



## OPEN Comparative performance evaluation of machine learning models for predicting the ultimate bearing capacity of shallow foundations on granular soils

Jalal Shah<sup>1</sup>, Mehtab Alam<sup>3</sup>✉, Faisal Javed<sup>1,2</sup>, Muhammad Umar<sup>4,5</sup> & Furqan Ahmad<sup>6</sup>✉

Accurate estimation of the ultimate bearing capacity (UBC) of shallow foundations is critical for safe and economical geotechnical design. Traditional approaches depend heavily on extensive and costly field and laboratory investigations, while numerical simulations, though effective, are computationally intensive and time-consuming. To address these limitations, this study investigates the application of machine learning (ML) models for efficient and reliable prediction of the ultimate bearing capacity of shallow foundations. Although numerous studies have explored individual ML techniques for this purpose, a comprehensive and consistent comparison of widely used models under identical conditions remains limited. This research evaluates six ML algorithms; k-Nearest Neighbors (kNN), Artificial Neural Network (NN), Random Forest (RF), Extreme Gradient Boosting (xGBoost), Adaptive Boosting (AdaBoost), and Stochastic Gradient Descent (SGD), using a dataset of 169 experimental results collected from literature. The input features include foundation width (B), depth (D), length-to-width ratio (L/B), soil unit weight ( $\gamma$ ), and angle of internal friction ( $\phi$ ). Model performance was assessed using multiple evaluation metrics: coefficient of determination ( $R^2$ ), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and objective function (OBJ). To enhance model interpretability, SHapley Additive Explanations (SHAP) and Partial Dependence Plots (PDPs) were employed to analyze feature importance and input-output relationships, highlighting the influence of both soil properties and foundation geometry on predicted bearing capacity. Among the evaluated models, AdaBoost demonstrated the best overall performance, achieving  $R^2$  values of 0.939 and 0.881 on the training and testing sets, respectively. Based on the cumulative ranking of the models across all evaluation metrics, the models were ranked in the following order of performance: AdaBoost > kNN > RF > xGBoost > NN > SGD. While the results are promising, a key limitation is the use of single-layer soil data, which restricts applicability to more complex, multilayered soil profiles. Future studies should incorporate multilayer datasets and account for spatial variability to enhance the generalizability and robustness of predictive models.

**Keywords** Ultimate bearing capacity, Machine learning, Shallow foundation

The ultimate bearing capacity (UBC) is a fundamental parameter in foundation engineering, forming the basis for designing and evaluating foundation performance. Foundation failure primarily occurs due to excessive settlement or bearing capacity failure, with the latter often governing design. The UBC is defined as the maximum stress the ground can sustain before experiencing shear failure or unacceptable settlement. Kohestani et al.<sup>1</sup> defined shallow foundations as those with a depth-to-width ratio (D/B) of  $\leq 4$ . For such foundations, the accurate determination of the ultimate bearing capacity (UBC) is crucial for ensuring structural stability.

<sup>1</sup>Department of Civil Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23460, Khyber Pakhtunkhwa, Pakistan. <sup>2</sup>Western Caspian University, Baku, Azerbaijan. <sup>3</sup>Department of Geotechnical Engineering, College of Civil Engineering, Tongji University, Shanghai 20009, China. <sup>4</sup>Department of Civil Engineering, National University of Computer Emerging Sciences Lahore Campus, 852-B Milaad St, Block B Faisal Town, Lahore 54770, Pakistan. <sup>5</sup>Institute of Industrial science, The University of Tokyo, Tokyo, Japan. <sup>6</sup>UNHCR, Jalalabad, Afghanistan. ✉email: engr\_mehtab@yahoo.com; engr.furqanahmad123@gmail.com

Traditional approaches for estimating UBC rely on theoretical formulations incorporating numerous simplifying assumptions. Terzaghi<sup>2</sup> pioneered a comprehensive bearing capacity theory in 1943 using the limit equilibrium method, followed by advancements from Meyerhof<sup>3</sup>, Hansen<sup>4</sup> and Vesic<sup>5</sup>. While multiple UBC theories exist, they share a common fundamental structure and can be expressed as:

$$Q_u = cN_c S_c d_c i_c + qN_q S_q d_q i_q + 0.5\gamma B N_\gamma S_\gamma d_\gamma i_\gamma \quad (1)$$

Where  $B$  denotes the width of foundation,  $q$  denotes the effective stress at the foundation base,  $c$  is the soil cohesion,  $\gamma$  is the soil's unit weight,  $N_c$ ,  $N_q$ ,  $N_\gamma$  are the bearing capacity factors,  $S_c$ ,  $S_q$ ,  $S_\gamma$ , are the foundation shape factors,  $d_c$ ,  $d_q$ ,  $d_\gamma$  are depth factors,  $i_c$ ,  $i_q$ ,  $i_\gamma$  are the load inclination factors.

Despite the extensive development of these classical equations, they present inherent limitations. The various factors such as bearing capacity, depth, shape, and inclination factors are significantly influenced by foundation geometry, slope inclination and in the case of granular soil, the internal friction angle ( $\phi$ ). However, conventional methods often oversimplify soil behavior by assuming homogeneity, isotropy, and linear stress-strain responses, which can lead to conservative designs, increased construction costs, and potential inaccuracies when compared to experimental data<sup>6</sup>. These limitations highlight the need for alternative predictive approaches that are efficient, fast, accurate, and less costly, while also accounting for soil variability and site-specific complexities.

Numerical techniques such as the finite element method (FEM) and finite difference method (FDM) offer more sophisticated simulations<sup>7–12</sup> but require extensive computational resources and complex material modeling. In contrast, machine learning (ML) has recently emerged as a powerful tool for geotechnical engineering applications<sup>13–15</sup> enabling data-driven predictions without the need for explicit soil behavior assumptions. ML models can process large datasets, identify complex patterns, and provide more flexible predictive frameworks compared to conventional methods<sup>13</sup>.

Several studies have demonstrated the effectiveness of ML in predicting the UBC of shallow foundations. Padmini et al. compared artificial neural networks (ANNs), fuzzy logic, and neuro-fuzzy models against classical UBC equations, concluding that data-driven models achieved superior accuracy. Zhao and Yin<sup>16</sup> applied chaotic particle swarm optimization with support vector machines (CPSO-SVM) using input parameters identical to this study and reported improved performance over conventional models. Similarly, Adarsh. S<sup>17</sup>, et al. explored support vector machines (SVMs) and genetic programming (GP), highlighting the potential of soft computing techniques. Gupta. R., et al.<sup>18</sup> used ANNs to identify the minimum necessary input variables for accurate UBC prediction. More recently, Alzabeebee. S<sup>19</sup>, et al. developed a multi-objective genetic algorithm-based model that outperformed classical theories from Terzaghi, Vesic, and Hansen, as well as several existing ML approaches.

Kohestani, V., et al.<sup>1</sup> utilized 112 data points from the dataset used in this study to develop a Random Forest (RF) model for predicting the UBC of shallow foundations on cohesionless soils. The model achieved an impressive correlation coefficient (R) value of 0.9871. However, while this indicates high accuracy, the small size of the dataset raises concerns about the model's generalizability and practical applicability. Machine learning models developed with small datasets tend to overfit, leading to reduced performance when deployed in real-world applications with new, unseen data. Furthermore, this study focused on achieving accurate predictions without addressing the interpretability or explainability of the model, which are critical for understanding how the model derives its predictions, quantifying the importance of input parameters, and showing the relationships between the inputs and the output (i.e., the UBC of shallow foundations).

Khajehzadeh M. et al.<sup>20</sup> conducted a review of artificial intelligence applications for the UBC evaluation of shallow foundations, focusing on studies published between 2015 and 2023. The review concluded that AI has emerged as one of the most acceptable and suitable approaches for estimating the bearing capacity of shallow foundations. However, the authors also observed that most of the studies relied on small datasets, with the largest dataset comprising only 97 real experimental data points. This highlights the need for larger datasets to improve the accuracy and reliability of ML models, thus identifying a significant research gap related to the quantity of data used in training these models.

Liu Y. et al.<sup>21</sup> developed two artificial neural network (ANN)-based models for predicting the UBC of shallow foundations on two-layered soil profiles: the backtracking search algorithm artificial neural network (BSA-ANN) and equilibrium optimizer artificial neural network (EO-ANN). Their study utilized 901 data points derived from 2D finite element analysis. They also created a monolithic predictive formula based on the EO-ANN model, showing that the training root mean square error (RMSE) of the ANN was reduced by 11.72% and 17.46%, respectively.

Although numerous studies have explored machine learning (ML) models for predicting the ultimate bearing capacity (UBC) of shallow foundations, most have relied on heterogeneous datasets and varied evaluation metrics. These inconsistencies hinder direct, fair comparisons across models and limit the generalizability of their findings. Additionally, many existing comparisons involve models trained on different datasets, often with varying data quality, size, feature distributions, and experimental setups, further undermining the reliability of conclusions drawn.

To address these limitations, the present study develops and evaluates six widely adopted ML models; Neural Network (NN), Random Forest (RF), k-Nearest Neighbors (kNN), Stochastic Gradient Descent (SGD), Extreme Gradient Boosting (xGBoost), and Adaptive Boosting (AdaBoost), within a unified and consistent framework. Using a comprehensive dataset comprising 169 experimental results compiled by Khorrami. R., et al.<sup>22</sup> that include input features: internal friction angle ( $\phi$ ), unit weight ( $\gamma$ ), foundation width ( $B$ ), depth ( $D$ ), and length-to-width ratio ( $L/B$ ), which have been widely recognized in literature as critical variables for UBC estimation<sup>23</sup>. This uniformity ensures objective, head-to-head performance comparisons under controlled conditions.

Model performance is assessed using widely accepted statistical metrics, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error

(RMSE), Coefficient of Determination ( $R^2$ ), and Objective Function (OBJ). Beyond predictive accuracy, the study emphasizes model explainability and interpretability, areas often underexplored in geotechnical ML research. A comprehensive analysis using SHapley Additive Explanations (SHAP) and Partial Dependence Plots (PDPs) is conducted to uncover the influence of each input parameter on model predictions.

Ultimately, this study establishes a robust benchmark for evaluating ML models in UBC prediction and offers valuable insights into selecting appropriate algorithms in geotechnical applications. By enhancing both predictive reliability and transparency, the proposed framework contributes to more informed and cost-effective foundation design practices.

## Methodology

### Data collection and description

In numerous studies, the width (B), length (L), and depth of the foundation (D), along with key soil parameters such as unit weight ( $\gamma$ ) and the angle of internal friction ( $\phi$ ), have been identified as the primary determinants of the ultimate bearing capacity (UBC) of shallow foundations on cohesionless soil<sup>24–29</sup>. These factors are crucial as they influence stress distribution and failure mechanisms, which are essential for accurately estimating UBC. Foye et al. emphasized that B, D, L,  $\gamma$ , and  $\phi$  are the principal governing parameters in UBC estimation. Among the geometric factors, foundation depth (D) has the most significant impact, whereas among all the parameters considered, the angle of internal friction ( $\phi$ ) exerts the greatest influence on the UBC of granular soils<sup>3,30</sup>.

The dataset used in this study was compiled by Khorrami. R., et al.<sup>22</sup> from existing literature and consists of 169 data points. The selected input parameters include foundation width (B), depth (D), and length-to-width ratio (L/B), as well as soil properties; namely unit weight ( $\gamma$ ) and angle of internal friction ( $\phi$ ). Of the 169 data points, 65 are derived from small-scale experiments, while the remaining data points originate from large-scale experiments. The statistical analysis of the dataset is presented in Table 1.

