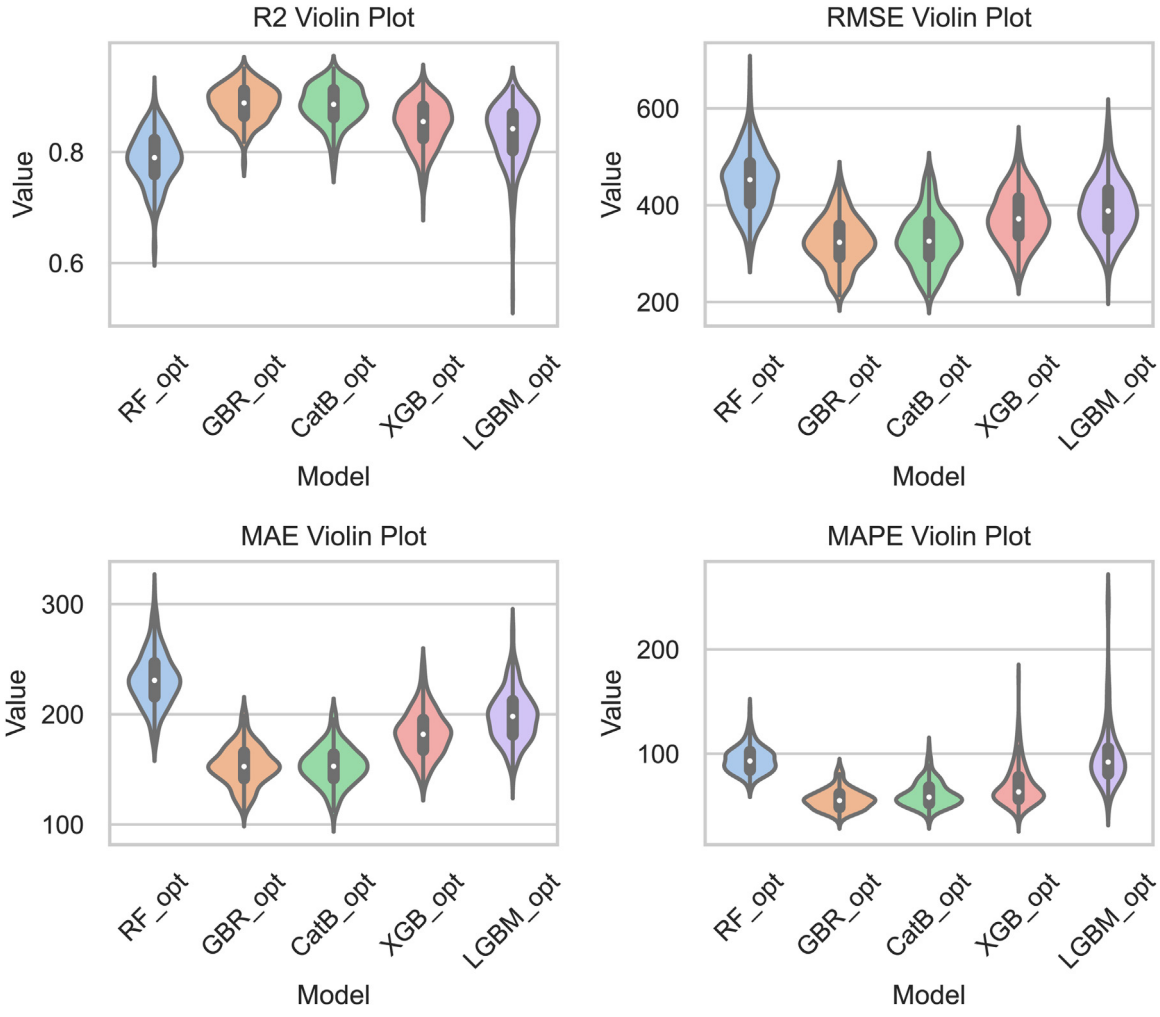


Fig. 9. Performance of optimized models using input Set II (testing phase).

