Vietnam Journal of Earth Sciences, 47(3), 447–462, https://doi.org/10.15625/2615-9783/23484

Vietnam Journal of Earth Sciences
https://vjs.ac.vn/index.php/jse

Prediction of Optimum Water Content of soil using advanced machine learning methods: A comparative study of RF, SVM, and ANN models

Binh Thai Pham[1], Le Huyen Trang*[,1], Indra Prakash[2]

[1]*University of Transport Technology, Hanoi 100000, Vietnam*
[2]*Formerly Dy. Director General, Geological Survey of India, Gujarat, India*

Received 06 June 2025; Received in revised form 12 September 2025; Accepted 23 September 2025

ABSTRACT

In construction, achieving adequate soil compaction is essential for ensuring the strength and stability of geotechnical structures, with Optimum Water Content (OWC) being a critical parameter. Traditional laboratory methods for determining the OWC are accurate but often time-consuming and resource-intensive. This study investigates the potential of advanced machine learning methods: Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks with Multilayer Perceptron (ANN-MLP) to predict the OWC of soil using a curated dataset of over 214 soil samples collected from the Van Don - Mong Cai expressway construction project (Vietnam). The models were developed using input factors such as specific gravity, grain size distribution, organic content, and Atterberg limits. Among the three approaches, the RF model exhibited the best performance (R^2 = 0.84, RMSE = 1.07% and MAE = 0.78%) compared with other models such as ANN (MLP) (R^2 = 0.44, RMSE = 2.02% and MAE = 1.61%) and SVM (R^2 = 0.63, RMSE = 1.65% and MAE = 1.17%). Partial Dependence Plot (PDP) analysis further highlighted fines content, plasticity indices, and organic matter as key influencing factors with a high impact on the predictive capability of the model. The findings demonstrated that the RF model offers an accurate and efficient tool for estimating the OWC of soil, with potential to reduce reliance on extensive laboratory testing and support faster, data-driven geotechnical decision-making.

Keywords: Optimum Water Content (OWC), advanced machine learning Methods, RF, SVM, ANN, soil compaction, Partial Dependence Plot (PDP).

1. Introduction

In geotechnical engineering, determining the Optimum Water Content (OWC) of soil is essential for achieving effective compaction, which in turn governs the strength, stability, and durability of structures such as embankments, road subgrades, and foundations (Blotz et al., 1998). The OWC refers to the moisture content at which a soil achieves its maximum dry density under a given compactive effort. An accurate estimation of the OWC of soil is critical, as deviations can lead to under-compaction, reduced load-bearing capacity, long-term settlement, and, in severe cases, structural failure (Mueller et al., 2003). Beyond compaction control, knowledge of the OWC supports a range of applications, including

*Corresponding author, Email: lehuyentrang500@gmail.com

slope stability assessment, landfill design, and ground improvement techniques.

Conventionally, the OWC is determined through standardized laboratory tests such as the standard and modified Proctor compaction tests (Aragón et al., 2000). While reliable, these methods are labor-intensive, time-consuming, and impractical for rapid field assessments, particularly in large-scale construction projects. Moreover, they often overlook the complex, nonlinear interactions among various soil properties, such as grain size distribution, plasticity, and organic content, that significantly influence compaction behavior. To address these limitations, the researchers have proposed empirical and statistical models to estimate the OWC from easily measurable soil parameters (Hassan et al., 2017; Lai et al., 2017). Although these models are simple and accessible, they are typically derived from small, site-specific datasets and often lack generalizability across diverse soil types and conditions.

Recent advancements in data science have led to the growing application of machine learning (ML) techniques in geotechnical engineering, offering a promising alternative for predictive modeling (Prakash et al., 2024). For example, Benbouras and Lefilef (2023) advanced the field by employing different progressive ML models to predict the OWC and maximum dry density (MDD). The study compiled a database of 147 samples and implemented K-fold cross-validation to ensure robustness. Random Forest (RF) emerged as the top performer compared with other models such as Gaussian Process (GP), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Li et al. (2024) compared the efficacy of four popular ML algorithms, including SVM, ANN, RF, and Extreme Gradient Boosting (XGBoost), for predicting MDD and OWC using 168 soil samples. XGBoost emerged as the most accurate model ($R^2 > 0.92$), with liquid limit (LL) and plastic limit (PL) identified as the most influential input features. Khatti and Grover (2023) presented a comparative evaluation of four ML models, such as Least-Square Boost Random Forest (LSBoostRF), Long Short-Term Memory (LSTM), Least-Square Support Vector Machine (LSSVM), and ANN to predict compaction parameters of soil, namely MDD and OWC. For MDD prediction, the LSSVM model demonstrated the highest accuracy and outperformed other models. For the OWC, LSTM showed the best performance. In general, ML models such as ANN, SVM, and RF are particularly effective in capturing complex, nonlinear relationships in large and heterogeneous datasets without the need for predefined functional forms. Previously mentioned studies have demonstrated the potential of ML models to outperform traditional empirical approaches in predicting various geotechnical properties, including the OWC.

However, there is limited understanding of how different ML models perform relative to one another when applied to the OWC prediction, and how soil parameters influence the prediction outcomes (Khatti and Grover, 2023; Li et al., 2024). In addition, it is necessary to evaluate ML models in specific regions to identify the most suitable model for each dataset. In this context, the present study aims to develop and compare advanced ML models: RF, SVM, and ANN (MLP) for predicting the OWC of soils in Vietnam. A curated dataset of 214 soil samples gathered from the Van Don - Mong Cai expressway construction project was collected and used to train and evaluate the models. Various validation metrics, such as R^2, RMSE, and MAE, were selected for validation and comparison. Python software was utilized for data processing and modeling in this study.

2. Materials and Methods

2.1. Data collection and analysis

The data of this study were collected from the Van Don - Mong Cai expressway construction project (Vietnam). The research data were collected and consist of 214 sets of experimental results, evenly distributed across various locations along the entire expressway. Soil samples were directly taken from the construction site and material quarries, then transported to the laboratory for the determination of geotechnical and physical properties. It includes results from standard or modified Proctor compaction tests and routine soil classification tests, such as grain size analysis and Atterberg limits. Only datasets with complete records of the OWC and essential soil index properties were included to ensure consistency and reliability for model training and evaluation. The data is organized in a tabular format, with each row representing a unique soil sample and each column corresponding to a soil parameter. Eight input variables were selected for their geotechnical relevance to compaction behavior: Plastic Limit (PL), Silt Liquid Limit (LL), and Clay content (SC), Fine Sand content (FS), Coarse Sand content (CS), Specific Gravity (G), Organic content (O), and Plasticity Index (PI). The target variable is the OWC.