Statistical analysis provides insights into the distribution, central tendency, and variability of the dataset. Kurtosis, a statistical measure, describes the prominence of a distribution's tails relative to the overall distribution, indicating whether the data exhibit heavier or lighter tails compared to a normal distribution. Skewness, on the other hand, quantifies the asymmetry of a probability distribution. The acceptable range for skewness is  $-3$  to  $+3$ , while for kurtosis, it is  $-10$  to  $+10$ <sup>31,32</sup>. The values of both measures for all features in this study fall within these recommended bounds, confirming the dataset's suitability for analysis<sup>33</sup>.

### Correlation matrix for input parameters

A Pearson correlation matrix was constructed to evaluate the relationships and dependencies between input and output parameters. This analysis provides insights into the relative influence of different parameters on UBC and aids in assessing ML model performance. As illustrated in Fig. 1, the depth of foundation (D) emerges as the most influential factor, followed by the width of foundation (B), while unit weight ( $\gamma$ ) has the least effect on UBC.

Multicollinearity and interdependency among input features can present challenges in ML modeling. When two or more features exhibit a high degree of correlation, it becomes difficult for ML models to distinguish the individual contributions of each variable, potentially leading to overfitting and reduced predictive accuracy. Strong correlations among input features may cause the model to overemphasize redundant information, impairing its ability to generalize effectively. Mitigating multicollinearity is crucial for enhancing the interpretability and performance of models, thus ensuring that the effect of each feature is captured correctly<sup>34–36</sup>.

It is generally recommended that the correlation coefficient of any two variables must be less than 0.80 to prevent multicollinearity problems<sup>37,38</sup>. In this study, the correlation values for all feature pairs remain below this threshold, indicating that multicollinearity is unlikely to affect model performance significantly.

### Background of utilized ML methods

This section provides an overview of the machine learning (ML) models used for predicting the UBC of shallow foundations.

#### Random forest (RF)

Random Forest (RF) is an ensemble learning algorithm for predictive modeling. RF builds numerous decision trees (DTs) at training time, with each of them being built based on a random subset of data and input features. Randomness is utilized to lower overfitting and improve generalization<sup>36</sup>. In contrast to a single decision tree, which may be subject to high variance, RF improves predictive accuracy and stability by averaging the

	Inputs					Output
	B (m)	D (m)	L/B	$\gamma$ (kN/m <sup>3</sup> )	$\Phi$ (°)	$q_u$ (kN/m <sup>2</sup> )
Minimum	0.030	0.000	1.000	9.850	31.950	14.000
Maximum	3.016	0.890	6.000	20.800	45.700	2847.000
Standard Deviation	0.563	0.208	1.914	2.923	3.429	470.978
Mean	0.532	0.119	2.217	15.637	39.208	481.527
Skewness	2.507	2.160	1.223	-0.315	-0.338	2.066
Kurtosis	8.186	4.021	-0.202	-0.534	-0.422	5.215

**Table 1.** Shows statistical analysis of the data.

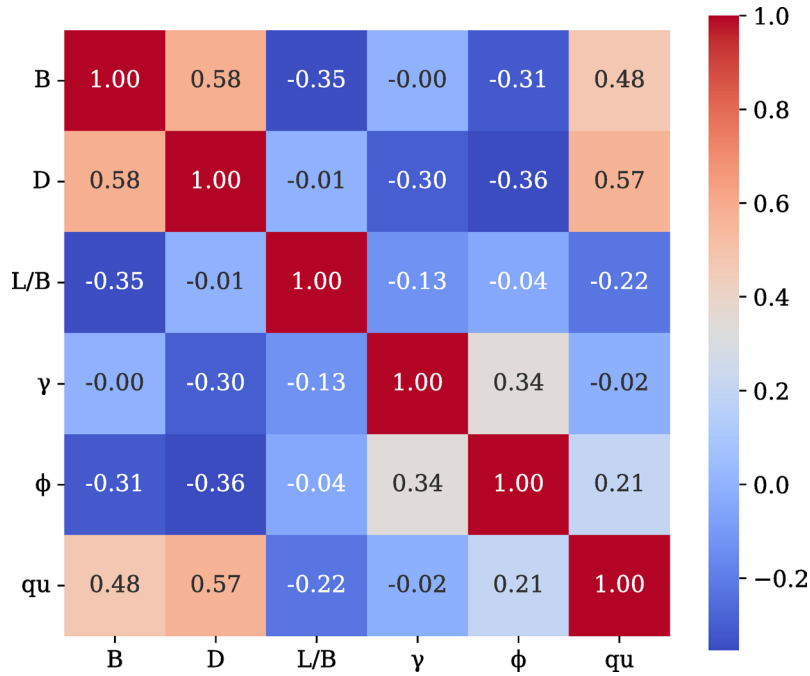


Fig. 1. Correlation matrix for the input dataset.

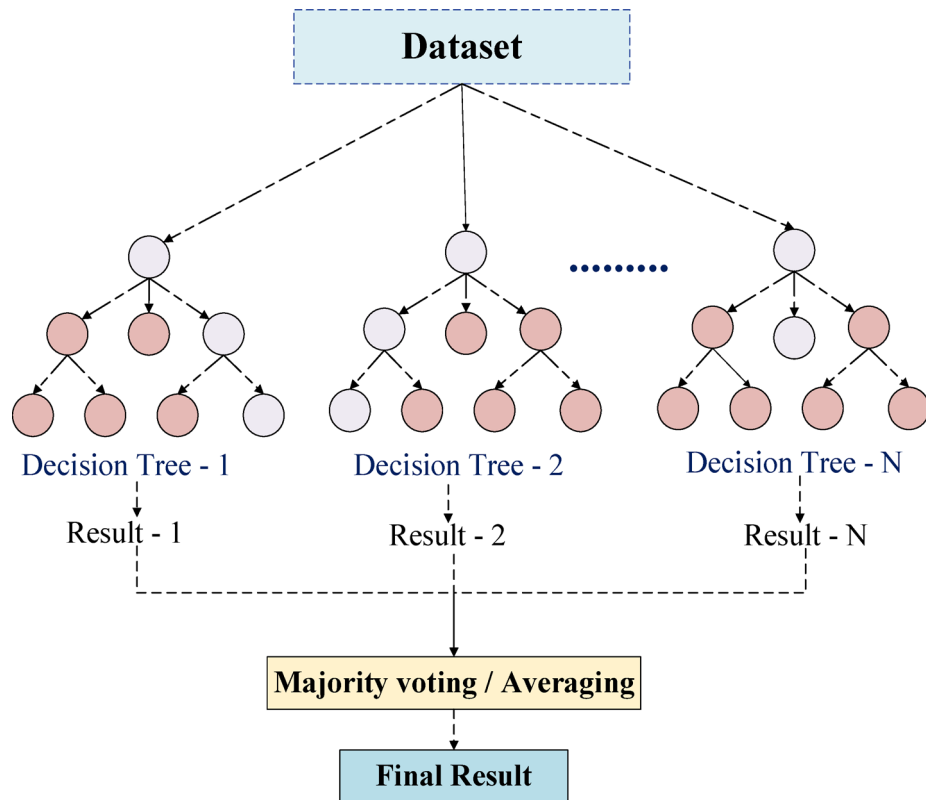


Fig. 2. Workflow of Random Forest.

output of multiple trees. RF is especially suited for dealing with big data and high-dimensional features, with an improvement in balance between bias and variance. However, its computational cost is higher than that of individual decision trees, and it can be less interpretable<sup>34,35</sup>. The schematic workflow of RF model is given in Fig. 2.

*Extreme gradient boosting (xGBoost)*

Extreme Gradient Boosting (xGBoost) is a boosting-based ensemble learning algorithm commonly used for classification and regression problems. It sequentially constructs models, where every additional tree attempts to minimize residual errors of previous trees<sup>38</sup>. xGBoost repeatedly aggregates weak learners to produce a strong predictive model that can learn intricate data patterns (Fig. 3). Although it performs better, it overfits unless regularized and needs fine-tuning of its hyperparameters, e.g., learning rate and tree depth. Contrary to RF, where individual decision trees are trained separately and then combined through averaging, xGBoost constructs trees sequentially that attempt to rectify previous errors<sup>37</sup>. Although computationally intensive, xGBoost is commonly applied owing to its strong performance across a vast range of ML applications.

*k-Nearest neighbors (kNN)*

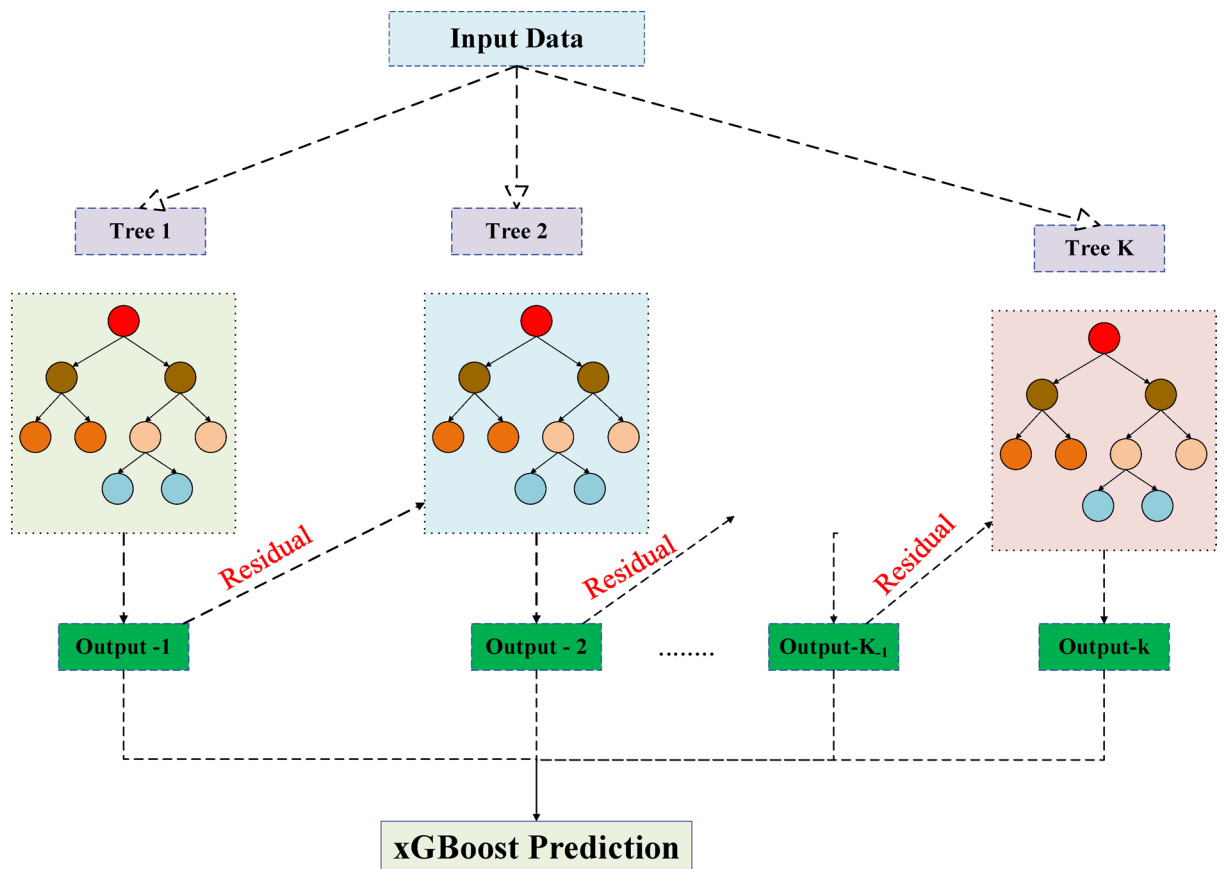
k-Nearest Neighbors (kNN) is a simple yet effective ML algorithm employed for regression as well as classification problems. It is an instance-based learning that makes predictions based on the majority class (in the case of classification) or mean (in the case of regression) of the k-nearest points in the feature space<sup>39</sup>. Unlike the majority of ML models, kNN does not need any explicit training process, which makes it computationally efficient in model building. Nevertheless, its performance is extremely sensitive to the selection of the distance measure as well as the value of k. kNN is also computationally intensive in the case of large datasets since it entails pairwise distance computations for each new prediction<sup>40</sup>. The schematic workflow of KNN model is given in Fig. 4.