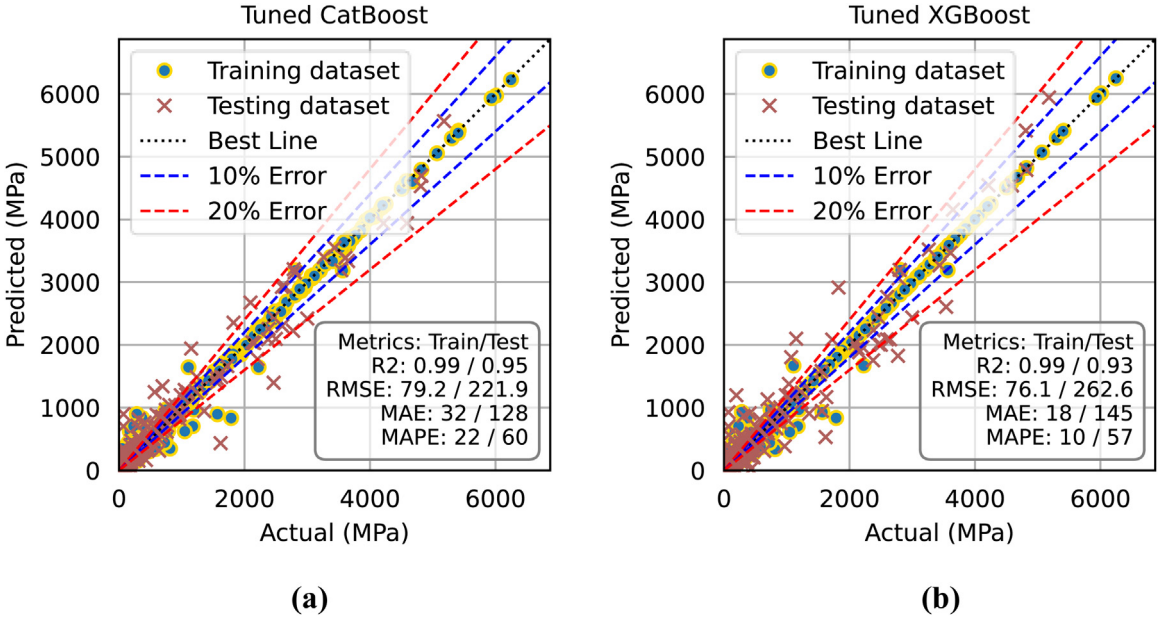


Fig. 10. Predicted vs actual values of base resistance using best models built on data Set II by: (a) CatBoost, and (b) XGBoost algorithms.

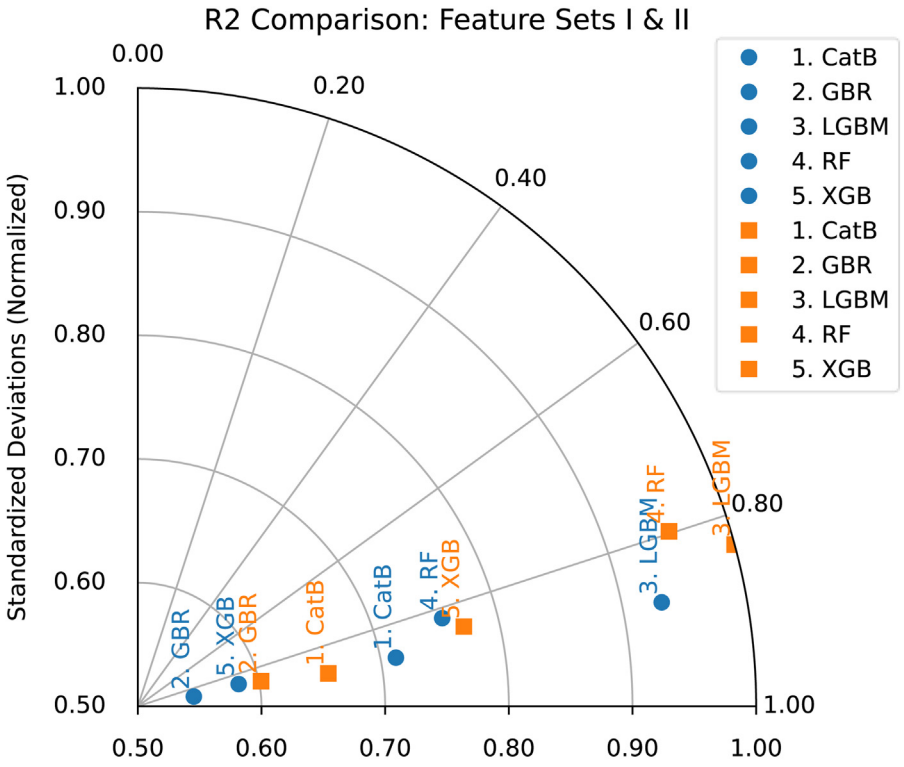


Fig. 11. Taylor's Chart comparing model performances using input feature Set I and Set II. Note: blue - Set I, and orange - Set II.

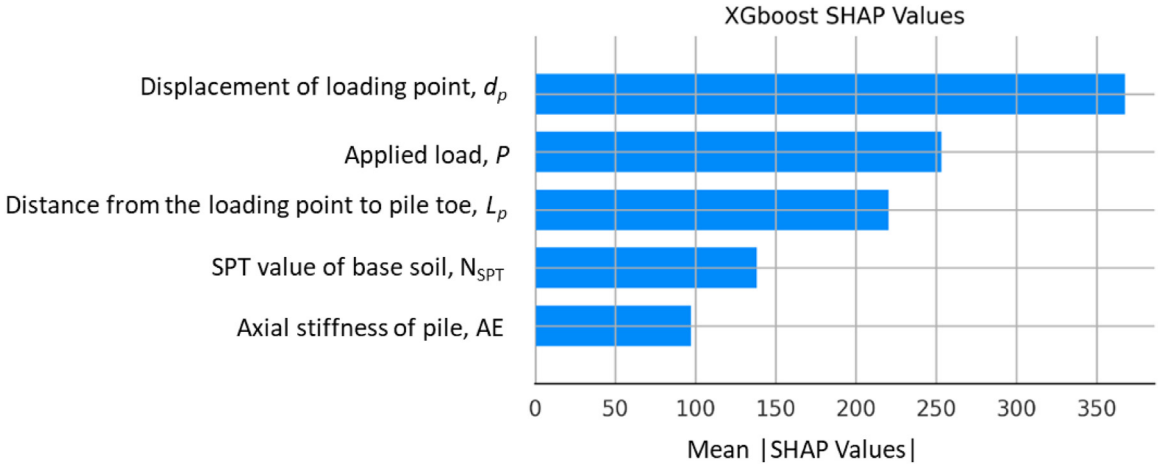


Fig. 12. Influence of input features (Set I) on the predicted base resistance (Q_b) by XGBoost model.