The selected input variables capture both physical and physicochemical soil properties influencing moisture-holding capacity and compaction. More specifically, G reflects the density of soil solids relative to water. Higher G typically correlates with lower OWC due to reduced pore space. CS and FS affect drainage and packing. CS promotes drainage, lowering the OWC, while fine sand increases water retention, potentially raising the OWC. SC with high surface area and electrochemical activity in finer particles increases water adsorption, leading to higher OWC. Even in small amounts, the porous nature increases water retention, elevating the OWC. LL and PL indicate moisture boundaries for soil consistency. Higher values suggest greater clay activity and moisture requirements for compaction PI measures the moisture range for plastic behavior. Higher PI, linked to greater clay content, correlates with higher OWC. This selection ensures a comprehensive representation of soil behavior, enhancing the models' ability to predict the OWC across diverse geotechnical conditions using ML.

Table 1 provides a statistical overview of the dataset, detailing the mean, standard deviation (std), minimum, 25th percentile, median (50%), 75th percentile, and maximum for each variable. More specifically, G: Mean is 22.06, but a minimum of 0 suggests potential data errors requiring preprocessing. CS and FS: Moderate variability (CS: 3–46.3%, FS: 2.5–41.5%) reflects diverse grain sizes. SC: The high mean (44.81%) and wide range (17.87–88.7%) indicate the inclusion of both coarse- and fine-grained soils. O: Low mean (1.51%) but significant for water retention. LL, PL, PI: Moderate to high plasticity (LL: 2.08–48.45%, PI: 0.91–27.48%) supports diverse soil types. OWC: Mean of 14.01% and range of 9.3–21.5% align with typical compaction requirements. The dataset's diversity across soil types supports robust predictive modeling.

Figure 1 illustrates the frequency distributions of the variables, highlighting their statistical characteristics. For G, a bimodal distribution with peaks at 20–25 and 35–40, and a spike at 0, indicating potential outliers. With CS, moderately right-skewed, with a cluster at 20–30%, reflecting coarse-grained dominance. For FS, highly right-skewed, with a peak at 2.5–10%, suggesting limited acceptable sand content. With SC, approximately normal, centered at 45–50%, ideal for modeling fine-grained soils. For O, left-skewed, concentrated at 1–2%, consistent with low organic matter in geotechnical contexts. For LL and PL, right-skewed, clustered at 35–45% (LL) and 18–22% (PL), indicating moderate to high plasticity. With

PI, relatively symmetric, uniform between 10% and 25%, capturing cohesive soils. With OWC, the distribution is symmetric, slightly platykurtic, and centered at 14–15%, making it suitable for predictive modeling. Skewed distributions (G, FS, O) and outliers suggest the need for normalization or outlier treatment to optimize ML performance.

Table 1. Statistical summary of the dataset used for OWC prediction

No.	Variable	Mean	SD	Min.	25%	50%	75%	Max.
1	G	22.057	13.296	0	9.075	24.75	31.7	51.4
2	CS	24.101	7.017	3	20.7	23.7	27.775	46.3
3	FS	9.035	6.468	2.5	4.6	7.25	11	41.5
4	SC	44.807	10.447	17.87	37.75	44.55	49.2	88.7
5	O	1.509	0.373	0.12	1.2525	1.51	1.77	2.94
6	LL	39.515	6.173	2.08	36.638	39.99	43.508	48.45
7	PL	20.318	3.068	1.17	19.293	20.835	21.888	28.49
8	PI	19.198	4.078	0.91	16.83	18.435	22.32	27.48
9	OWC	14.01	2.619	9.3	12.19	14.275	15.4	21.5

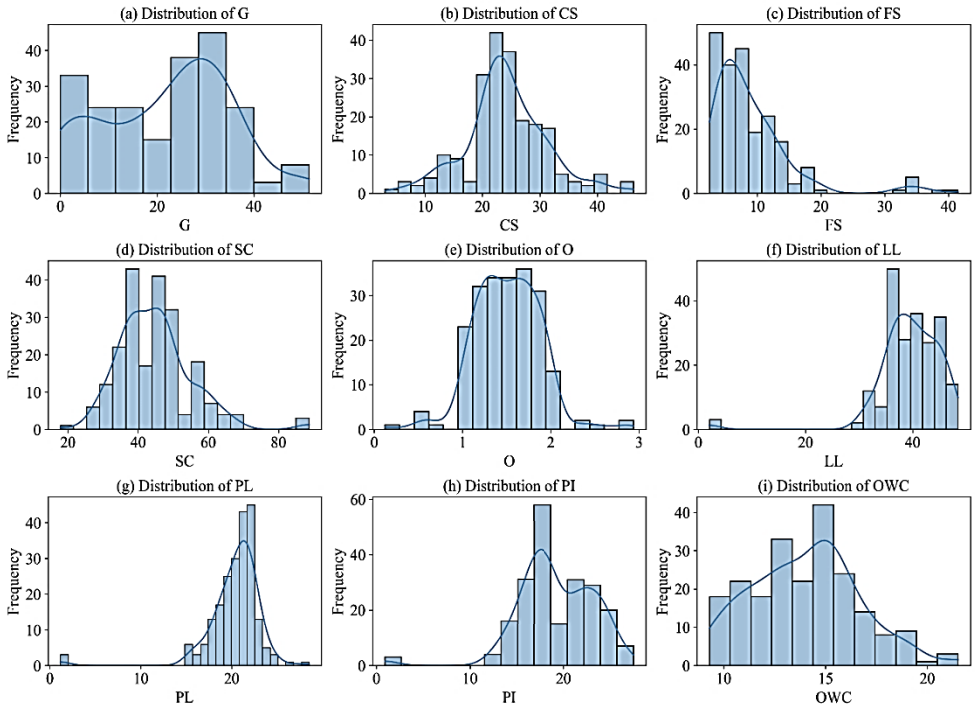

Figure 1. Frequency distributions of input and output variables used in the modeling process

Figure 2 presents a correlation matrix quantifying pairwise linear relationships between variables and the OWC, with values varying from −1 (strong negative) to +1 (strong positive). For G, there is a moderate negative correlation with OWC (−0.41), reflecting denser particles reducing water retention. Strong negative correlations with SC (−0.74) and FS (−0.59) indicate lower G in finer soils. With CS, there is a negligible correlation with OWC (−0.01), but an inverse correlation with SC (−0.29) and a positive correlation with PL (0.27). For FS, there is a weak positive correlation with OWC (0.16), as well as mild positive correlations with SC (0.22) and PL (0.21). For SC, the strongest positive correlation with OWC (0.43) is driven by high water adsorption in the fines. For O,

mild positive correlation with OWC (0.11), reflecting moisture retention by organic matter. In the case of LL, PL, PI, moderate positive correlations with OWC (LL: 0.39, PL: 0.37, PI: 0.31), emphasizing the plasticity role. Strong intercorrelations (LL–PL: 0.82, LL–PI: 0.90, PL–PI: 0.48) reflect a shared dependence on clay content. The matrix highlights the dominance of SC, LL, PL, and PI in predicting OWC, validating their inclusion. Nonlinear relationships suggest ML models must capture complex interactions for accurate predictions.