*Neural networks (NN)*

Neural Networks (NN) is a class of deep learning models inspired by the structure and functional processes of the human brain. NN models consist of layers of artificial neurons that are connected and enable processing and interpretation of input data to generate predictive outcomes. NN models are ideally suited to find complex, nonlinear relationships between data and hence are suitable for a range of applications, from classification to regression and clustering. While their flexibility has long been established, neural networks require vast amounts of data and considerable computational resources in training. Furthermore, their intrinsic “black box” nature raises formidable interpretability challenges<sup>41</sup>. The schematic workflow of NN model is given in Fig. 5.

*Adaptive boosting (AdaBoost)*

Adaptive Boosting (AdaBoost) is a boosting technique designed to enhance the performance of weak learners, typically decision trees in most instances, by blending them together to form a robust ensemble model. AdaBoost



**Fig. 3.** Workflow of xGBoost.

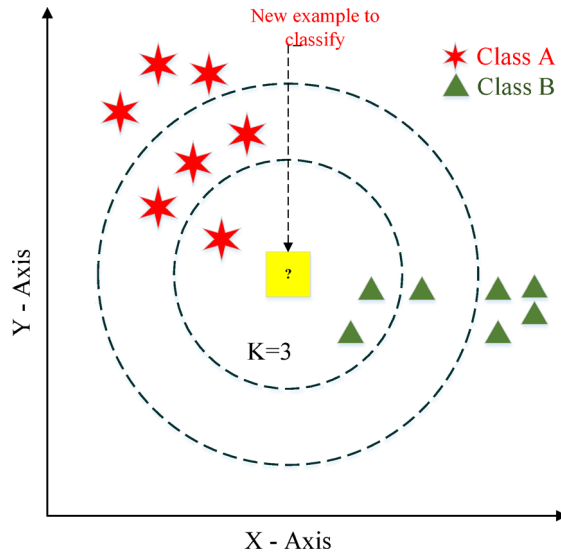


Fig. 4. Workflow of kNN.

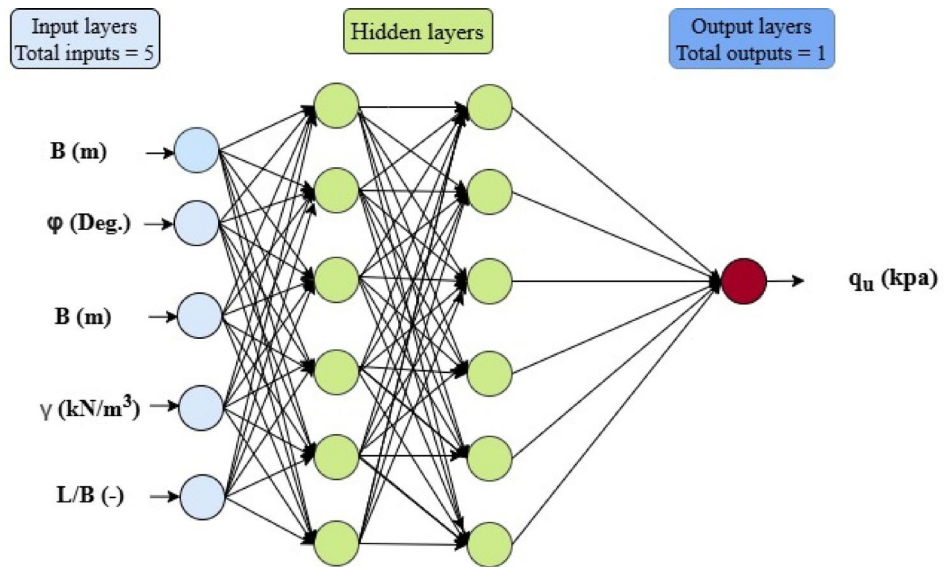


Fig. 5. Workflow of Neural Network.

assigns weights to instances in the data and hence targets misclassified instances in subsequent iterations<sup>42</sup>. The recursive approach effectively reduces bias as well as variance, hence improving overall predictive accuracy. (Fig. 6). While robust in an overwhelming majority of applications, AdaBoost is highly susceptible to noisy data and outliers, which can degrade performance<sup>43</sup>.

*Stochastic gradient descent (SGD)*

Stochastic Gradient Descent (SGD) is a widely used optimization technique for ML model training, particularly in big data. Unlike the regular gradient descent that updates model parameters from the whole dataset, SGD updates parameters step by step from a randomly selected data point, significantly improving computational efficiency<sup>44</sup> schematically as shown in Fig. 7. This makes it well-suited for high-dimensional data and large-scale optimization problems<sup>45</sup>. However, the convergence path of SGD might be highly unstable, demanding careful tuning of hyperparameters such as learning rate and momentum to obtain stable performance.

**Model development**

In machine learning, dataset splitting is a fundamental step for evaluating model performance, with standard split ratios commonly ranging from 70/30 to 80/20<sup>46</sup>. This study assessed each model under various split ratios to identify the optimal configuration, as shown in Table 2. As the training set percentage increases, error metrics

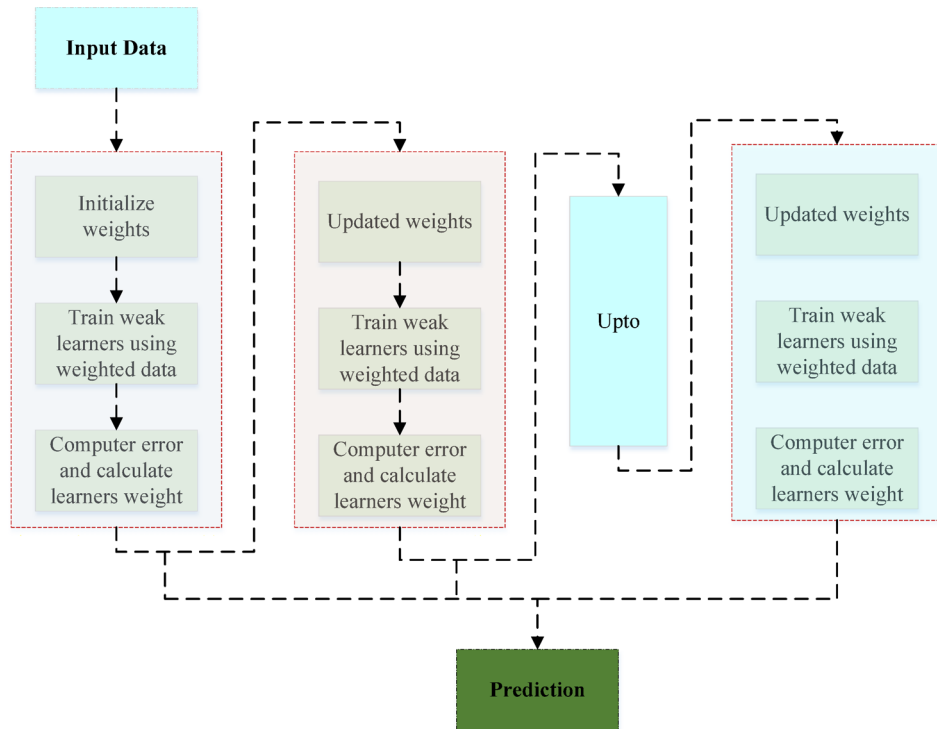


Fig. 6. Workflow of AdaBoost.

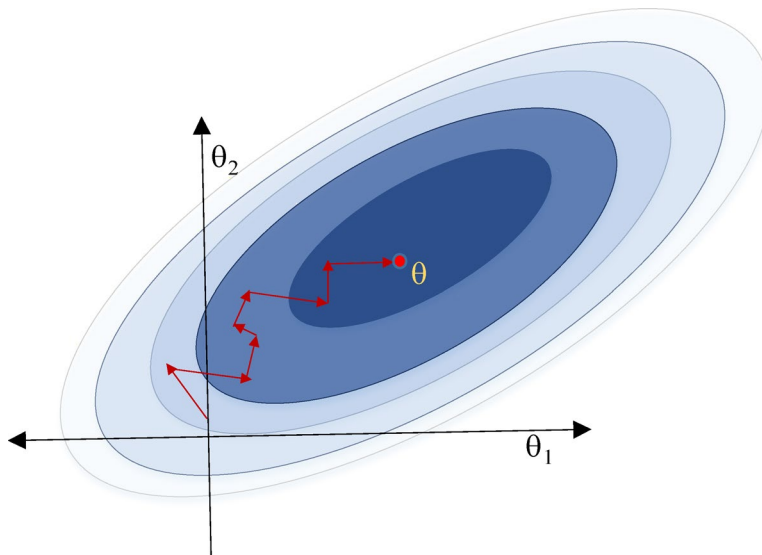


Fig. 7. Workflow of Stochastic Gradient Descent.

(MSE, RMSE, MAE, MAPE) decrease while  $R^2$  improves, indicating more robust model learning. For instance, Random Forest at a 70/30 split yields an MSE of 25497.87 and an  $R^2$  of 0.716, whereas at 80/20, its MSE declines to 11508.981 and  $R^2$  rises to 0.881; a similar trend is observed for AdaBoost, which improves from an MSE of 27455.316 and  $R^2$  of 0.694 to an MSE of 11521.525 and  $R^2$  of 0.881. The same trend is observed across all the models. These findings corroborate earlier studies, suggesting that allocating more than 70% of the data to training provides sufficient examples for capturing complex relationships while maintaining a sizable test set for generalization<sup>47,48</sup>. Consequently, the 80/20 split emerges as a balanced choice, delivering superior predictive accuracy for most models compared to smaller training sets.

Machine learning models were developed using Orange Data Mining, a visual, coding-free software that simplifies model implementation. The dataset was imported into the software via the File widget (as an.xlsx file) and then connected to the Data Sampler widget, which randomly splits the dataset into training and

Split Ratio	Model	Results of Test on Test data				
		MSE	RMSE	MAE	MAPE	R <sup>2</sup>
70/30	kNN	34287.543	185.169	116.057	0.439	0.618
	RF	25497.87	159.681	89.319	0.284	0.716
	xGBoost	31148.671	176.49	96.696	0.258	0.653
	AdaBoost	27455.316	165.696	89.584	0.282	0.694
	SGD	84194.051	290.162	185.52	0.742	0.091
	NN	31619.798	177.82	102.312	0.284	0.647
72/28	kNN	36331.766	190.609	119.592	0.418	0.608
	RF	26366.387	162.377	92.314	0.28	0.715
	xGBoost	32643.497	180.675	99.644	0.26	0.648
	AdaBoost	28321.204	168.289	97.624	0.298	0.694
	SGD	79649.7	282.223	186.389	0.768	0.126
	NN	30640.126	174.974	100.613	0.28	0.664
76/24	kNN	25283.363	159.007	104.36	0.398	0.723
	RF	21965.11	148.206	82.017	0.236	0.759
	xGBoost	30710.244	175.243	94.917	0.254	0.663
	AdaBoost	24472.549	156.437	82.51	0.248	0.732
	SGD	70389.815	265.311	148.03	0.468	0.215
	NN	29660.454	172.128	98.913	0.275	0.68
78/22	kNN	26761.89	163.591	106.777	0.396	0.72
	RF	23477.5	153.224	86.017	0.24	0.754
	xGBoost	29494.292	171.74	96.41	0.281	0.691
	AdaBoost	28210.849	167.919	94.007	0.262	0.704
	SGD	53242.363	230.743	156.642	0.654	0.442
	NN	28680.782	169.283	97.2135	0.271	0.697
80/20	kNN	13563.433	116.462	83.968	0.370	0.860
	RF	11508.981	107.28	63.765	0.203	0.881
	xGBoost	16178.28	127.194	71.784	0.246	0.833
	AdaBoost	11521.525	107.338	64.203	0.202	0.881
	SGD	37444.387	193.506	133.337	0.62	0.613
	NN	27701.11	166.437	95.514	0.266	0.713

**Table 2.** Shows the model assessment for different split ratios.

testing subsets based on a specified ratio. The training subset is passed to the selected machine learning model widgets, while the Prediction widget is used to evaluate model performance on either the training or testing data, depending on the evaluation phase. The Data Table widget displays the predictions alongside the actual values. This pipeline-based architecture ensures a strict separation of training and testing data, effectively preventing data leakage. At every stage, such as data importing, splitting the data into training and testing sets, and passing the training set from the Data Split widget to the machine learning models for training, the software ensures transparency by displaying the data and its description. Additionally, the Data Table widget provides metadata, including the number of instances, features, and data types, which further confirms the integrity of the data split and confirms data leakage.