SHAP value of the applied load is about 530 which is 3 times larger than the second impactful feature, i.e., the distance from the loading point to pile toe L_p. In the predictions by CatBoost model, the decisive role of the applied load on the base resistance even becomes more predominant as the mean SHAP value by the applied load is 4 times larger than that by L_p. Although the applied load is well understood to have a direct impact on the base resistance of pile, as described through past mathematical formulae such as the hyperbolic equations (Hirayama, 1990; Bohn et al., 2017), the SHAP analysis enabled this impact to be measured based on real-life data of pile response. It is also interesting to observe that shear strength of soils made the least contribution among all secondary factors. Among different soils surrounding the pile, it is understandable that shear strength of the soil which directly contacts the pile toe influences most the base resistance. Analytical methods for determining base resistance of pile foundation often adopt the conventional Tezaghi's theory where the shear strength parameters of soil play a decisive role (Wrana, 2015).

It is also interesting when comparing the impact charts of input features Set I (Fig. 12) and Set II (Fig. 13). While the applied load P shows its outstanding influence on the predicted Q_b in the case of data Set II, the analysis based on the predictions from Set I indicates the driving role of the displacement d_p on the predicted outcome. The primary importance of P in Set I-based prediction drops to the second rank in Set II-based prediction with its mean SHAP value that is only approximately 2/3 of the value by the displacement d_p as shown in Fig. 12b This finding is crucial as it provides a quantitative testament

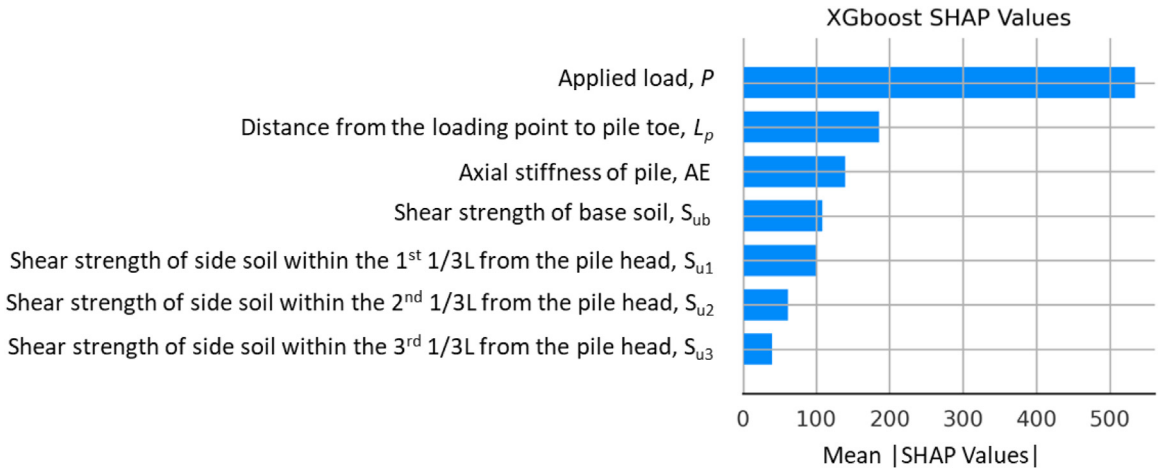


Fig. 13. Influence of input features (Set II) on the predicted base resistance (Q_b) by XGBoost model.