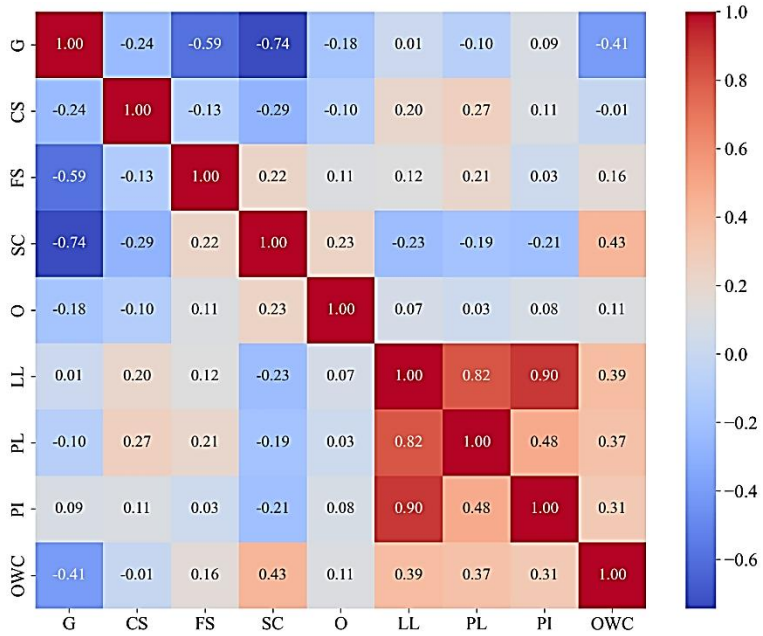

Figure 2. Correlation matrix of variables used in the modeling process

To enhance the performance and stability of ML models, especially those sensitive to feature scaling, standard normalization was applied in this study. It standardizes the data by removing the mean and scaling to unit variance. This process ensures that all features contribute equally to the learning process, reducing potential bias due to differing scales. While standard scaling does not eliminate outliers, it mitigates their impact by centering and scaling the data, which can enhance the convergence and accuracy of the ML models.

In the modeling, the data was randomly split into two parts, including a training part (70%) and a testing part (30%) used for training and validating the models, respectively. This ratio for splitting the training and testing data was proposed by previous published works (Hoang et al., 2025; Nguyen et al., 2021).

2.2. Methods used

2.2.1. Random Forest (RF)

RF, introduced by Breiman (2001), is an ensemble learning method that enhances decision tree accuracy and reduces overfitting by averaging predictions from multiple trees. RF employs bootstrap aggregating (bagging) and random feature selection to create diverse trees (Zhou et al., 2023). For each tree, a random subset of training data is sampled with replacement, and a random subset of features is used at each split, minimizing correlation and improving generalization.

Mathematically, given a training dataset $D = (x_i, y_i)_{i=1}^{N}$, where $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ represents

the vector of input features for the ^{i-}th soil sample and is the corresponding target OWC, the RF regression prediction $f(x)$ for a new x input is given by the average of the predictions from M individual trees (Breiman, 2001; Kim, 2024):

$$\hat{f}(x) = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \quad (1)$$

Where $T_m(x)$ denotes the prediction of the $m-th$ decision tree.

Each decision tree T_m is built through recursive binary splitting of the feature space to minimize the variance of the target variable within each terminal node (leaf). At each split, the algorithm selects the feature and split point that minimizes the sum of squared residuals:

$$\min_{j,s}\left[\sum_{x_i \in R_1(j,s)} (y_i - \bar{y}_{R_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}_{R_2})^2\right] \quad (2)$$

Where j indexes the features, s is the split threshold, and $R_1(j,s)$ and $R_2(j,s)$ are the two regions created by splitting the data on feature at point s. $\bar{y}R_1$ and $\bar{y}R_2$ represent the mean target values in areas R_1 and R_2, respectively.

RF uses out-of-bag (OOB) error for internal validation, leveraging unsampled data to estimate performance (Breiman, 2001). RF excels in capturing nonlinear relationships between soil parameters and OWC, is robust to noise, and provides variable importance scores. However, it can be computationally intensive and less interpretable than single trees. In this study, the hyperparameters used to train the RF include the number of trees (100), the maximum depth of the tree (set to none), and the bootstrap (set to true).

2.2.2. Support Vector Machines (SVM)

SVM is a supervised learning algorithm introduced initially by Vapnik (1995). SVMs are primarily designed for classification tasks and are based on the principle of finding an optimal hyperplane that maximizes the margin between classes in a high-dimensional feature space (Nhat et al., 2025; Vapnik, 1995). The core idea is to transform input data into a higher-dimensional space using kernel functions, where a linear separation between classes becomes possible (Vapnik, 2013). The data points that lie closest to the decision boundary, known as support vectors, are crucial in defining this hyperplane (Navidi et al., 2022).

To handle non-linear relationships between features, SVM employs kernel functions such as the linear kernel, polynomial kernel, and radial basis function (RBF) kernel (Pal et al., 2024; Yin et al., 2023). The RBF kernel is especially popular for its flexibility in modeling complex, non-linear boundaries (Nguyen et al., 2022). SVM incorporates regularization parameters to control the trade-off between maximizing the margin and minimizing classification errors, thus reducing the risk of overfitting. This makes it particularly useful for datasets with limited samples but complex feature interactions. In this study, the hyperparameters used to train the SVM include: regularization parameter (1), kernel type (RBF), epsilon (0.1).