The performance of ML models depends significantly on proper hyperparameter tuning, as incorrect configurations can lead to suboptimal predictions by failing to minimize the error function effectively. Additionally, the training process dynamics plays a crucial role in determining overall model performance<sup>49,50</sup>. For illustrative purposes, Table 3 presents selected results from the hyperparameter tuning process for the AdaBoost model. Parameters such as the number of estimators, regression loss function (Linear, Square, Exponential), and learning rate were varied to assess their impact on model performance. The results demonstrate how changes in these hyperparameters influenced evaluation metrics, guiding the selection of the most effective configuration. Table 4 presents the optimal hyperparameters selected for the evaluated models.

Hyperparameter tuning was performed to identify the optimal configuration at which each model delivers the best predictive performance. This process involved systematically varying the values of key hyperparameters and observing changes in performance metrics such as R<sup>2</sup>, RMSE, and MAE. The tuning was conducted iteratively: for each set of hyperparameters, the model was evaluated, and adjustments were made until the configuration yielding the highest accuracy was identified.

### Model accuracy assessment

To ensure a comprehensive evaluation of the predictive performance of the machine learning models, six (6) statistical metrics were employed, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE),

No. of estimator	Regression Loss function	Learning Rate	R <sup>2</sup>	RMSE	MAE	
19	Linear	0.10000	0.865	114.296	65.270	
	Square		0.863	115.088	64.989	
	Exponential		0.861	115.821	68.733	
19	Linear	0.110000	0.861	115.804	65.885	
	Square		0.846	122.153	72.447	
	Exponential		0.866	113.703	68.312	
19	Linear	0.15000	0.872	111.064	68.757	
	Square		0.861	116.031	68.751	
	Exponential		0.867	113.185	63.702	
14	Linear	0.15000	0.875	109.841	64.542	
	Square		0.863	114.937	69.748	
	Exponential		0.854	118.649	68.665	
15	Linear	0.15000	0.881	107.338	64.203	
	Square		0.863	115.180	69.128	
	Exponential		0.859	116.591	68.104	

**Table 3.** Hyperparameters tuning for AdaBoost.

Model	Hyperparameters	Optimal
Random Forest	No. of Trees	515
	No. of attributes considered at each split	3
Neural Network	No. of Hidden Layers	10
	No. of Neurons in each layer	10
	Activation Function	Relu
	Solver Function	Adam
	Regularization $\alpha$	0.0006
	Maximum No. of iterations	200
Gradient Boost	No. of Trees	80
	Learning Rate	0.09
	Regularization Lambda	0.008
	Limit Depth of Individual Tress	3
	Fraction of training instances	1
	Fraction of features for each tree	1
	Fraction of features for each level	1
Fraction of features for each split	1	
Ada Boost	No. of estimators	15
	Learning Rate	0.1500
	Fixed seed for random generator	12
	Regression loss function	Exponential
kNN	No. of Neighbors	14
	Metric	Mahalanobis
	Weight	by distances
Support Gradient Descent	Classification loss Function	Modified Huber
	Regression loss function	$\epsilon$ insensitive, $\epsilon=0.11$
	Regularization	Elastic Net, Mixing = 0.14
	Strength $\alpha$	0.00001
	Learning Rate	Optimal
	Number of iterations	11,000
	Tolerance (Stopping Criteria)	0.001

**Table 4.** Optimal hyperparameters for the evaluated models.

R-square ( $R^2$ ), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE)<sup>51–53</sup>. Each metric provides distinct insights into model accuracy and generalization ability. The mathematical formulations for these metrics are presented in Eqs. (2)–(6).

$R^2$  is the proportion of variance in observed values accounted for by the model predictions, and is a measure of the strength of the linear relationship. The greater the value of  $R^2$ , the closer the fit. MAE is the average absolute difference between predicted and actual values and is an easy-to-compute measure of prediction accuracy. RMSE, being the square root of MSE, punishes large errors more than MAE and is thus good for a sense of how large differences in predictions are. MAPE captures the error as a proportion of observed values and thus can be compared across datasets of varying scales. Lastly, MSE calculates the average squared difference between predicted and observed values and is more concerned with large errors and is useful in general model assessment.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

Where;

$y_i$  Actual value of the dependent variable for the  $i$ -th data point.

$\hat{y}_i$  Predicted value for the  $i$ -th data point.

$\bar{y}$ : Mean of the actual values of the dependent variable.

$n$ : Total number of data points.

Additionally, we utilized a thorough metric, OBJ<sup>54</sup> to assess the accuracy of the model. This metric integrates RMSE, MAE, and  $R^2$  evaluating them simultaneously for both the training and validation datasets as given in Eq. 7.

$$OBJ = \frac{n_{tr}}{n} \left( \frac{RMSE_{tr} + MAE_{tr}}{R_{tra}^2 + 1} \right) + \frac{n_{test}}{n} \left( \frac{RMSE_{test} + MAE_{test}}{R_{test}^2 + 1} \right) \quad (7)$$

A precise model is expected to demonstrate high scores in  $R^2$  while simultaneously maintaining lower scores in RMSE, MSE, MAPE, MAE, and OBJ.

### Models overfitting prevention and control

Overfitting typically occurs when a model performs significantly better on the training dataset compared to the testing dataset<sup>55</sup>. While there is no universally strict threshold, substantial discrepancies in performance metrics; such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination ( $R^2$ ), are widely considered indicative of overfitting.

To mitigate overfitting, several well-established practices are recommended, such as randomizing the dataset prior to training, adopting a suitable train-test split ratio<sup>48</sup> increasing the dataset size (through the addition of real-world data or augmentation), and using ensemble models that are less prone to overfitting<sup>36</sup> such as those based on bagging and boosting techniques.

In this study, the following preventive measures were implemented:

- The dataset was randomized by shuffling the rows to eliminate any inherent order or bias.
- A commonly recommended 80/20 train-test split was applied to ensure sufficient training while retaining adequate data for model evaluation. The advantage of this split is proved by evaluating model performance in testing in the model development section.
- Ensemble models including Random Forest, AdaBoost, and xGBoost were selected due to their inherent robustness against overfitting by aggregating predictions from multiple weak learners.
- Close monitoring of model performance in both training and testing phases.

Increasing the dataset size by collecting more real-world data was not feasible, and data augmentation was deliberately avoided to maintain data integrity and prevent the introduction of artificial noise.

### Model interpretability

Interpreting machine learning models is essential for understanding their decision-making process and ensuring reliability<sup>56</sup>. This study utilizes two widely recognized interpretability techniques: SHapley Additive exPlanations (SHAP) and Partial Dependence Plots (PDP).

SHAP, derived from Shapley values in cooperative game theory, quantifies the contribution of each feature to the model's predictions. It provides locally accurate and consistent attributions by distributing the predicted outcome fairly among input features. This makes SHAP a powerful tool for identifying key variables influencing the model's decisions<sup>56</sup>.

PDP, on the other hand, offers a global perspective by illustrating the marginal effect of one or more features on the predicted output while holding other variables constant. This visualization helps in understanding feature influence, detecting nonlinear relationships, and capturing feature interactions within the model<sup>57</sup>.

By combining SHAP and PDP, this study ensures a comprehensive interpretability framework. SHAP provides granular, instance-level insights into feature contributions, while PDP offers a broader understanding of overall feature trends. Together, these techniques enhance the model's transparency and facilitate a more informed analysis of its predictions.

## Results and discussion

### Regression slope analysis

Figure 8 illustrates the regression plots for the developed machine learning models. The optimal prediction line represents a perfect alignment between observed and predicted data points. Predictive accuracy improves when data points get closer to the ideal line, indicating a high correlation between actual and predicted values.

A Regression slope (RS) value greater than 0.8 is generally taken to indicate strong correlation between model-estimated values and actual values<sup>18,57,58</sup>. For the present study, all models produced high RS values for training and testing phases, except for the SGD model in both phases and the NN model during testing. Of the six machine learning models that were developed and evaluated, AdaBoost, kNN, RF, and xGBoost were found to perform most reliably, showing the highest concordance between predicted and actual values.

### Performance evaluation of the proposed models

The performance of all developed machine learning models was evaluated using five statistical metrics, with their values presented in Table 5 for both training and testing. In the training phase, all models, except for SGD, demonstrated strong predictive capabilities, achieving  $R^2$  values exceeding 0.92. Similarly, in the testing phase, most models-maintained  $R^2$  values above 0.83, with the exception of the NN and SGD models.

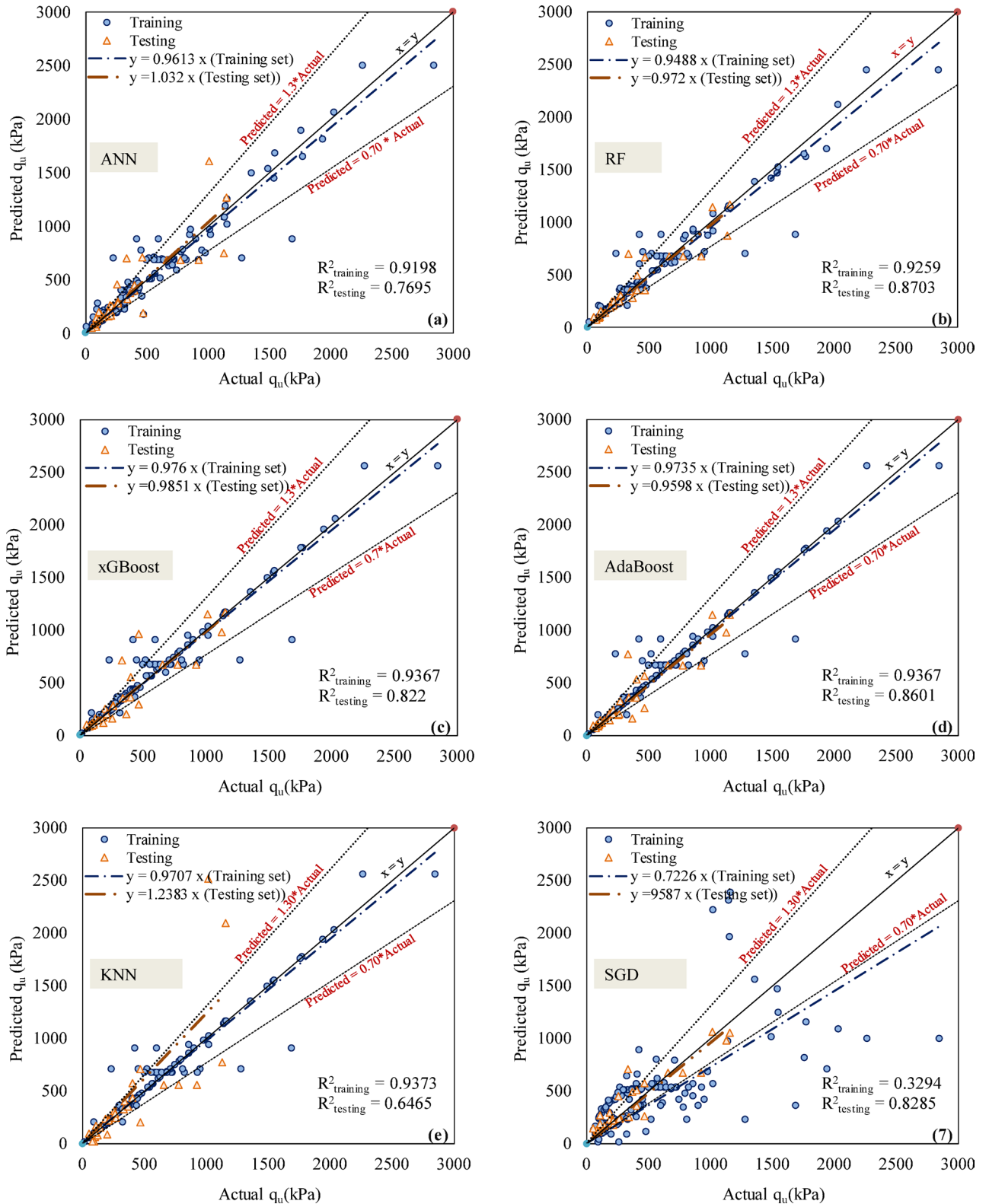
Certain models, such as the NN, exhibited high accuracy during training but did not generalize as well in testing. Conversely, some models that performed moderately well in training, such as RF, showed remarkable improvement in testing, achieving  $R^2$  values of 0.931 and 0.881 for training and testing, respectively. Overall, AdaBoost emerged as the best-performing model across both phases, with  $R^2$  values of 0.939 in training and 0.881 in testing, indicating its robustness and superior generalization ability.