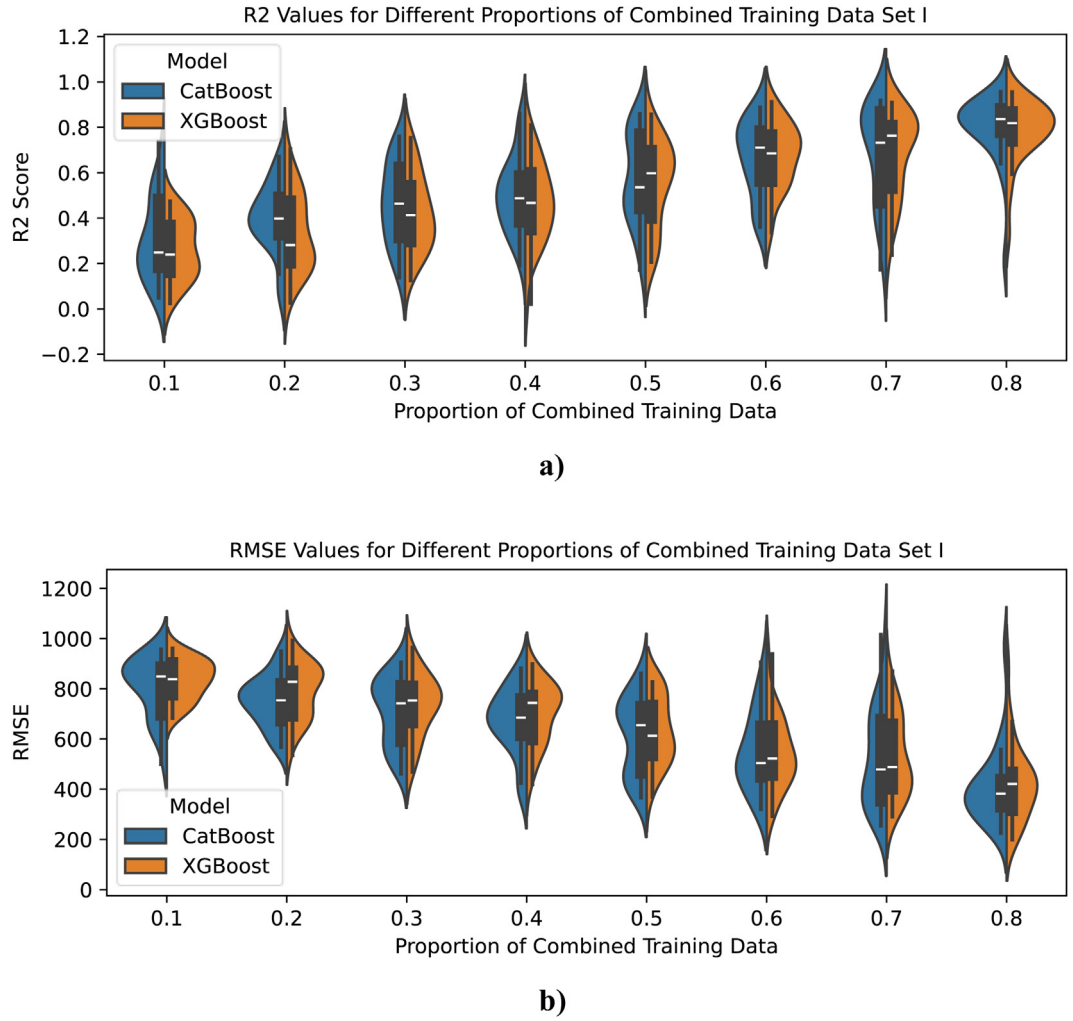


Fig. 14. Improved ML model through reinforced learning with updated training data using Set I: a) R^2; and b) RMSE.

to the critical role of considering d_p in predicting base resistance of pile foundation.

The results from feature impact analysis above bring significant implications to understanding load transfer mechanism as well as assisting designs. In practice of calculating base resistance, design guidelines often consider shear strength of soil underneath the pile toe (base soil) as the major parameter while ignoring the impact of soil resistances along the pile. For example, various national and professional practice codes (ECP202/4, 2005; AASHTO Specifications, 2010; TCVN 10304, 2014) only use SPT value of base soil for calculating base resistance. Fig. 13, on the other hand, shows that the shear strength of soils along the pile play a considerable role in the response of base resistance. For instance,

the outcome from XGBoost model shows that the mean SHAP value of S_{u1} (i.e., shear strength of soil within the first 1/3 pile length) is 100 point which is just slightly smaller than the largest score induced by the base soil. Meanwhile, the contributions of soils at deeper layer (i.e., S_{u2} and S_{u3}) decrease. This result reflects very well the load transfer mechanism of long piles where skin resistance develops downward with later and smaller contributions from deeper soils. It suggests the importance of considering soil resistance along the pile when calculating base resistance in design practice.

Performance of set I and set II-based models when predicting independent cases of long pile foundation: reinforced learning process

Three case studies of long pile foundation constructed in Mekong Delta (Ho Chi Minh City) reported by (Fellenius & Nguyen, 2013; Nguyen & Fellenius, 2014, 2015) were collected for validating the models developed based on data Set I and Set II (see Supplementary Material S3 and S4). The independent test piles selected for validation were 71–96 m long and subjected to O–Cell pile load test. The loading point positioned in the range 15–20 m from the pile toe, and the applied load increased up to 3000 tonnes, which is relevant to testing the developed model.

In practice of machine learning predictions, the training database is expected to be enriched continuously until achieving a desired confidence. Newly collected data are added to the existing database and the training process is repeated, thus the reinforced learning. Following this approach, the external independent dataset shown above were added to the original database for reinforced learning process. Specific steps can be described as follows:

- ***Step 1***: Fine tune the models hyperparameters of each machine learning models as presented in the previous section;
- ***Step 2***: Split the external validation data into 2 parts, one was combined with the original training data to retrain the models while the other was used for testing. The proportion of this division was varied gradually from 0.1 to 0.8 to examine the response of model performance;
- ***Step 3***: Use the optimized models (i.e., XGBoost & CatBoost) defined in Step 1 for retraining adopting new training data generated in Step 2;
- ***Step 4***: To ensure the repeatability of process, for each proportion of data split in Step 2, we run 30 simulations corresponding to 30 random splits of dataset using different random state parameters in the stratified split algorithm.

Fig. 14 shows how the performance of CatBoost and XGBoost models change with different proportions of data split, represented via R^2 and RMSE metrics. For only 10 % of the validation data that was included into the training process, both the models exhibit moderate R^2 scores and significant variability, among which, the CatBoost model shows slightly better performance. As the proportion increases to 20 % and 30 %, R^2 scores become larger with decreasing variability, indicating enhanced model performance. From 40 % to 60 % of inclusion, the R^2 scores stabilize with consistently high values, and CatBoost out-