2.2.3. Multilayer Perceptron neural network (MLP)

ANN was inspired by the structure and functioning of the human brain's neural networks (Wu and Feng, 2018). Among the various ANN architectures, the Multilayer Perceptron neural network (MLP) is one of the most extensively applied (Gardner and Dorling, 1998). Initially conceptualized in the 1960s and significantly advanced in the 1980s with the development of the backpropagation learning algorithm, the MLP has become a cornerstone in the modeling of complex, nonlinear systems across multiple domains (Gardner and Dorling, 1998), including

geotechnical engineering (Pham et al., 2019). An MLP includes an input layer, one or more hidden layers, and an output layer, each composed of interconnected processing units known as neurons (Pham, 2024). Each neuron calculates a weighted sum of its inputs and applies a nonlinear activation function to introduce nonlinearity into the model. This nonlinearity is critical, as it allows the network to learn and represent complex relationships in the data. In geotechnical applications, such as predicting the OWC of soils, MLP are particularly effective. They can model the intricate, nonlinear relationships between multiple input variables - such as grain size distribution, Atterberg limits, and other soil properties and the target output. This capability makes MLPs valuable tools for tasks involving complex data interactions that are not easily captured by traditional empirical methods.

Mathematically, the output of the *j*-th neuron in the *l*-th layer, $o_j^{(l)}$, is computed as (Wu and Feng, 2018):

$$o_j^{(l)} = f\left(\sum_{i=1}^{n_{l-1}} w_{ij}^{(l)} o_i^{(l-1)} + b_j^{(l)}\right) \quad (1)$$

Where: n_{l-1} is the number of neurons in the previous layer *(l-1),* $W_{ij}^{(l)}$ is the weight connecting the ^{i}*–th* neuron in layer *(l-1)* to the ^{j}-th neuron in layer, $o_j^{(l-1)}$ is the output from the ^{i}-th neuron in the previous layer, $b_j^{(l)}$ is the bias term for neuron *j* in layer *l*, $f(.)$ and is the activation function, typically nonlinear (e.g., sigmoid, ReLU).

For regression tasks like OWC prediction, the output layer usually employs a linear activation function to produce continuous values. In this study, the hyperparameters used to train the ANN (MLP) include: number of neurons in the hidden layer (100), maximum number of iterations (1000).

2.2.4. Validation metrics

In this study, three widely recognized validation metrics were employed to assess the regression models developed for predicting the OWC of soil: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).

R^2 measures the proportion of variance in the observed data explained by the model. As a normalized metric ranging from 0 to 1, higher values indicate a better model fit (Duc et al., 2025; Nguyen et al., 2025). R^2 provides an intuitive sense of how closely the predicted values align with the actual observations, serving as an indicator of the model's explanatory power. Mathematically, R^2 is defined as (Phan and Ly, 2024; Phung et al., 2023):

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

Where $\bar{y}$ is the mean of observed values, $\hat{y}_i$ is the predicted value, y_i is the observed value, and N is the number of samples.

RMSE represents the standard deviation of the prediction errors, effectively quantifying the average magnitude of the residuals (Pham et al., 2021). Because errors are squared before averaging, RMSE penalizes larger deviations more heavily, making it sensitive to outliers. It is expressed in the same units as the target variable, offering a direct interpretation of typical prediction error (Ngo et al., 2022; Nguyen et al., 2023):

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

MAE captures the average absolute difference between predicted and actual values. Unlike RMSE, MAE treats all errors equally, making it a more robust measure in the presence of outliers. It reflects the typical size of prediction errors without disproportionately emphasizing extreme values. MAE is expressed as (Prakash et al., 2022):

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

In this study, the combined use of R^2, RMSE, and MAE ensures a robust evaluation of the ML models' ability to accurately predict the OWC across diverse soil types, thereby supporting informed model selection and reliable deployment in geotechnical practice.

2.2.5. Taylor Diagram

The Taylor Diagram, proposed by Taylor (2001), is a graphical tool to evaluate and compare the performance of multiple ML models in predicting the OWC of soils. In the Taylor Diagram's polar coordinate system, the radial distance from the origin represents the model's standard deviation, indicating the spread or variability of the predictions (Taylor, 2001). The angle corresponds to the Pearson correlation coefficient between the predicted and observed values, reflecting the degree of similarity between the pattern and the observed values (Ghorbani et al., 2025).

In this study, the Taylor Diagram facilitated direct visual comparison of the predictive performance of various models, including ANN (MLP), SVM, and RF, against the measured OWC values. By plotting each model on the diagram, it was possible to quickly identify those that most accurately reproduced the observed soil behavior. The use of the Taylor Diagram enabled a more nuanced evaluation than traditional single-metric approaches, revealing trade-offs between correlation strength and variability representation (Jose et al., 2022). In addition, the Taylor Diagram proved to be a valuable validation tool, enabling an integrated and visually intuitive comparison of model performance and thereby supporting the selection of the most suitable ML model for geotechnical prediction tasks.

2.2.6. Partial Dependence Plots (PDP)

PDP, introduced by Friedman (2001), was employed in this study as a model-agnostic interpretability technique to analyze the influence of individual input variables on the predictions made by complex ML models (Friedman, 2001). PDP is beneficial for interpreting "black-box" models such as RF, SVM, and ANN (MLP), allowing for visualization of the marginal effect of selected features on the predicted outcome without altering the original model structure.

The main principle of PDP is to isolate and visualize the effect of selected input features on the predicted response by averaging out the influence of all other features (Johnson et al., 2022). Specifically, for a prediction model $\hat{f}(X)$ trained on input features (X_S, X_C), where X_S is a subset of features of interest and X_C represents the complementary set of all other features, the partial dependence function is defined as the expected prediction over the marginal distribution of X_C (Friedman, 2001):

$$\hat{f}_{X_S}(x_S) = \mathrm{E}_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, x_C) dP(x_C) \quad (1)$$

Where X_S is a fixed value or range of the features of interest, and $P(x_C)$ is the marginal probability distribution of X_C.

In practice, the partial dependence is estimated using the available data by averaging predictions across all observations while fixing x_C to specific values:

$$\hat{f}_{X_S}(x_S) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(x_S, x_{C_i}) \quad (2)$$

Where N is the number of observations and x_{C_i} are the observed values of the complementary features.

3. Results and discussion

3.1. Model validation and comparison

Validation of the models (RF, SVM, and ANN [MLP]) was implemented on both training and testing datasets using R^2, RMSE, and MAE metrics. Figure 3 presents scatter plots of predicted versus observed OWC values for the SVM, RF, and ANN (MLP) models, along with R^2 values. The SVM model

achieved moderate predictive accuracy, with an R^2 of 0.64 during training and 0.63 during testing, indicating consistent performance without overfitting. The RF model substantially outperformed the others, achieving an R^2 of 0.96 during training and 0.84 during testing. This shows excellent model fit and strong generalization. In contrast, the ANN (MLP) model showed the weakest performance. The R^2 was 0.60 for training and dropped to 0.44 in testing, revealing limited generalization and possible overfitting.