The ranking system in Table 5 integrates multiple model evaluation metrics, including MSE, RMSE, MAE, MAPE, and  $R^2$ , to assess the overall performance of different models. In the training dataset, kNN outperforms other models by ranking first in all metrics, achieving the lowest cumulative rank of 5, indicating the best performance. AdaBoost follows closely with a cumulative rank of 10, while xGBoost, RF, and NN rank third, fourth, and fifth with cumulative ranks of 16, 19, and 25, respectively. SGD performs the worst, with a cumulative rank of 30. In the testing, RF and AdaBoost excel, both achieving the best cumulative rank of 9, demonstrating strong generalization capabilities. kNN and xGBoost follow with cumulative ranks of 19 and 21, respectively, while NN ranks fifth with a cumulative rank of 28. SGD again performs poorly, ranking last with a cumulative rank of 35. Notably, kNN, which performed exceptionally well in training, shows a decline in testing, suggesting potential minimal overfitting, as the decline is small. Conversely, Random Forest and AdaBoost demonstrate consistent performance across both datasets, highlighting their reliability. Overall, based on the combined ranking system of both training and testing datasets, the models can be ordered as follows: AdaBoost > kNN > Random Forest > xGBoost > Neural Network > Stochastic Gradient Descent. A common practice is to rank the models based on its performance in the testing. The model ranking based on performance in the testing phase is as; AdaBoost = RF > kNN > xGBoost > NN > SGD. This analysis underscores the importance of evaluating models on both training and testing datasets to ensure robustness and generalization.

The absolute error representation in the developed models' predictions for the UBC of shallow foundations is illustrated in Fig. 9. The comparison of the evaluated models' results is also shown in the figure with distinct datasets for training, and testing. Among the models, AdaBoost had the lowest average error (48.05), followed by xGBoost (53.94), RF (64.01), and KNN (65.02). NN had a slightly higher error (81.54), while SGD performed the worst with 204.16. This highlights AdaBoost as the most accurate model, with SGD showing significantly higher errors.

For maximum errors (Fig. 9), AdaBoost (771.79) and xGBoost (786.27) had the lowest values, while kNN (1502.01) and SGD (1850.20) exhibited the highest. In terms of minimum error, AdaBoost (0.00) and xGBoost (0.09) performed best, while NN (1.79) and SGD (1.17) showed slightly higher values.

Overall, AdaBoost provided the most precise predictions, while SGD demonstrated the least accuracy. The results confirm that the selected ML models can effectively predict the UBC of shallow foundations, with varying degrees of precision.



**Fig. 8.** Regression Slope Analysis of established models.

Models	MSE	RMSE	MAE	MAPE	R <sup>2</sup>	OBJ	Total Rank
<b>Training Data</b>							
Neural Network	18878.64 [5]	137.40 [5]	78.154 [5]	0.221 [5]	0.924 [5]	-	25
Random Forest	17124.865 [4]	130.862 [4]	64.073 [4]	0.152 [3]	0.931 [4]	-	19
Extreme Gradient Boosting	16557.398 [3]	128.676 [3]	61.563 [3]	0.16 [4]	0.933 [3]	-	16
AdaBoost	15309.387 [2]	123.731 [2]	44.798 [2]	0.086 [2]	0.938 [2]	-	10
kNN	14971.747 [1]	122.359 [1]	42.698 [1]	0.079 [1]	0.939 [1]	-	5
Stochastic Gradient Descent	112458.28 [6]	335.348 [6]	222.424 [6]	0.677 [6]	0.544 [6]	-	30
<b>Testing Data</b>							
Neural Network	27701.110 [5]	166.437 [5]	95.514 [5]	0.266 [4]	0.713 [4]	135.47 [5]	28
Random Forest	11508.981 [1]	107.280 [1]	63.705 [1]	0.203 [2]	0.881 [1]	103.07 [3]	9
Extreme Gradient Boosting	16178.28 [4]	127.194 [4]	71.784 [3]	0.246 [3]	0.833 [3]	104.79 [4]	21
AdaBoost	11521.525 [2]	107.338 [2]	64.203 [2]	0.202 [1]	0.881 [1]	91.01 [1]	9
kNN	13563.433 [3]	116.462 [3]	83.968 [4]	0.370 [5]	0.860 [2]	92.91 [2]	19
Stochastic Gradient Descent	37444.387 [6]	193.506 [6]	133.337 [6]	0.620 [6]	0.613 [5]	343.11 [6]	35

**Table 5.** Shows assessment factors values for model training dataset

### Overfitting analysis

As outlined in the methodology section, four strategies were adopted to control overfitting: data randomization, an 80/20 train-test split, the selection of some ensemble models, and close monitoring of training and testing performance metrics.

The dataset was first randomized and then split using the Data Sampler widget in Orange Data Mining, with 80% allocated to training and 20% to testing, in accordance with recommendations in relevant literature and as proved in the model development section. Model generalization was assessed by comparing evaluation metrics across both datasets with a special focus on model performance in testing.

Furthermore, ensemble models such as AdaBoost, xGBoost, and RF were intentionally chosen due to their well-known resistance to overfitting<sup>36</sup>. These models combine the outputs of multiple base learners, thereby reducing variance and improving generalization.

Most models demonstrated nearly consistent performance between training and testing datasets, suggesting very slight to no overfitting. For example, the AdaBoost model achieved closely aligned R<sup>2</sup> values of 0.939 (training) and 0.881 (testing), highlighting its strong generalization capability and no sign of overfitting. Similarly, RF and kNN also maintained stable performance across both phases. In contrast, the NN model showed signs of overfitting, with a high training R<sup>2</sup> (0.924) and a notably lower testing R<sup>2</sup> (0.713), indicating that it memorized the training data but failed to generalize. The SGD model, on the other hand, exhibited poor results on both datasets, indicating underfitting and a failure to learn the underlying data patterns.

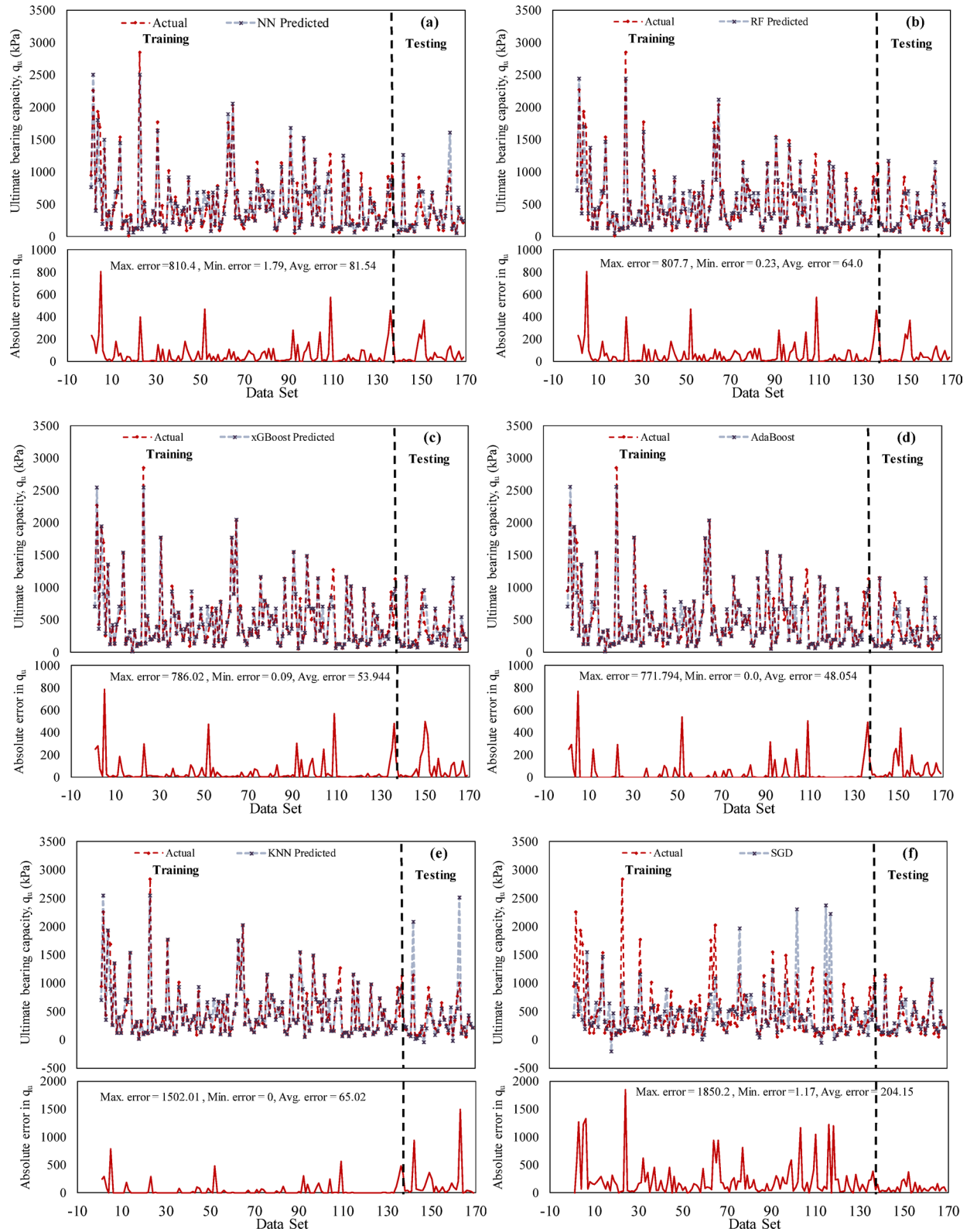
### Enhanced explainability of the model

Interpreting machine learning predictions is often challenging without incorporating mathematical reasoning, theoretical validation, and an understanding of the mechanisms driving the model's outputs<sup>59</sup>. To address this, this study employs Shapley Additive Explanations (SHAP) and Partial Dependence Plots (PDP) to enhance the interpretability of the developed models. These techniques provide both local and global insights into feature importance and model behavior.