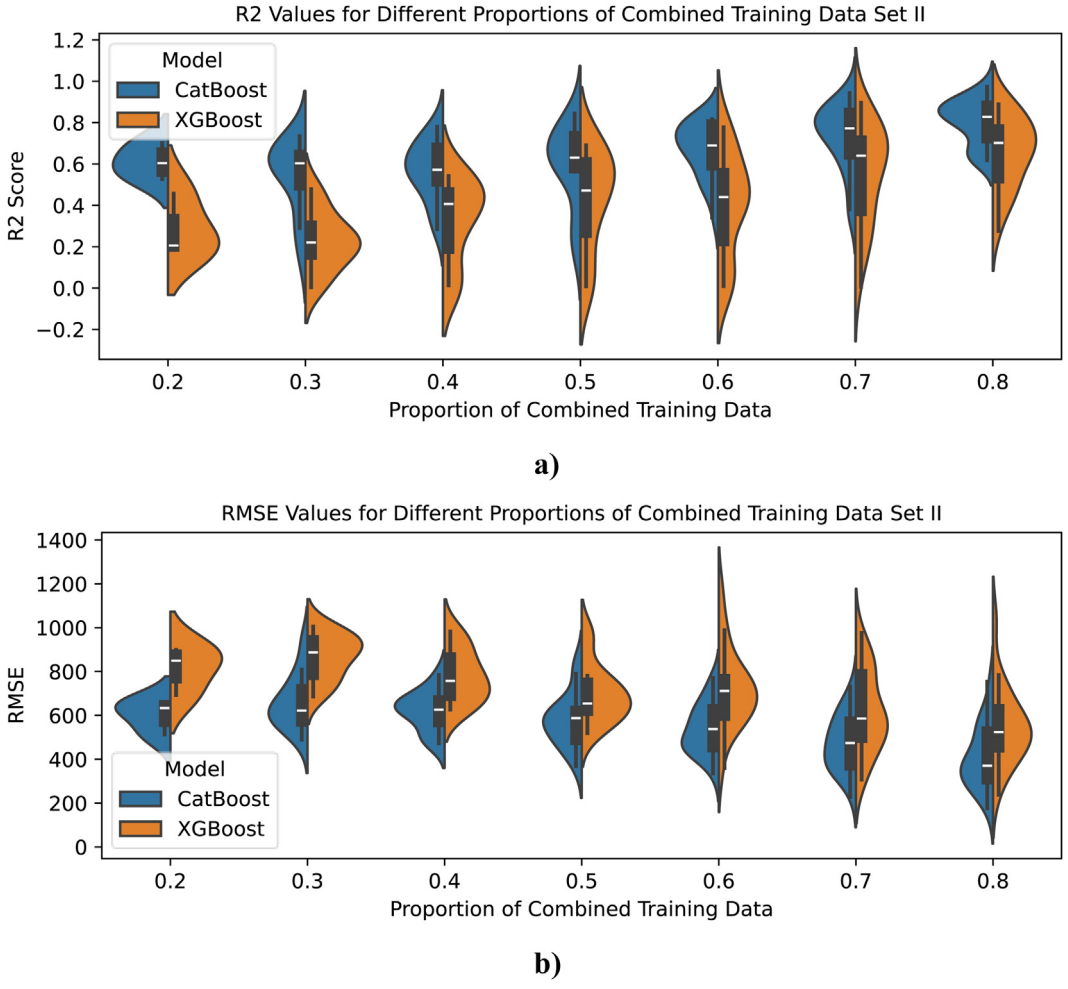


Fig. 15. Improved ML model through reinforced learning with updated training data using Set II: a) R^2; and b) RMSE.

performs XGBoost. At higher proportions of 70 % and 80 %, the R^2 scores undoubtedly reach their peak with tight distributions, suggesting robust and reliable model performance, with CatBoost maintaining a slight edge. In terms of RMSE, both models result in large error values when only 10 % of validation data was considered for training, reflecting low reliability of predictions. However, as the proportion increases to 20 % and 30 %, RMSE values decrease, and the distribution narrows, indicating improved prediction accuracy. This trend continues with the models achieving their lowest RMSE values for the proportion between 40 % and 80 %, where CatBoost consistently outperforms XGBoost.

Similarly, Fig. 15 displays the performance of CatBoost and XGBoost models for data Set II with R^2 and RMSE metrics shown across different proportions of combined training data. The R^2 score distributions reveal that both models exhibit moderate R^2 scores with substantial variability at the 20 % and 30 % proportions of validation data being used for training. The CatBoost model shows a higher mean R^2 value compared to the XGBoost, indicating better performance. For a proportion varying between 40 % and 50 %, the models continue to show enhanced R^2 scores with further reduced variability, while the CatBoost consistently maintains a higher median R^2 score. As the proportions increases from 60 % to 80 %, both the models reach their peak performance with high R^2 scores and tight distributions, suggesting robust and reliable predictions. The CatBoost slightly outperforms XGBoost across these proportions, demonstrating greater reliability and stability. On the other hand, the RMSE diagram shows that both CatBoost and XGBoost models initially have relatively high RMSE values at the proportion of 20 % and 30 %. However, the CatBoost has a lower median RMSE than the XGBoost, suggesting higher accuracy. As the proportion increases to 40 % and 50 %, the RMSE values decrease for both models, and the distribution narrows, indicating improved prediction accuracy. This trend continues for the proportion from 60 % to 80 % with decreasing RMSE and tighter distributions.

In short, increasing the proportion of new data for training enhanced the performance of both ML models. While the current study indicated that an acceptable prediction (mean $R^2 > 0.7$) could be assured when > 60 % new dataset is adopted for the adaptive training, the optimal proportion required to achieve an expected level of accuracy varies depending on the specific characteristics of the dataset. Interestingly, although the CatBoost model showed slightly lower performance compared to the XGBoost model when using the original data for training, it consistently delivered better results when updated and applied to new datasets. This finding underscores the adaptability of CatBoost algorithm in handling new and varying data conditions, making it a more robust choice for practical applications where data characteristics may change over time. This adaptability is particularly valuable in the field of geotechnical engineering, where site-specific conditions can significantly influence the accuracy of predictive models. The enhanced performance of CatBoost suggests that it may be better suited for real-world applications where continuous updates and varying data conditions are common. Therefore, integrating strategies that allow for the periodic updating of training data can lead to substantial improvements in model reliability and effectiveness, ultimately contributing to more informed decision-making and optimized resource allocation in civil engineering practices.