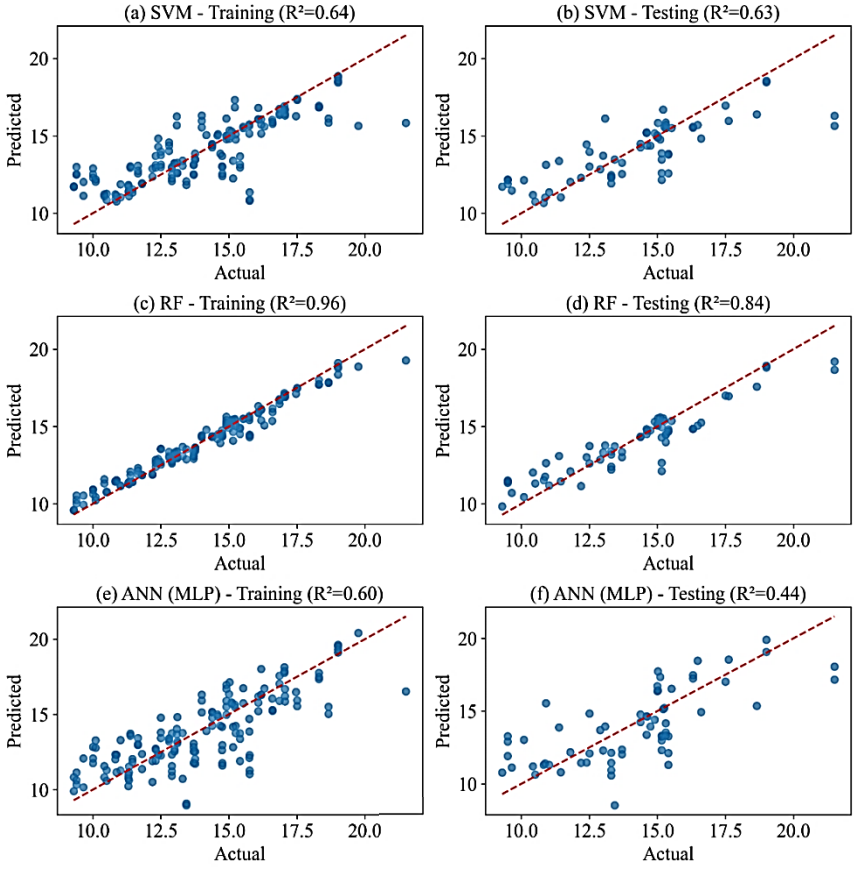

Figure 3. R^2 values of the models: (a) SVM training, (b) SVM testing, (c) RF training, (d) RF testing, (e) ANN (MLP) training, and (f) ANN (MLP) testing

Figure 4 provides a comparison of the predicted versus observed OWC values. During the training phase, all models follow the general trend of the data, but the RF predictions most closely align with actual values across both high and low OWC ranges. SVM maintains acceptable trend tracking, though with increased deviation at local extremes. The ANN (MLP) model shows more erratic behavior, with significant fluctuations and poorer tracking, especially during testing.

Figure 5 compares the models using three key metrics: R^2, RMSE, and MAE across training and testing datasets. RF consistently achieves the highest R^2 values (0.96 training, 0.84 testing), the lowest RMSE (0.53 training, 1.07 testing), and the lowest MAE (0.39 training, 0.78 testing). SVM shows intermediate performance, while ANN (MLP) yields the highest error values in both RMSE and MAE.