### SHAP explanation

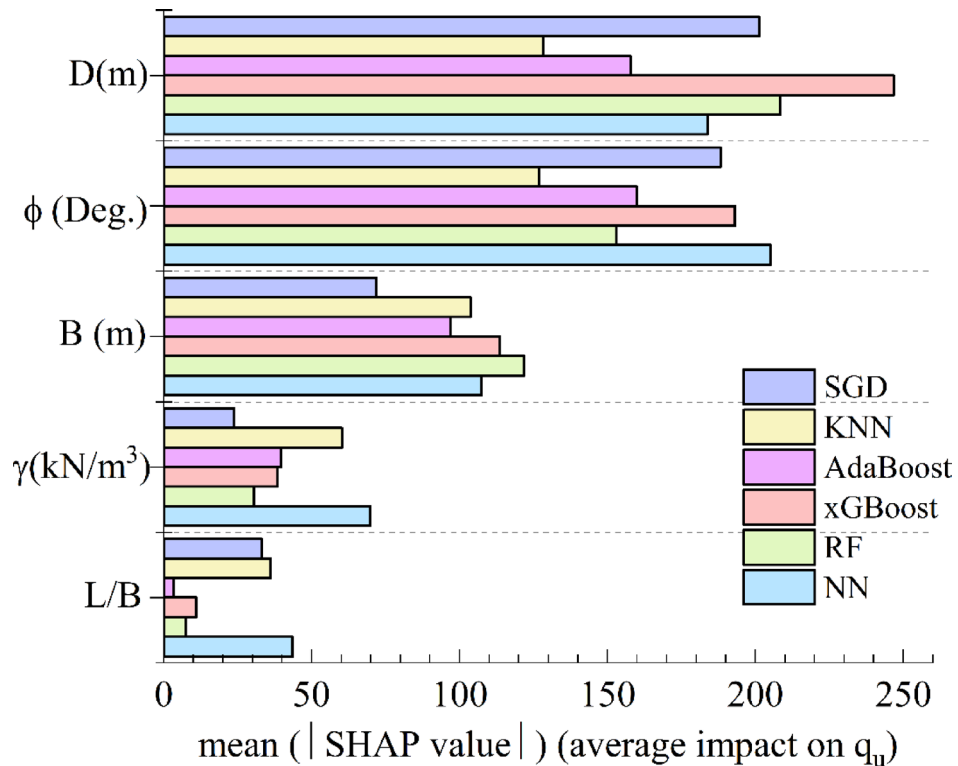
Two types of SHAP plots are generated for all the developed models: the Mean SHAP Value Plot and the SHAP Summary Plot. The Mean SHAP Value Plot (Fig. 10) illustrates the average impact of input parameters on the predicted UBC across various machine learning models. Foundation depth (D) exhibits the highest SHAP value in most models, reaffirming its dominant influence. However, AdaBoost ranks the angle of internal friction ( $\phi$ ) as the most influential factor, followed by D, following the order  $\phi > D > B > \gamma > L/B$ . This aligns with previous studies<sup>3,30</sup> which identified  $\phi$  as the most critical geotechnical parameter and D as the most significant geometric factor. Given its alignment with prior studies and experimental findings, AdaBoost effectively captures the most accurate trend, resulting in superior predictive performance. This study highlights a limitation in R. Zhang et al.'s<sup>60</sup> ranking of parameters ( $\phi > B > D > \gamma > L/B$ ), which contradicts Meyerhof's findings. The results from this research show that D has a more significant impact than suggested by Zhang et al.<sup>60</sup>, aligning more closely with Meyerhof's framework and reinforcing the need for a reevaluation of parameter significance in UBC predictions.

xGBoost, NN, RF, and K-Nearest kNN prioritize D over  $\phi$ , following the order  $D > \phi > B > \gamma > L/B$ . In contrast, SGD and NN exhibit distinct trends, ranking parameters as  $D > \phi > B > L/B > \gamma$  and  $D > \gamma > B > \phi > L/B$ , respectively.



**Fig. 9.** Absolute error representation in the established models for predicting bearing capacity of shallow foundation.

The lower ranking of  $\gamma$  in SGD suggests its limited contribution to linear regression-based models. Conversely, NN assigns greater importance to  $\gamma$ , indicating that deep learning models capture complex interactions between  $\gamma$  and other parameters, which traditional models may overlook. These findings reinforce the dominance of D



**Fig. 10.** SHAP features.

and  $\phi$  in bearing capacity estimation while demonstrating how different ML models interpret feature importance with varying sensitivities.

To investigate the impact of input parameters on the predicted UBC, the SHAP summary plot was employed to illustrate each feature's contribution to the model output, as depicted in Fig. 11. Red points represent higher feature values, while blue points indicate lower ones; a positive SHAP value signifies a favorable effect on UBC, and a higher absolute value denotes stronger influence. Conversely, negative SHAP values reduce UBC, highlighting detrimental effects of specific parameter ranges.

In the xGBoost model, foundation depth (D) and friction angle ( $\phi$ ) exhibit the widest SHAP distributions, emphasizing their dominant roles. Foundation width (B) exerts a moderate impact, whereas unit weight ( $\gamma$ ) and length-to-width ratio (L/B) show narrower spreads, suggesting lesser influence. Positive SHAP values generally increase UBC, aligning with established geotechnical insights that underscore the synergy between soil properties and foundation geometry.

In the AdaBoost model,  $\phi$  emerges as the most influential parameter, followed closely by D, indicating a balanced interplay between soil friction and foundation depth. B and  $\gamma$  have moderate effects, while L/B remains least critical. This ordering supports prior literature emphasizing  $\phi$  as a pivotal soil property, and the model's sensitivity to friction angle aligns with its robust predictive accuracy.

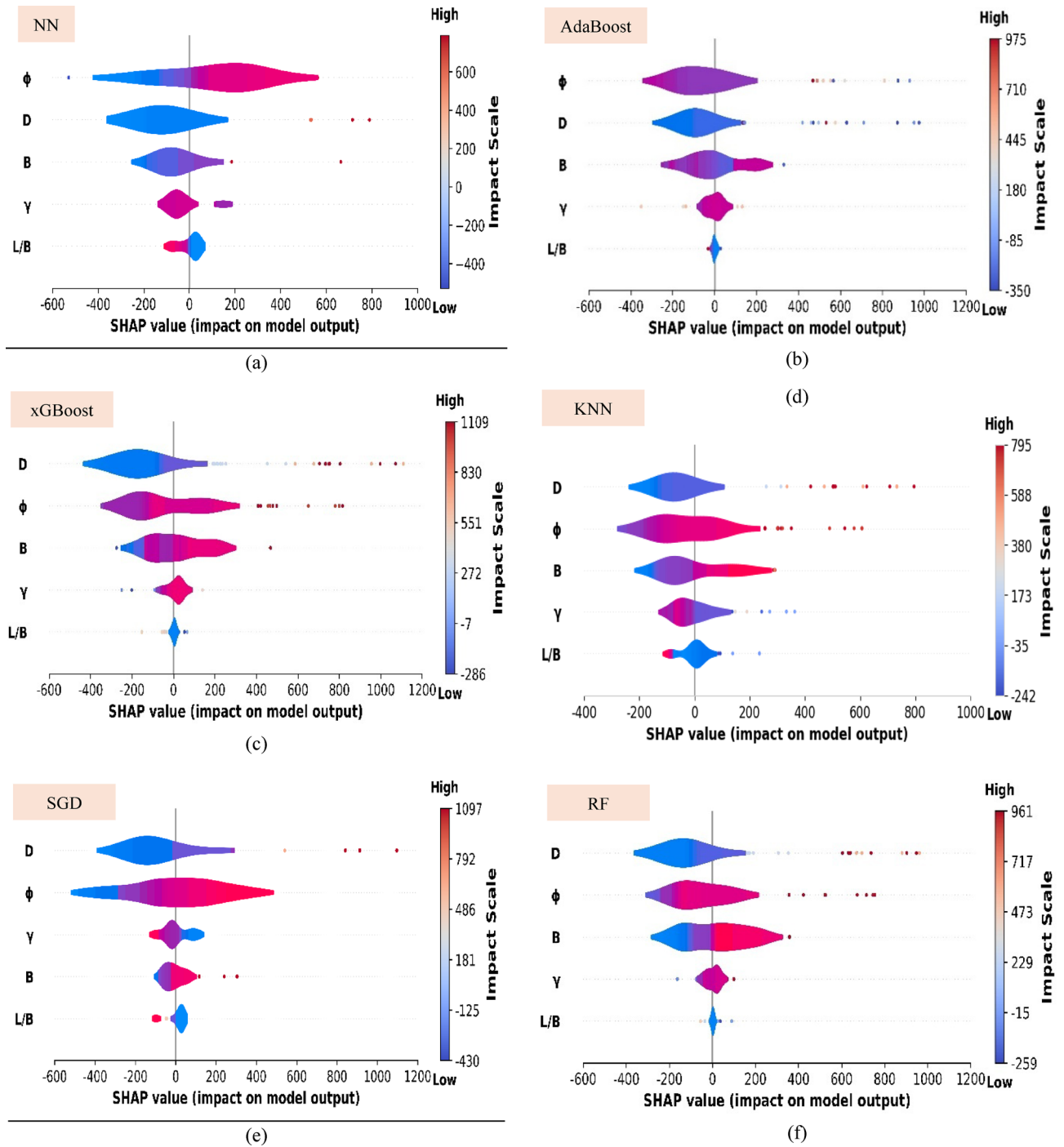
In the kNN model, D and  $\phi$  dominate the SHAP distribution, confirming their importance. B demonstrates moderate influence, whereas  $\gamma$  and L/B appear comparatively minor. Red regions at higher SHAP values typically increase UBC, reflecting the local, distance-based nature of kNN, which captures interactions between soil and geometric factors to varying degrees.

In the NN model,  $\phi$  has the largest spread of SHAP values, highlighting its primary role, while D also contributes substantially. B and  $\gamma$  display moderate effects, and L/B remains marginal. Positive SHAP values shift UBC upward, illustrating the NN's capacity to model nonlinear relationships between soil properties and foundation characteristics.

In the RF model, D yields the greatest SHAP range, followed by  $\phi$ , indicating a balanced focus on geometry and soil friction. B shows moderate significance, whereas  $\gamma$  and L/B exhibit narrower distributions. Positive SHAP values raise UBC, mirroring ensemble-based insights that consistently emphasize D and  $\phi$  as key drivers in bearing capacity estimation.

In the SGD model, D again ranks highest, with  $\phi$  next in importance, but  $\gamma$ , B, and L/B display narrower spreads. Positive SHAP values correlate with increased UBC, while negative ones reduce it. The linear optimization framework of SGD may limit its ability to capture more complex interactions, accounting for its distinct feature hierarchy relative to other models.

The SHAP analysis clearly identified the angle of internal friction ( $\phi$ ) as the most influential parameter across the developed models, particularly in AdaBoost, where it contributed most significantly to the model's predictive output. This is consistent with classical bearing capacity theories for cohesionless soils, where  $\phi$  directly governs



**Fig. 11.** Showing SHAP summary plots for (a) NN (b) AdaBoost (c) xGBoost (d) kNN (e) SGD (f) RF.

shear strength and, consequently, the ultimate bearing capacity (UBC). A higher  $\phi$  enhances resistance along potential failure surfaces, making it a critical parameter in both empirical and data-driven models. Conversely, the length-to-width ratio (L/B) exhibited the lowest SHAP values in all models, indicating minimal influence on the predicted UBC. While L/B may affect stress distribution patterns, its indirect effect renders it less impactful relative to soil strength and geometric parameters such as  $\phi$ , D, and B. This ranking of features not only enhances model interpretability but also offers practical insights, underscoring the need to prioritize accurate assessment of  $\phi$  in field investigations while suggesting that simplifications in L/B may be acceptable in preliminary design phases.