Conclusions

This study explored the use of multiple ensemble learning algorithms to develop prediction models for base resistance of pile foundation with special attention to the role of feature selection. Six different ensemble ML algorithms, including the most fundamental decision trees (DT), bagging method Random Forest (RF) and 4 boosting algorithms were adopted for training and assessment. Through an extensive literature review and site investigations (86 different real cases), the study identified 2 major groups of input features for model development, i.e., with and without having information of soil-pile interaction (e.g., the displacement of loading point d_p under loading). The developed models were applied to independent cases of long pile foundation, followed by a reinforced learning process that was proposed to enhance prediction performance. Salient findings and conclusions from the current study can be highlighted as follows:

- The 4 boosting algorithms CatBoost, XGBoost, GBR and LGBM were found to provide more reliable and accurate predictions of base resistance of piles compared to RF and DT. Among them, GBR, CatBoost and XGBoost achieved the largest mean values of R^2 with smaller range of errors, thus recommended choices for practical application. The use of Bayesian-Gaussian coupling optimization significantly promoted model performance, especially for the XGBoost algorithm though the training and test subsets were randomly changed over 300 times.
- Examining model performance over different random divisions of training and testing was found crucial to confirm the reliability and stability of prediction outcomes. The current study showed that the assessment matrices such as R^2 and error indices varied widely over different training and testing subsets. This led to a conclusion that probability analysis considering the non-uniformity and random distribution of data on the predicted outcomes plays a critical role in the practice of decision making.
- Models developed based on the soil-pile interaction detail, i.e., the displacement feature of pile (Set I) resulted in more accurate and reliable predictions with the mean R^2 that could exceed 0.9 with lower standard deviation for XGBoost and GRB algorithms. In both cases, i.e., Set I (representative of data Group (*i*)) and Set II (representative of data Group (*ii*)) of input features, the GRB proved it as the best performing model. This indicated the importance of incorporating pile's displacement into the input features to promote the reliability of ML models,
- Using data Set II, i.e., Group (*ii*) did not require information of soil-pile interaction, reducing the demand for pile load tests, thus considerable saving in practical design and construction of pile foundation. One might conclude that the approach based on data Set II would bring more benefits in real-life applications, especially when the difference in prediction performance between models based on data Set I and Set II was not always considerable.
- The feature analysis showed that the displacement of piles provided the most comprehensive and reliable indication to the response of pile's base resistance. While the applied load played a decisive role in the development of base resistance with its SHAP mean value that was 4 times larger than other features in data Set II, it dopped to the second rank when the displacement was considered. Moreover, the explainable ML techniques scored the contribution from shear strength of side and base soils to tip resistance, reasonably representing the load-transfer process of pile foundation under loading.
- The proposed reinforced learning process was found effective to improve prediction of new unfamiliar cases when > 60 % new dataset were combined to the existing training database. The CatBoost model showed its better adaptability to new data compared to XGBoost, even though it had poorer performance when predicting original data. This investigation addressed an inherent matter of data-based prediction method when there is considerable difference between the training and new datasets.

This study has achieved considerable success in developing ML models for predicting base resistance of pile foundation, especially giving valuable implications to the selection of impact features and building database, which is often the most important step to construct a ML model. Nevertheless, there are some limitations that require further effort in the future study. First, the current data of base resistance Q_b mainly concentrate in the range < 4000 kPa, resulting in a lack of confidence when predicting pile foundation where Q_b can exceed 4000 kPa. Second, the poor performance of trained models when predicting independent cases of pile foundations from past studies needs more in-depth analysis and insightful understanding. How the difference be-

tween datasets can be characterized in a more quantitative approach and being used as indicators for assessing prediction quality are an intrinsic matter needs to be solved to achieve a more complete and robust solution for developing ML models of pile foundation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Thanh T. Nguyen reports administrative support was provided by University of Technology Sydney. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Thanh T. Nguyen: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Khuong Le Nguyen:** Software, Methodology, Investigation, Formal analysis. **Thien Q. Huynh:** Validation, Resources, Data curation. **Quangdung Tran:** Resources, Data curation.

Data availability

The data used in the current study can be found in Supplementary Materials of this paper.

Acknowledgements

The field investigation and data collection were supported by a great number of local civil/geotechnical contractors and consultants in Vietnam such as Hoa Binh Construction Group, FECON South JSC., (FECON Corporation), Bachy Soletanche VN, among others.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.geoai.2025.100019.

References

AASHTO Specifications. (2010). *LRFD bridge design specification*. Washington: American Association of State Highway Officials.

Al-Atroush, M. E., Hefny, A. M., & Sorour, T. M. (2022). Modified Meyerhof approach for forecasting reliable ultimate capacity of the large diameter bored piles. *Scientific Reports, 12*(1), 8541. 10.1038/s41598-022-12238-w.

Alkroosh, I. S., Bahadori, M., Nikraz, H., & Bahadori, A. (2015). Regressive approach for predicting bearing capacity of bored piles from cone penetration test data. *Journal of Rock Mechanics and Geotechnical Engineering, 7*(5), 584–592. 10.1016/j.jrmge.2015.06.011.

Amjad, M., Ahmad, I., Ahmad, M., Wróblewski, P., Kamiński, P., & Amjad, U. (2022). Prediction of pile bearing capacity using XGBoost algorithm: Modeling and performance evaluation. *Applied Sciences, 12*(4), 2126.

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*, Article 103991. 10.1016/j.earscirev.2022.103991.

Benbouras, M. A., Petrişor, A.-I., Zedira, H., Ghelani, L., & Lefilef, L. (2021). Forecasting the bearing capacity of the driven piles using advanced machine-learning techniques. *Applied Sciences, 11*(22), Article 10908.