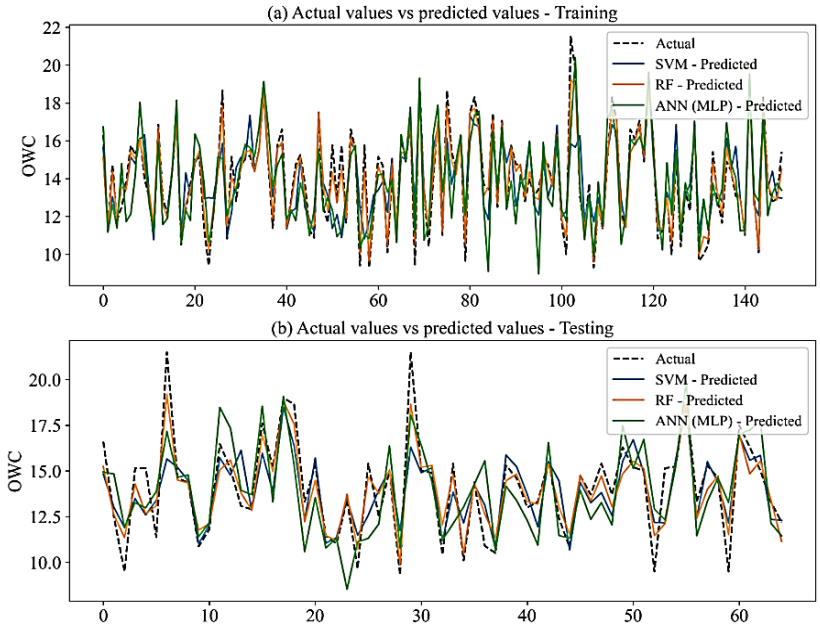

Figure 4. Actual vs predicted values of the models: (a) training and (b) testing

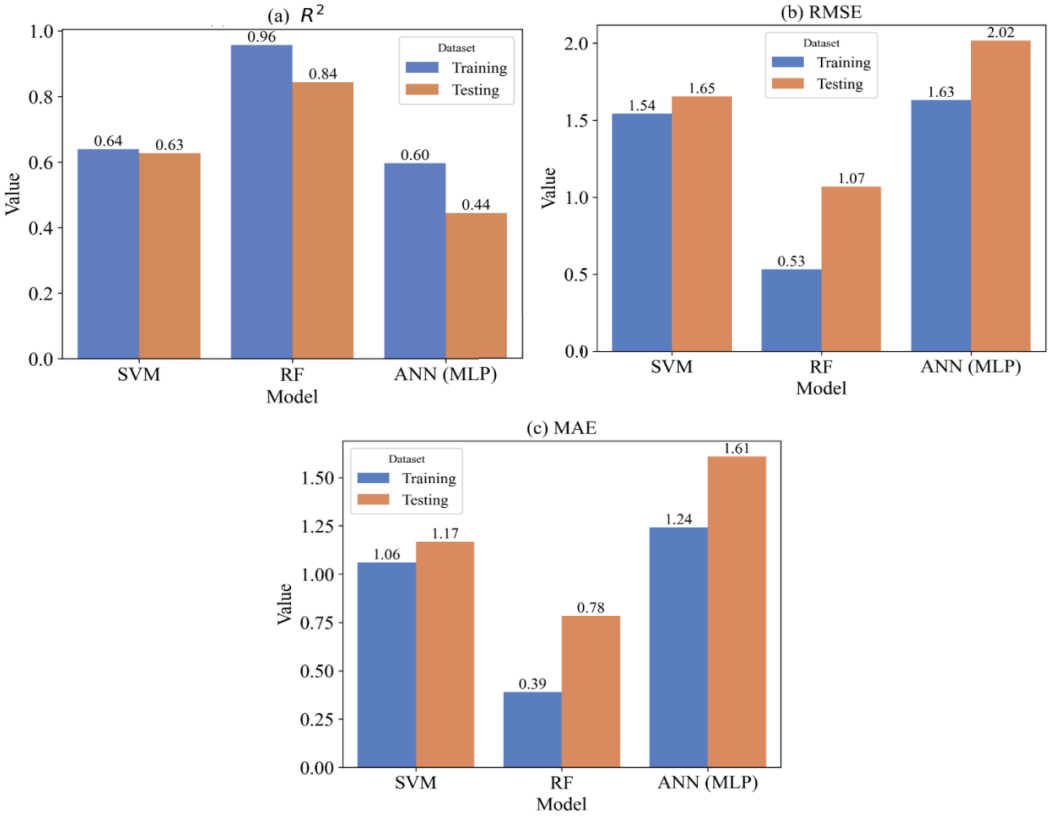

Figure 5. Comparison of the validation metrics of the models: (a) R^2, (b) RMSE, and (c) MAE

Figure 6 displays a Taylor Diagram-based comparative analysis. The RF model appears closest to the reference point in all panels, indicating the highest correlation and lowest error. SVM is moderately placed, while M ANN (MLP) LP lags behind. The visual clustering around the reference point supports the numerical validation metrics, consolidating the conclusion that RF provides the most accurate and stable performance among the evaluated models.

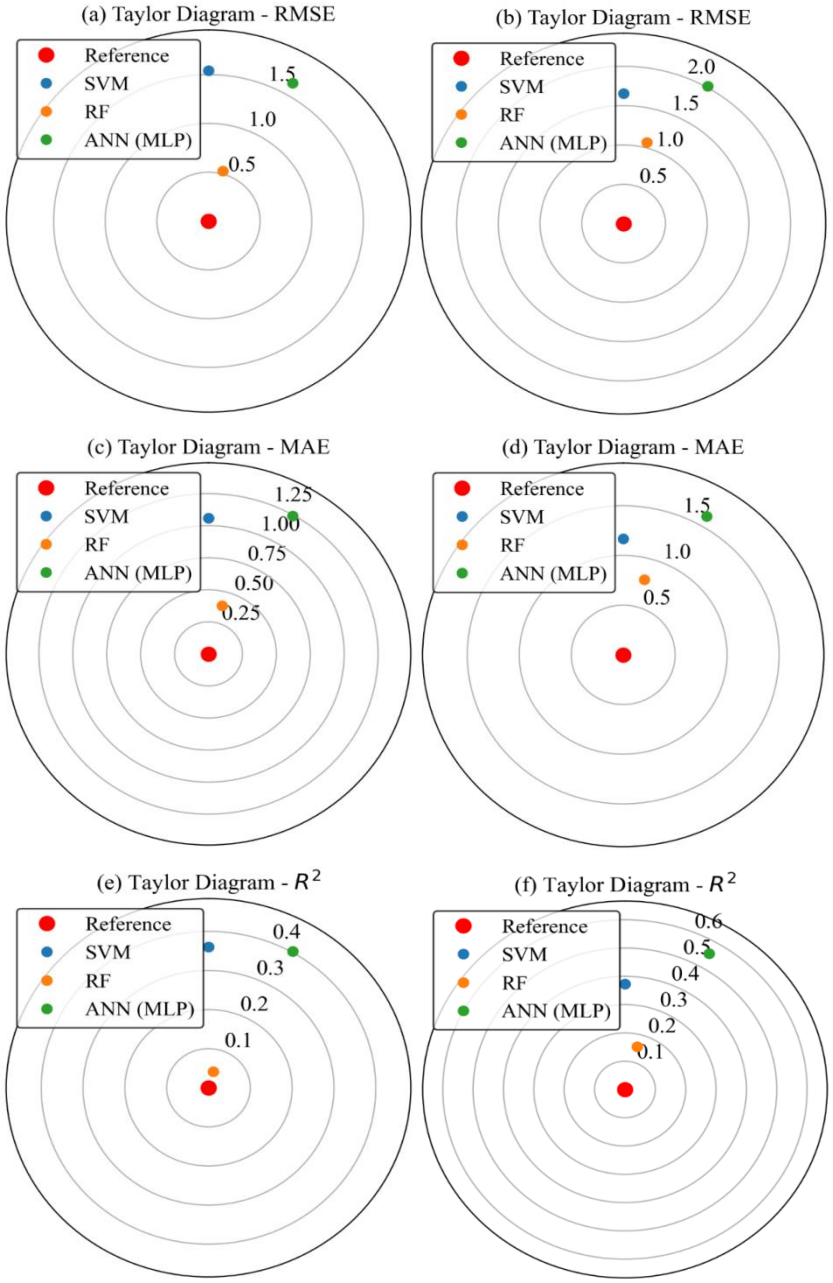

Figure 6. Taylor diagram of the models: (a) RMSE training, (b) RMSE testing, (c) MAE training, (d) MAE testing, (e) R^2 training, and (f) R^2 testing

In summary, the findings of this study highlight the superior predictive performance of the RF model in estimating the OWC of soils. Its ability to generalize well to unseen data highlights its suitability for practical geotechnical applications. From a computational perspective, RF's ensemble structure, which leverages bootstrap aggregation (bagging) and randomized feature selection, enhances model stability and reduces variance, allowing it to handle nonlinearities and variable interactions effectively (Ahmad et al., 2017; Breiman, 2001). This is particularly advantageous for modeling complex geotechnical phenomena where explicit functional relationships are difficult to define. In contrast, SVM model requires careful kernel selection and parameter tuning and may be less adaptable to heterogenous datasets (Han et al., 2018). ANN (MLP) model, though theoretically powerful, exhibited challenges in this study, including overfitting and poor convergence, possibly due to limited dataset size and the high dimensionality of input features (Ahmad et al., 2017).

Table 2 shows the comparison of the validation metrics of the RF model used in this study with previous and published works using different combination of input variables. It can be observed that the models used in this study is comparative with the models used in the previous and published works.