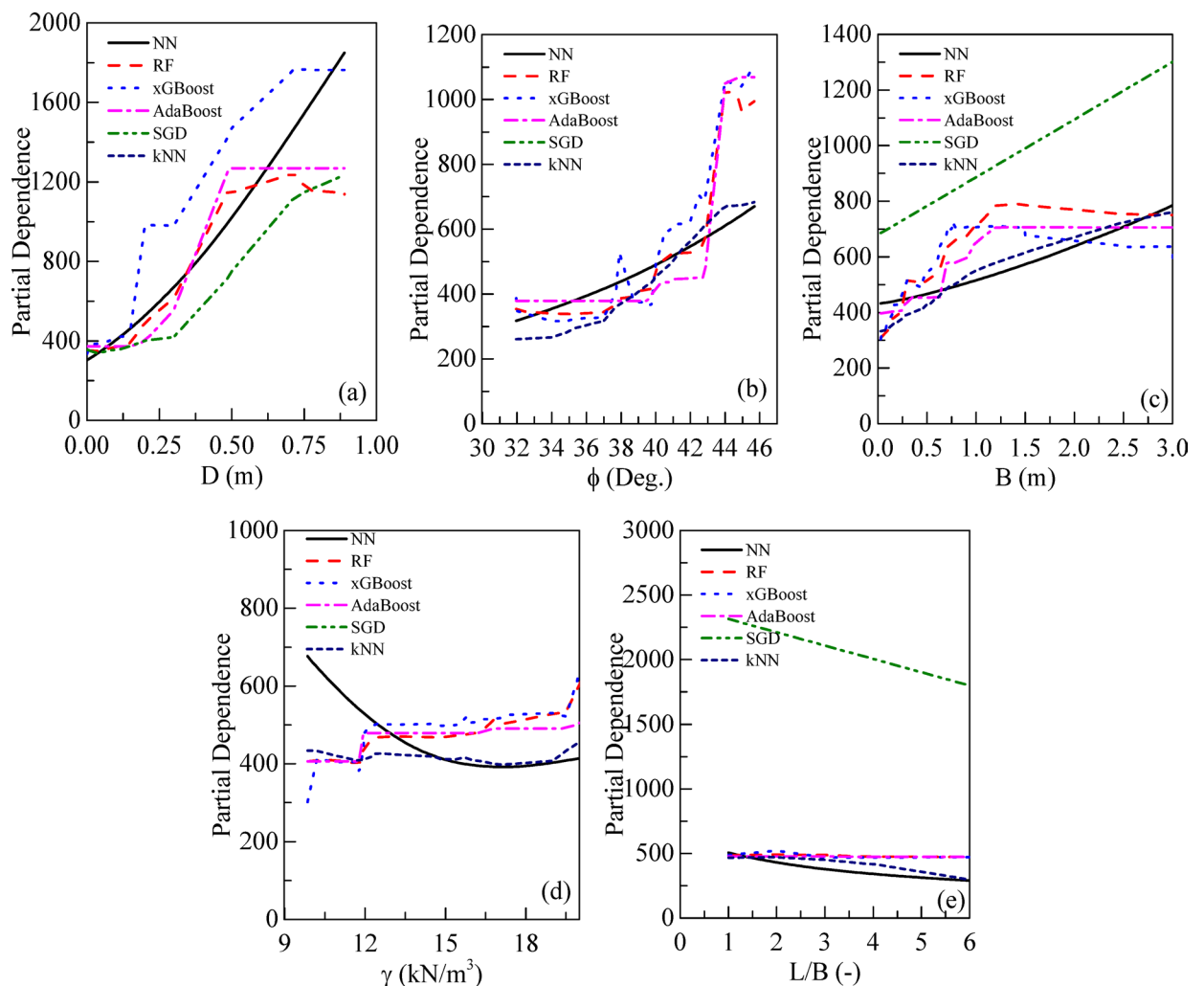
### Partial dependence plots (PDPs) interpretation

Partial dependence plots (PDPs) illustrate how variations in input parameters influence a model's predicted output. These plots provide insights into how different machine learning models capture trends between input variables and the target variable. The PDPs for all models are presented in Fig. 12, and their interpretations are summarized as follows:

All models exhibit a nonlinear, nearly exponential increase in UBC with increasing  $D$ , confirming that deeper foundations generally provide greater resistance. This behavior aligns with both theoretical<sup>3</sup> and physical expectations and is further substantiated by SHAP analysis, which indicates a consistently positive impact of depth on UBC. Notably, the AdaBoost model initially follows this trend but eventually stabilizes, suggesting a diminishing marginal influence of depth beyond a certain threshold, a phenomenon that may be attributed to soil confinement effects or bearing capacity limits.

The influence of the angle of internal friction ( $\phi$ ) follows a sharp nonlinear increasing trend across most models, indicating its critical role in governing bearing capacity. The AdaBoost model, however, displays a distinct pattern, remaining constant up to a specific threshold before gradually increasing, followed by a sudden surge. This behavior suggests that when cohesionless soil behavior dominates, the contribution of  $\phi$  to bearing capacity becomes highly pronounced, potentially approaching an exponential relationship. These observations align with established theoretical frameworks and previous empirical studies<sup>3,60</sup> with SHAP analysis further confirming the significant contribution of  $\phi$  to ultimate bearing capacity.

Foundation width ( $B$ ) exhibits a generally increasing trend, demonstrating that wider foundations contribute to higher bearing capacity. However, some models indicate a plateauing effect at larger widths, suggesting diminishing returns beyond a certain threshold. This behavior aligns with established geotechnical principles<sup>60</sup> as increasing width enhances UBC primarily by increasing the effective stress distribution, but excessive width may lead to settlement effects that counterbalance the gains. These findings highlight the need to consider optimal width ( $B$ ) dimensions in foundation design to maximize efficiency.



**Fig. 12.** Shows PDP for the developed models for input parameter (a) Depth ( $D$ ), (b) Angle of internal friction ( $\phi$ ), (c) Width ( $B$ ), (d) Unit weight ( $\gamma$ ) and (e)  $L/B$ .

Approach	Error measures					
	R (%)	RMSE	MAE	SI (%)	I <sub>a</sub>	DR (mean ± std)
Terzaghi (1943)	84.9	264.49	160.36	56.46	0.91	0.99 ± 0.41
Meyerhof (1963)	93.5	163.142	108.33	34.82	0.96	0.85 ± 0.28
Hansen (1970)	94.3	211.17	140.83	45.07	0.92	0.69 ± 0.25
Vesic (1975)	93.6	164.94	111.07	35.2	0.96	0.84 ± 0.28
Shahnazari-MLR (2012)	72	349.86	292.2	74.68	0.82	0.80 ± 1.61
Shahnazari-MGP (2012)	93.7	256.72	132.83	54.8	0.94	0.84 ± 0.39
Tsai-GP (2012)	73.2	522.19	242.31	109.46	0.45	0.75 ± 0.27
Sadrossadat-LGP (2013)	95.9	177.99	120.11	37.99	0.96	0.72 ± 0.37
Omar-ANN (2018)	94.6	27.434	11.117	-	-	-
Khorrami-M5 (2020)	96	130.14	86.08	27.78	0.98	0.99 ± 0.23
Zhang and Xue- MEP (2022)	88	-	-	-	-	-
Alzabeebee. S., et. al.EPR-MOGA <sup>19</sup>	98.48	110.5	61.5	-	-	-
Kumar-ANN-ICA(2024)	95.47	0.164	0.133	-	-	-
AdaBoost - This study	95.35	115.534	54.500	-	-	-

**Table 6.** Comparison of the proposed model in this study with literature.

The effect of unit weight of soil ( $\gamma$ ) on UBC is relatively moderate compared to other parameters. Except for the NN model, all models depict a steady but gradual increase, indicating that while higher  $\gamma$  contributes positively to bearing capacity, its impact remains less pronounced. This aligns with theoretical expectations, as UBC is more dominantly controlled by parameters such as B and  $\phi$ . The comparatively weaker influence of  $\gamma$  is likely due to its indirect role in influencing stress distribution rather than directly governing failure mechanisms.

The length-to-width ratio (L/B) exhibits a consistent decreasing trend across all models, reinforcing theoretical predictions that elongated foundations experience reduced UBC. This trend can be attributed to stress redistribution effects, where an increased L/B ratio leads to a more uniform load distribution but reduces the confining effects that contribute to higher resistance. These findings align with existing literature<sup>60</sup> suggesting that optimizing L/B is crucial in foundation design to achieve an efficient balance between load-bearing performance and structural stability.

### Comparison with models from literature

The predictive performance of the proposed AdaBoost model was compared with empirical, statistical, and machine learning-based models from the literature using key error metrics, including RMSE, MAE, Correlation Coefficient (R) (Table 6). The results indicate that AdaBoost outperforms classical empirical models, such as Terzaghi, Meyerhof, Hansen, and Vesic, which rely on simplified assumptions and exhibit higher RMSE and MAE values.

Compared to Khorrami-M5<sup>22</sup> and Zhang & Xue-MEP<sup>60</sup> AdaBoost outperformed both models with lower RMSE and MAE, and higher R<sup>2</sup> making it a more reliable and stable alternative. Omar-ANN<sup>61</sup> has lower R values than AdaBoost, although it has achieved lower RMSE and MAE, required extensive hyperparameter tuning, making AdaBoost a more practical and efficient option. The AdaBoost model ranked second overall among machine learning models, following Kumar-ANN-ICA<sup>62</sup> which achieved the lowest error values. However, AdaBoost exhibited a high R<sup>2</sup> value in training (96.90%), beating Kumar ANN-ICA in this metric.

### Conclusion

This study developed six machine learning models; NN, kNN, xGBoost, AdaBoost, NN, and SGD, to predict the UBC of shallow foundations. A comprehensive dataset comprising 169 experimental results was utilized from existing literature to train and validate these models. The input parameters included foundation width (B), depth (D), length-to-width ratio (L/B), unit weight of soil ( $\gamma$ ), and angle of internal friction ( $\phi$ ).

The models' predictive performance was assessed using five statistical metrics: R<sup>2</sup>, MAE, mean MAPE, MSE, and RMSE. Furthermore, SHAP and PDPs were utilized to enhance model interpretability.

Among the six models, AdaBoost, RF, kNN and xGBoost consistently demonstrated high predictive accuracy, with AdaBoost emerging as the top performer in both testing phase and cumulative ranking. RF, despite not having the highest accuracy in training, showed strong generalization performance in testing. On the other hand, SGD and Neural Network underperformed, struggling to capture the complex relationships between input parameters and UBC, likely due to their sensitivity to data variations.

The SHAP analysis confirmed that angle of  $\phi$  and D are the most influential factor affecting the UBC, with as the  $\phi$  top most influential parameter, aligning with insights from the previous studies. In contrast, the length-to-width ratio (L/B) had the least impact, except for the SGD model, which exhibited an inconsistent feature importance distribution. PDP analysis further revealed that most models captured a generally increasing trend between input parameters and UBC, while SGD exhibited unstable and fluctuating predictions.

This study highlights the effectiveness of AdaBoost in geotechnical applications, demonstrating their robustness in capturing nonlinear dependencies in foundation behavior. The integration of ML explainability techniques (SHAP and PDP) provides valuable insights into feature importance and model decision-making,

bridging the gap between black-box predictions and engineering reasoning. Future work can explore the impact of feature engineering, data augmentation, and hybrid ML-physics models to further enhance prediction reliability for foundation design in geotechnical engineering.

## Data availability

Data will be made available on request from the corresponding author.