Bohn, C., Santos, A. L. d., & Frank, R. (2017). Development of axial pile load transfer curves based on instrumented load tests. *Journal of Geotechnical and Geoenvironmental Engineering, 143*(1), Article 04016081. 10.1061/(ASCE)GT.1943-5606.0001579.

Budi, G. S., Kosasi, M., & Wijaya, D. H. (2015). Bearing capacity of pile foundations embedded in clays and sands layer predicted using PDA test and static load test. *Procedia Engineering, 125*, 406–410. 10.1016/j.proeng.2015.11.101.

Cao, M.-T., Nguyen, N.-M., & Wang, W.-C. (2022). Using an evolutionary heterogeneous ensemble of artificial neural network and multivariate adaptive regression splines to predict bearing capacity in axial piles. *Engineering Structures, 268*, Article 114769. 10.1016/j.engstruct.2022.114769.

Chan, W. T., Chow, Y. K., & Liu, L. F. (1995). Neural network: An alternative to pile driving formulas. *Computers and Geotechnics, 17*(2), 135–156. 10.1016/0266-352X(95)93866-H.

Dai, G., Salgado, R., Gong, W., & Zhang, Y. (2012). Load tests on full-scale bored pile groups. *Canadian Geotechnical Journal, 49*(11), 1293–1308. 10.1139/t2012-087.

Das, S. K., & Suman, S. (2015). Prediction of lateral load capacity of pile in clay using multivariate adaptive regression spline and functional network. *Arabian Journal for Science and Engineering, 40*(6), 1565–1578. 10.1007/s13369-015-1624-y.

DCCEEW. 2021. Australia State of the Environmenment Report. *In* Department of Climate Change, Energy, the Environment and Water.

ECP202/4. 2005. Egyptian code for soil mechanics–Design and construction of foundations. Part 4, Deep Foundations.

El Gendy, M., El-Arabi, I. A., & Kamal, M. A. (2014). Comparative analyses of large diameter bored piles using international codes. *DFI Journal - The Journal of the Deep Foundations Institute, 8*(1), 15–26. 10.1179/1937525514Y.0000000001.

Fellenius, B. H., & Nguyen, M. H. (2013). Large diameter long bored piles in the Mekong Delta. *International Journal of Geoengineering Case Histories*, (3), 196–207.

Goh, A. T. C., Kulhawy, F. H., & Chua, C. G. (2005). Bayesian Neural network analysis of undrained side resistance of drilled shafts. *Journal of Geotechnical and Geoenvironmental Engineering, 131*(1), 84–93. 10.1061/(ASCE)1090-0241(2005)131:1(84).

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data, 7*(1), 94. 10.1186/s40537-020-00369-8.

Hazewinkel, A. (2022). *Driveability predictions in vibratory pile driving*. TU Delft Mechanical. Masters thesis. https://repository.tudelft.nl/record/uuid:281c7676-346c-4e88-96cf-7e36c4719b15.

Hirayama, H. (1990). Load-settlement analysis for bored piles using hyperbolic transfer functions. *Soils and Foundations, 30*(1), 55–64. 10.3208/sandf1972.30.55.

Huynh, T. Q., Nguyen, T. T., & Nguyen, H. (2023). Base resistance of super-large and long piles in soft soil: Performance of artificial neural network model and field implications. *Acta Geotechnica, 18*(5), 2755–2775. 10.1007/s11440-022-01736-w.

Kardani, N., Zhou, A., Nazem, M., & Shen, S.-L. (2020). Estimation of bearing capacity of piles in cohesionless soil using optimised machine learning approaches. *Geotechnical and Geological Engineering, 38*(2), 2271–2291. 10.1007/s10706-019-01085-8.

Li, L., Lai, N., Zhao, X., Zhu, T., & Su, Z. (2023). A generalized elastoplastic load-transfer model for axially loaded piles in clay: Incorporation of modulus degradation and skin friction softening. *Computers and Geotechnics, 161*, Article 105594. 10.1016/j.compgeo.2023.105594.

Millán, M. A., Picardo, A., & Galindo, R. (2023). Application of artificial neural networks for predicting the bearing capacity of the tip of a pile embedded in a rock mass. *Engineering Applications of Artificial Intelligence, 124*, Article 106568. 10.1016/j.engappai.2023.106568.

Moayedi, H., & Hayati, S. (2018). Applicability of a CPT-based neural network solution in predicting load-settlement responses of bored pile. *International Journal of Geomechanics, 18*(6), Article 06018009. 10.1061/(ASCE)GM.1943-5622.0001125.

Momeni, E., Nazir, R., Armaghani, D. J., & Maizir, H. (2015). Application of artificial neural network for predicting shaft and tip resistances of concrete piles. *Earth Sciences Research Journal, 19*(1). 10.15446/esrj.v19n1.38712.

Momeni, E., Nazir, R., Jahed Armaghani, D., & Maizir, H. (2014). Prediction of pile bearing capacity using a hybrid genetic algorithm-based ANN. *Measurement, 57*, 122–131. 10.1016/j.measurement.2014.08.007.

Muduli, P. K., Das, S. K., & Das, M. R. (2013). Prediction of lateral load capacity of piles using extreme learning machine. *International Journal of Geotechnical Engineering, 7*(4), 388–394. 10.1179/1938636213Z.00000000041.

Nguyen, M. H., & Fellenius, B. H. (2014). O-cell tests on two 70 m long bored piles in Vietnam. In *From soil behavior fundamentals to innovations in geotechnical engineering: Honoring roy E. olson* (pp. 482–496). *Edited*. pp..