Table 2. Comparison of validation metrics: current study vs. previous works

Works	Models	R^2 (Testing)	RMSE (Testing)	MAE (Testing)
Taffese and Abegaz (2022)	ANN	0.55	-	-
	OME	0.56	-	-
Liu et al. (2023)	RF	0.82	4.96	-
	SVM	0.72	5.33	-
This study	RF	0.84	1.07	0.78
	ANN	0.44	2.02	1,61
	SVM	0.63	1.65	1.17

3.2. PDP analysis

Figure 7 illustrates PDP for the best RF model, highlighting how key input features individually influence the predicted OWC values. PDP for G shows a decreasing trend, consistent with the physical principle that denser soils retain less moisture. CS content displays a slight peak followed by a decline in the OWC, reflecting reduced water retention at higher sand contents. FS shows a mild increase in the OWC with higher values, as finer particles tend to hold more water. SC content has a strong positive influence on the OWC, with higher values substantially increasing predicted moisture content, aligning with the known behavior of fine-grained soils. O also shows a marked increase in the OWC, consistent with the water-retention capacity of organic matter. LL and PL both show upward trends, indicating that plastic soils require more water for compaction. PI displays a stepped increase in the OWC, further emphasizing the role of soil plasticity in determining moisture needs.

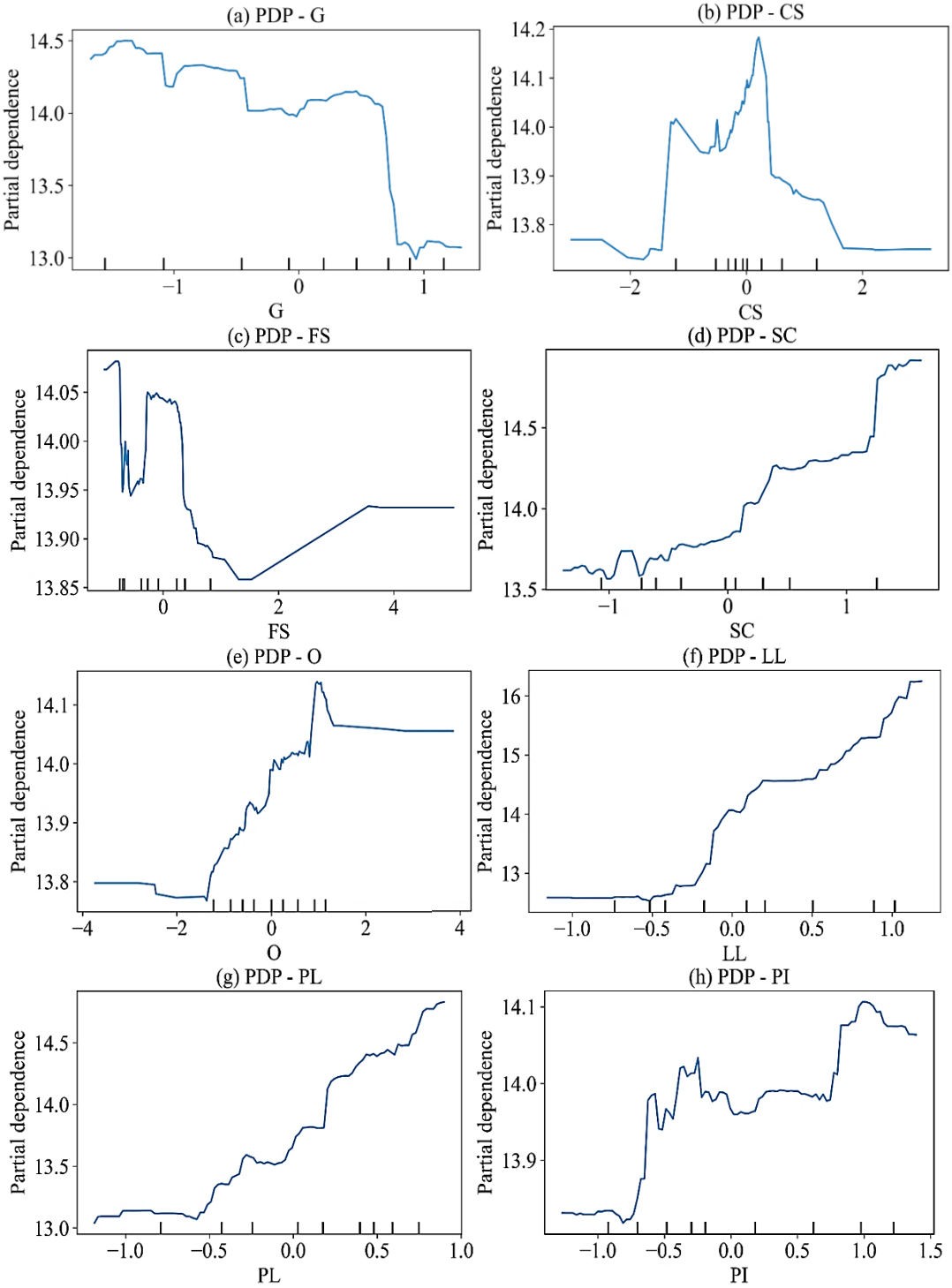

Figure 7. PDP analysis of the variables using RF model: (a) G, (b) CS, (c) FS, (d) SC, (e) O, (f) LL, (g) PL, and (h) PI

4. Conclusions

In this study, a comparative evaluation of advanced ML models, including RF, SVM, and ANN (MLP) for predicting the OWC of soils using data of 214 soil samples collected from the Van Don - Mong Cai expressway construction project (Vietnam) and commonly measured geotechnical properties such as G, CS, FS, SC, O, LL, and PL. Results showed that among the three, the RF model consistently demonstrated the highest predictive accuracy and generalization capability, outperforming both SVM and ANN (MLP). PDP analysis highlighted fines content, plasticity indices, and organic matter as the most influential predictors, in line with established geotechnical knowledge.

The findings of this study underscore the potential of RF as an accurate and efficient alternative to traditional laboratory-based OWC determination methods. By significantly reducing the reliance on extensive testing, RF can accelerate decision-making in geotechnical engineering while maintaining reliability. Nevertheless, future work should focus on expanding the dataset to include diverse soil types, field-scale data, and varying environmental conditions to enhance model robustness further. Additionally, integrating explainable ML methods and hybrid modeling approaches may strengthen interpretability and broaden the scope of ML applications in geotechnical practice.

References

Ahmad M.W., Mourshed M., Rezgui Y., 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77–89.

Aragón A., Garcıa M., Filgueira R., Pachepsky Y.A., 2000. Maximum compactibility of Argentine soils from the Proctor test;: The relationship with organic carbon and water content. Soil and Tillage Research, 56, 197–204.

Benbouras M.A., Lefilef L., 2023. Progressive machine learning approaches for predicting the soil compaction parameters. Transportation Infrastructure Geotechnology, 10, 211–238.