Received: 1 April 2025; Accepted: 28 July 2025

Published online: 21 October 2025

## References

- Kohestani, V. R., Vosoughi, M., Hassanlourad, M. & Fallahnia, M. Bearing capacity of shallow foundations on cohesionless soils: A random forest based approach. *Civil Eng. Infrastruct. Journal* **50**(1), 35–49 (2017).
- Terzaghi, K. Theoretical soil mechanics. *Theoretical Soil Mech.* <https://doi.org/10.1002/9780470172766> (1943).
- Meyerhof, G. G. Some recent research on the bearing capacity of foundations. *Can. Geotech. J.* **1**, 16–26 (1963).
- Hansen, J. B., Brinch, J., bullet, H. & nan, S. A revised and extended formula for bearing capacity. (1970).
- Vesic, A. S. & ANALYSIS OF ULTIMATE LOADS OF SHALLOW FOUNDATIONS. *Journal Soil Mech. & Found. Div* **99** 45–73.
- Braja, M. D. Principle Of Foundation Engineering 7th edition. *Cengage Learning* 7, (2010).
- Wang, K., Ye, J., Wang, X. & Qiu, Z. The Soil-Archiving effect in Pile-Supported embankments: A review. *Build. 2024.* **14**, 126 (2024).
- Chen, D. et al. Randomly generating realistic calcareous sand for directional seepage simulation using deep convolutional generative adversarial networks. *J. Rock Mech. Geotech. Eng.* <https://doi.org/10.1016/j.jrmge.2025.01.055> (2025).
- Hua, L., Tian, Y., Gui, Y., Liu, W. & Wu, W. Semi-Analytical study of pile-Soil interaction on a permeable pipe pile subjected to rheological consolidation of clayey soils. *Int. J. Numer. Anal. Methods Geomech.* **49**, 1058–1074 (2025).
- Dong JF, Wang QY, Guan ZW, Chai HK. High-temperature behaviour of basalt fibre reinforced concrete made with recycled aggregates from earthquake waste. *Journal of Building Engineering* **48**, 103895 (2022).
- Amin, F. et al. Sustainable strengthening of concrete deep beams with openings using ECC and bamboo: an equation and data-driven approach through Abaqus modeling and GEP. *Results Eng.* **26**, 104813 (2025).
- Wang, K., Cao, J., Ye, J., Qiu, Z. & Wang, X. Discrete element analysis of geosynthetic-reinforced pile-supported embankments. *Constr. Build. Mater.* **449**, 138448 (2024).
- Shao, W. et al. The Application of Machine Learning Techniques in Geotechnical Engineering: A Review and Comparison. *Mathematics* vol. 11 Preprint at (2023). <https://doi.org/10.3390/math11183976>
- Rabbani, A., Samui, P., Kumari, S. & Optimized ANN-based approach for Estimation of shear strength of soil. *Asian J. Civil Eng.* **24**, 3627–3640 (2023).
- Rabbani, A., Samui, P. & Kumari, S. A novel hybrid model of augmented grey Wolf optimizer and artificial neural network for predicting shear strength of soil. *Model. Earth Syst. Environ.* **9**, 2327–2347 (2023).
- Zhao, H. B. & Yin, S. A CPSO-SVM model for ultimate bearing capacity determination. *Marine Georesources Geotechnology* **28**, 64–75 (2010).
- Adarsh, S., Dhanya, R., Krishna, G., Merlin, R. & Tina, J. Prediction of Ultimate Bearing Capacity of Cohesionless Soils Using Soft Computing Techniques. *ISRN Artificial Intelligence* (2012). (2012).
- Gupta, R., Goyal, K. & Yadav, N. Prediction of safe bearing capacity of noncohesive soil in arid zone using artificial neural networks. *International J. Geomechanics* **16**, 04015044 (2016).
- Alzabeebee, S., Alshkane, Y. M. A. & Keawsawasvong, S. New model to predict bearing capacity of shallow foundations resting on cohesionless soil. *Geotechnical Geol. Engineering* **41**, 3531–3547 (2023).
- Khajehzadeh, M. & Keawsawasvong, S. Artificial intelligence for bearing capacity evaluation of shallow foundation: an overview. *Geotech. Geol. Eng.* **42**, 5401–5424 (2024).
- Liu, Y. & Reports, Y. L. S. & undefined. Integrated machine learning for modeling bearing capacity of shallow foundations. *nature.com Y Liu, Y Liang Scientific Reports, 2024 nature.com* (123AD) (2024). <https://doi.org/10.1038/s41598-024-58534-5>
- Khorrami, R., Derakhshani, A. & Moayedi, H. New explicit formulation for ultimate bearing capacity of shallow foundations on granular soil using M5' model tree. *Measurement (Lond)* **163**, 108032 (2020).
- Foye, K. C., Salgado, R. & Scott, B. Assessment of variable uncertainties for Reliability-Based design of foundations. *Journal Geotech. Geoenvironmental Engineering* **132**, 1197–1207 (2006).
- GOLDER, H. Q. et al. THE ULTIMATE BEARING PRESSURE OF RECTANGULAR FOOTINGS. *Journal Institution Civil Engineers* **17**, 161–174 (1941).
- Mosallanezhad, M. & Moayedi, H. Comparison analysis of bearing capacity approaches for the strip footing on layered soils. *Arab J. Sci. Eng* **42**, 3711–3722 (2017).
- Briaud, J. L. & Gibbens, R. Large-scale load test and data base of spread footings on sand. *McLean, Virginia: Federal Highway Administration, HNR 10* (1997).
- Consoli, N. C., Schnaid, F., Phatak, D. R. & Dhonde, H. B. Behavior of five large spread footings in sand. *Journal Geotech. Geoenvironmental Engineering* **126**, 940–942 (2000).
- Cerato, A. B. & Lutenecker, A. J. Scale effects of shallow foundation bearing capacity on granular material. in *BGA International Conference on Foundations, Innovations, Observations, Design and Practice* (2003). [https://doi.org/10.1061/\(asce\)1090-0241\(2007\)133:10\(1192\)](https://doi.org/10.1061/(asce)1090-0241(2007)133:10(1192)).
- Akbas, S. O. & Kulhawy, F. H. Axial compression of footings in cohesionless soils. II: bearing capacity. *Journal Geotech. Geoenvironmental Engineering* **135**, 1575–1582 (2009).
- Padmini, D., Ilamparuthi, K. & Sudheer, K. P. Ultimate bearing capacity prediction of shallow foundations on cohesionless soils using neurofuzzy models. *Comput Geotech* **35**, 33–46 (2008).
- Khan, S. et al. Predicting the ultimate axial capacity of uniaxially loaded Cfst columns using multiphysics artificial intelligence. *Materials* **15**, 39 (2022).
- Khan, M. A. et al. Simulation of depth of wear of eco-friendly concrete using machine learning based computational approaches. *Materials* **15**, 58 (2022).
- Aldrees, A., Khan, M., Taha, A. T. B. & Ali, M. Evaluation of water quality indexes with novel machine learning and SHapley additive explanation (SHAP) approaches. *Journal Water Process. Engineering* **58**, 104789 (2024).
- Han, Q., Gui, C., Xu, J. & Lacidogna, G. A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm. *Constr. Build. Mater* **226**, 734–742 (2019).
- Biau, G. & Scornet, E. A random forest guided tour. *Test* **25**, 197–227 (2016).
- Breiman, L. Random forests. *Mach Learn* **45**, 5–32 (2001).
- Chen, T., Guestrin, C. & XGBoost A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* vols 13-17-August-2016 (2016).
- Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann Stat* **29**, 1189–1232 (2001).

39. Cover, T. M. & Hart, P. E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27 (1967).
40. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician* **46**, 175–185 (1992).
41. Schmidhuber, J. Deep Learning in neural networks: An overview. *Neural Networks* vol. 61 Preprint at (2015). <https://doi.org/10.1016/j.neunet.2014.09.003>
42. Freund, Y. & Schapire, R. E. A Decision-Theoretic generalization of On-Line learning and an application to boosting. *J Comput. Syst. Sci* **55**, 119–139 (1997).
43. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class adaboost. *Stat Interface* **2**, 349–60 (2009).
44. Bottou, L. Large-scale machine learning with stochastic gradient descent. in *Proceedings of COMPSTAT 2010–19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers* (2010). [https://doi.org/10.1007/978-3-7908-2604-3\\_16](https://doi.org/10.1007/978-3-7908-2604-3_16)
45. Kingma, D. P., Ba, J. L. & Adam A method for stochastic optimization. in *3rd International Conference on Learning Representations, ICLR - Conference Track Proceedings* (2015). (2015).
46. Kohavi, R. A. Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. in *IJCAI International Joint Conference on Artificial Intelligence* vol. 2 (1995).
47. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal Mach. Learn. Research* **7**, (2006) pp.1–30.
48. Bichri, H., Chergui, A. & Hain, M. Investigating the impact of Train / Test split ratio on the performance of Pre-Trained models with custom datasets. *International J. Adv. Comput. Sci. Applications* **15**, (2024).
49. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. in *32nd International Conference on Machine Learning, ICML 1* (2015). (2015) vol.
50. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun ACM* **64**, 107–115 (2021).
51. Hastie, T., Tibshirani, R. & Friedman, J. *Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer vol. 2nd (2009).
52. James, G., Witten, D., Hastie, T. & Tibshirani, R. ‘Data for an Introduction to Statistical Learning with Applications in R’ Package ‘ISLR’. CRAN (2017).
53. Willmott, C. J. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res* **30**, 79–82 (2005).
54. Alam, M., Chen, J., Umar, M., Ullah, F. & Shahkar, M. Predicting standard penetration test N-value from cone penetration test data using gene expression programming. *Geotech. Geol. Eng.* **42**, 5587–5613 (2024).
55. López, O. A. M., López, A. M., Crossa, D. J. & Overfitting Model tuning, and evaluation of prediction performance. *Multivar. Stat. Mach. Learn. Methods Genomic Prediction*. 109–139. [https://doi.org/10.1007/978-3-030-89010-0\\_4](https://doi.org/10.1007/978-3-030-89010-0_4) (2022).
56. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* vols 2017–December (2017).
57. Molnar, C. et al. Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. in *Communications in Computer and Information Science* vol. 1901 CCIS (2023).
58. Chen, L. et al. Development of predictive models for sustainable concrete via genetic programming-based algorithms. *Journal Mater. Res. Technology* **24**, 6391–6410 (2023).
59. Lipton, Z. C. The myths of model interpretability. *Commun ACM* **61**, 31–57 (2018).
60. Zhang, R. & Xue, X. Determining ultimate bearing capacity of shallow foundations by using multi expression programming (MEP). *Eng. Appl. Artif. Intell.* **115**, 105255 (2022).
61. Omar, M., Hamad, K., Al Suwaidi, M. & Shanableh, A. Developing artificial neural network models to predict allowable bearing capacity and elastic settlement of shallow foundation in sharjah, united Arab Emirates. *Arab. J. Geosci.* **11**, 1–11 (2018).
62. Kumar, D. R., Wipulanusat, W., Kumar, M., Keawsawasvong, S. & Samui, P. Optimized neural network-based state-of-the-art soft computing models for the bearing capacity of strip footings subjected to inclined loading. *Intelligent Syst. Applications* **21**, 200314 (2024).

## Author contributions

Jalal Shah: Investigation, Methodology, Visualization, Software, and Writing – original draft, Formal Analysis  
 Mehtab Alam: Supervision, Project administration, Software, Writing – original draft, Formal Analysis  
 Faisal Javed: Supervision, Project administration, Conceptualization, and Writing – review & editing. Muhammad Umar: Data curation, Writing – review & editing. Furqan Ahmad: Writing – review & editing.

## Funding

No funding was received for this study.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.A. or F.A.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025