Nguyen, M. H., & Fellenius, B. H. (2015). Bidirectional cell tests on non-grouted and grouted large-diameter bored piles. *Journal of Geo-Engineering Sciences, 2*, 105–117. 10.3233/JGS-140025.

Nguyen, T., Ly, D.-K., Huynh, T. Q., & Nguyen, T. T. (2023). Soft computing for determining base resistance of super-long piles in soft soil: A coupled SPBO-XGBoost approach. *Computers and Geotechnics, 162*, Article 105707. 10.1016/j.compgeo.2023.105707.

Nguyen, T., Ly, K.-D., Nguyen-Thoi, T., Nguyen, B.-P., & Doan, N.-P. (2022). Prediction of axial load bearing capacity of PHC nodular pile using Bayesian regularization artificial neural network. *Soils and Foundations, 62*(5), Article 101203. 10.1016/j.sandf.2022.101203.

Nguyen, T. T., Le, V. D., Huynh, T. Q., & Nguyen, N. H. T. (2024). Influence of settlement on base resistance of long piles in soft soil—Field and machine learning assessments. *Geotechnics, 4*(2), 447–469. 10.3390/geotechnics4020025.

Ouyang, W., Li, G., Chen, L., & Liu, S.-W. (2024). Machine learning-based soil–structure interaction analysis of laterally loaded piles through physics-informed neural networks. *Acta Geotechnica*. 10.1007/s11440-023-02179-7.

Pham, T. A., Tran, V. Q., Vu, H.-L. T., & Ly, H.-B. (2020). Design deep neural network architecture using a genetic algorithm for estimation of pile bearing capacity. *PLOS ONE, 15*(12), Article E0243030. 10.1371/journal.pone.0243030.

Pooya Nejad, F., & Jaksa, M. B. (2017). Load-settlement behavior modeling of single piles using artificial neural networks and CPT data. *Computers and Geotechnics, 89*, 9–21. 10.1016/j.compgeo.2017.04.003.

Pooya Nejad, F., Jaksa, M. B., Kakhi, M., & McCabe, B. A. (2009). Prediction of pile settlement using artificial neural networks based on standard penetration test data. *Computers and Geotechnics, 36*(7), 1125–1133. 10.1016/j.compgeo.2009.04.003.

Poulos, H. G. (1989). Pile behaviour—Theory and application. *Géotechnique, 39*(3), 365–415. 10.1680/geot.1989.39.3.365.

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

Reimann, L., Vafeidis, A. T., & Honsel, L. E. (2023). Population development as a driver of coastal risk: Current trends and future pathways. *Cambridge Prisms: Coastal Futures, 1*, E14. 10.1017/cft.2023.3.

Samui, P., & Kim, D. (2013). Least square support vector machine and multivariate adaptive regression spline for modeling lateral load capacity of piles. *Neural Computing and Applications, 23*(3), 1123–1127. 10.1007/s00521-012-1043-x.

Shahin, M. A. (2010). Intelligent computing for modeling axial capacity of pile foundations. *Canadian Geotechnical Journal, 47*(2), 230–243. 10.1139/t09-094.

Shahin, M. A. (2014). Load-settlement modeling of axially loaded drilled shafts using CPT-based recurrent neural networks. *International Journal of Geomechanics, 14*(6), Article 06014012. 10.1061/(ASCE)GM.1943-5622.0000370.

Sharo, A., Al-Shorman, B., Bani Baker, M., Nusier, O., & Alawneh, A. (2022). New approach for predicting the load-displacement curve of axially loaded piles in sand. *Case Studies in Construction Materials, 17*, Article E01674. 10.1016/j.cscm.2022.e01674.

TCVN 10304. (2014). *Pile Foundation- design standard*. Vietnam National Standard, Ministry of Science and Technology.

Truong, M. H., Indraratna, B., Nguyen, T. T., Carter, J., & Rujikiatkamjorn, C. (2021). Analysis of undrained cyclic response of saturated soils. *Computers and Geotechnics, 134*, Article 104095. 10.1016/j.compgeo.2021.104095.

Vahab, M., Shahbodagh, B., Haghighat, E., & Khalili, N. (2023). Application of physics-informed neural networks for forward and inverse analysis of pile–soil interaction. *International Journal of Solids and Structures, 277-278*, Article 112319. 10.1016/j.ijsolstr.2023.112319.

VIKTOR. 2022. Predict pile drive times with machine learning [Online]. Available: Https://www.viktor.ai/customer-cases/38/machine-learning-predict-pile-drive-time [Accessed].

Wrana, B. (2015). Pile load capacity – Calculation methods. *Studia Geotechnica et Mechanica, 37*(4), 83–93. 10.1515/sgem-2015-0048.

Wu, J., El Naggar, M. H., & Wang, K. (2023). Pile damage detection using machine learning with the multipoint traveling wave decomposition method. *Sensors, 23*(19), 8308.

Xiao, K., Guo, S., Wen, J., Han, J., & Yang, X. (2023). Three-stage analysis method for calculating the settlement of large-diameter extralong piles. *International Journal of Geomechanics, 23*(3), Article 04023001. 10.1061/IJGNAI.GMENG-8110.

Yousheng, D., Keqin, Z., Zhongju, F., Wen, Z., Xinjun, Z. O. U., & Huiling, Z. (2024). Machine learning based prediction model for the pile bearing capacity of saline soils in cold regions. *Structures, 59*, Article 105735. 10.1016/j.istruc.2023.105735.