Blotz L.R., Benson C.H., Boutwell G.P., 1998. Estimating optimum water content and maximum dry unit weight for compacted clays. Journal of Geotechnical and Geoenvironmental Engineering, 124, 907–912.

Breiman L., 2001. Random forests. Machine learning, 45, 5–32.

Duc N.D., Nguyen M.D., Prakash I., Van H.N., Van Le H., Thai P.B., 2025. Prediction of safety factor for slope stability using machine learning models. Vietnam Journal of Earth Sciences, 47(2), 182–200. https://doi.org/10.15625/2615-9783/22196.

Friedman J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.

Khatti J., Grover K.S., 2023. Prediction of compaction parameters of compacted soil using LSSVM, LSTM, LSBoostRF, and ANN. Innovative Infrastructure Solutions, 8, 76.

Kim J.-H., 2024. A study on the water content in distribution pole transformer using random forest model. Computers and Electrical Engineering, 120, 109823.

Lai X., Zhu Q., Zhou Z., Liao K., 2017. Influences of sampling size and pattern on the uncertainty of correlation estimation between soil water content and its influencing factors. Journal of Hydrology, 555, 41–50.

Li B., You Z., Ni K., Wang Y., 2024. Prediction of Soil Compaction parameters using machine learning models. Applied Sciences, 14, 2716.

Liu G., Tian S., Xu G., Zhang C., Cai M., 2023. Combination of effective color information and machine learning for rapid prediction of soil water content. Journal of Rock Mechanics and Geotechnical Engineering, 15, 2441–2457.

Mueller L., Schindler U., Fausey N.R., Lal R., 2003. Comparison of methods for estimating maximum soil water content for optimum workability. Soil and Tillage Research, 72, 9–20.

Navidi M.N., Seyedmohammadi J., Seyed Jalali S.A., 2022. Predicting soil water content using support vector machines improved by meta-heuristic algorithms and remotely sensed data. Geomechanics and Geoengineering, 17, 712–726.

Ngo T.Q., Nguyen L.Q., Tran V.Q., 2022. Predicting tensile strength of cemented paste backfill with aid of second order polynomial regression. Journal of Science and Transport Technology, 43–51.

Nguyen D.D., Nguyen H.P., Vu D.Q., Prakash I., Pham B.T., 2023. Using GA-ANFIS machine learning model for forecasting the load bearing capacity of driven piles. Journal of Science and Transport Technology, 3, 26–33.

Nguyen H.D., Pham V.T., Nguyen Q.-H., Bui Q.-T., 2025. Soil salinity prediction using satellite-based variables and machine learning: Case study in Tra Vinh province, Mekong Delta, Vietnam. Vietnam Journal of Earth Sciences, 47(2), 201–219. https://doi.org/10.15625/2615-9783/22438.

Nguyen Q.H., Ly H.-B., Ho L.S., Al-Ansari N., Le H.V., Tran V.Q., Prakash I., Pham B.T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 4832864.

Nguyen T.T., Nguyen D.D., Nguyen S.D., Prakash I., Van Tran P., Pham B.T., 2022. Forecasting construction price index using artificial intelligence models: support vector machines and radial basis function neural network. Journal of Science and Transport Technology, 9–19.

Nhat V.H., Trinh P.T., Cam L.V., Dieu B.T., Van Hiep L., Prakash I., Anh N.N., Van Hong N., Thanh N.D., Thao N.P., 2025. Mapping Cadmium Contamination Potential in Surface Soil for Civil Engineering Applications: A Comparative Study of Machine Learning and Deep Learning Models in the Gianh River Basin, Vietnam. Journal of Science and Transport Technology, 48–70.

Pal S., Hieu V.T., Nguyen D.D., Vu D.Q., Prakash I., 2024. Investigation of support vector machines with different kernel functions for the prediction of compressive strength of concrete. Journal of Science and Transport Technology, 55–68.

Pham B.T., Amiri M., Nguyen M.D., Ngo T.Q., Nguyen K.T., Tran H.T., Vu H., Anh B.T.Q., Van Le H., Prakash I., 2021. Estimation of shear strength parameters of soil using Optimized Inference Intelligence System. Vietnam Journal of Earth Sciences, 43(2), 189–198. https://doi.org/10.15625/2615-9783/15926.

Pham B.T., Nguyen M.D., Bui K.-T.T., Prakash I., Chapi K., Bui D.T., 2019. A novel artificial intelligence approach based on Multilayer Perceptron Neural Network and Biogeography-based Optimization for predicting the coefficient of consolidation of soil. Catena, 173, 302–311.

Pham T.A., 2024. Developing a Machine Learning Model for Predicting the Settlement of Bored Piles. Journal of Science and Transport Technology, 95–109.

Phan V.-H., Ly H.-B., 2024. RIME-RF-RIME: A novel machine learning approach with SHAP analysis for predicting macroscopic permeability of porous media. Journal of Science and Transport Technology, 58–71.

Phung B.-N., Le T.-H., Nguyen M.-K., Nguyen T.-A., Ly H.-B., 2023. Practical numerical tool for marshall stability prediction based on machine learning: an application for asphalt concrete containing basalt fiber. Journal of Science and Transport Technology, 26–43.

Prakash I., Kumar R., Nguyen T.-A., Vu P.T., 2022. Development of effective XGB model to predict the Axial Load Capacity of circular CFST columns. Journal of Science and Transport Technology, 26–42.

Prakash I., Nguyen D.D., Tuan N.T., Van Phong T., Van Hiep L., 2024. Landslide susceptibility zoning: integrating multiple Intelligent models with SHAP Analysis. Journal of Science and Transport Technology, 23–41.

Taffese W.Z., Abegaz K.A., 2022. Prediction of compaction and strength properties of amended soil using machine learning. Buildings, 12, 613.

Taylor K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of geophysical research: atmospheres 106, 7183–7192.

Vapnik V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York, 1–188.

Vapnik V., 2013. The nature of statistical learning theory. Springer science & business media.

Wu Y.-C., Feng J.-W., 2018. Development and application of artificial neural network. Wireless Personal Communications, 102, 1645–1656.

Yin D., Wang Y., Huang Y., 2023. Predicting soil moisture content of tea plantation using support vector machine optimized by arithmetic optimization algorithm. Journal of Algorithms & Computational Technology 17, 17483026221151198.

Zhou J., Zhang Y., Li C., Yong W., Qiu Y., Du K., Wang S., 2023. Enhancing the performance of tunnel water inflow prediction using Random Forest optimized by Grey Wolf Optimizer. Earth Science Informatics, 16, 2405–2420.