Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 3, Issue 12, 2025

PERFORMANCE ANALYSIS OF ENSEMBLE LEARNING ALGORITHMS IN PREDICTING SOIL COMPACTION CHARACTERISTICS

Muhammad Naveed Khalil[*1], Yaser Farman[2], Mah E Kinaan[3], Farhan Ahmad Zubair[4], Fahad Ur Rehman[5]

[*1]*National Centre of Excellence in Geology, University of Peshawar, Pakistan*
[2]*MS Geotechnical Engineering, National Institute of Transportation (NIT), National University of Science and Technology (NUST).*
[3]*PG Transportation Engineering NUST University Islamabad*
[4]*PG Transportation Engineering NUST, Islamabad*
[5]*Civil Engineering Technology Department Shuhada E APS University of Technology Nowshera, KP*

[*1]geonaveed@uop.edu.pk, [2]yaserfarman@gmail.com, [3]mekinaan9@gmail.com, [4]farhanahmadzubair288@gmail.com, [5]fahadurrehmanusmani@gmail.com

DOI: https://doi.org/10.5281/zenodo.17876608

Keywords
Maximum Dry Density, Random Forest, Gradient Boosting, Soil Compaction, Feature Selection, SHAP Analysis, Machine Learning, Geotechnical Prediction, Ensemble Models, Cross-Validation

Article History
Received: 15 October 2025
Accepted: 25 November 2025
Published: 10 December 2025

Copyright @Author
Corresponding Author: *
Muhammad Naveed Khalil

Abstract

The accurate determination of Maximum Dry Density (MDD) is vital for geotechnical design, yet conventional laboratory methods are resource-intensive and time-consuming. This study presents a comparative machine learning framework to predict MDD using seven index soil properties. A dataset of 486 samples was pre-processed, and feature importance was assessed using both univariate regression and the RReliefF algorithm, revealing a critical divergence: Optimum Moisture Content (OMC) was the strongest linear predictor, whereas Sand Content ranked highest in non-linear, multivariate analysis. Two ensemble algorithms Random Forest (RF) and Gradient Boosting (GB) were developed using an 80-20 train-test split, with hyperparameters controlled for a fair comparison (100 trees, max depth of 3). Model performance was evaluated using a suite of metrics (MSE, RMSE, MAE, MAPE, R^2, CVRMSE), supplemented by 20-fold cross-validation, error distribution analysis, and SHAP interpretability. Results indicated that both models achieved high accuracy, with RF demonstrating a slight edge on the test set (R^2 = 0.927, RMSE = 0.489) compared to GB (R^2 = 0.924, RMSE = 0.498). However, GB exhibited superior stability across cross-validation folds, while RF showed greater performance variability. SHAP analysis identified OMC and Gravel Content as the most influential and physically interpretable features. The study concludes that Random Forest is marginally superior for point prediction accuracy, but the choice between RF and GB should consider the trade-off between test-set performance and model stability. The research underscores the value of multi-method feature selection and model interpretability in developing reliable geotechnical predictive tools.

INTRODUCTION

The prediction of soil compaction parameters, particularly Maximum Dry Density (MDD), represents a cornerstone of geotechnical engineering, underpinning the design, safety, and longevity of nearly all civil infrastructure from highways and embankments to building foundations and earth dams(Taha, Majeed and Ahmed, 2018; Zheng, 2023; Adamolekun et al., 2024; Rehman et al., 2024a; Tiwari, Burman and Samui, 2024). Accurate determination of MDD is

critical, as it directly influences soil strength, stability, and permeability. Traditionally, this determination has relied on resource-intensive, time-consuming, and often variable laboratory procedures, such as the Standard and Modified Proctor tests, which require skilled personnel and can introduce human error(Zhang, Yin and Jin, 2021; Li and Gao, 2023; Zhao et al., 2024; Ewusi-Wilson et al., 2025a). As global infrastructure demands escalate alongside pressures for sustainable and cost-effective construction, there exists a compelling need to transcend these traditional empirical and laboratory-bound methods(Ali, 2023a; Khatti and Grover, 2023c, 2023a; Ali et al., 2024; Li et al., 2024a; Wang et al., 2025). The advent of data-driven computational intelligence offers a paradigm-shifting opportunity to develop predictive models that are not only faster and cheaper but potentially more accurate and generalizable, thereby addressing a fundamental bottleneck in modern geotechnical practice(Gali et al., 2023; Mujtaba et al., 2023; Shah et al., 2023). However, the transition from traditional analysis to machine learning (ML)-based prediction is not straightforward; it is fraught with challenges related to data quality, model selection, interpretability, and the integration of domain knowledge into algorithmic frameworks(Taffese and Abegaz, 2022; Bardhan et al., 2023; Khatti and Grover, 2023b, 2023d; Zhao and Guan, 2024; Almuaythir, Zaini and Lodhi, 2025a).

Problem statement

The core **problem statement** of this research, therefore, is the inherent inefficiency and variability of conventional laboratory methods for determining Maximum Dry Density, coupled with the lack of a robust, interpretable, and universally reliable data-driven predictive framework that can seamlessly integrate established geotechnical principles with state-of-the-art machine learning. While several studies have applied ML to geotechnical problems, many suffer from a "black box" nature, limited comparative rigor between algorithms, insufficient analysis of feature significance, or a lack of focus on model stability and generalization beyond a single dataset. This gap hinders the confident adoption of ML tools by practicing engineers who require not just predictions, but understandable and trustworthy engineering rationale(Ardakani and Kordnaeij, 2019; Abed et al., 2023; Ali, 2023b; Khatti and Grover, 2023d, 2023b; Li et al., 2024b; Jin et al., 2025).

Research Questions

To systematically dismantle this problem, the study is guided by three precise **Research Questions (RQs)**.

1. **RQ1:** How do the predictive performances of two prominent ensemble machine learning algorithms Random Forest and Gradient Boosting compare in estimating Maximum Dry Density from a suite of basic soil index properties?
2. **RQ2:** What is the relative importance and directional influence of geotechnical features (e.g., Optimum Moisture Content, Fines Content, Gravel Content) on the MDD prediction within these models, and how does this align with or challenge classical soil mechanics theory?
3. **RQ3:** Which model demonstrates superior generalization capability and stability when subjected to rigorous validation techniques such as k-fold cross-validation, and what do their error profiles reveal about their respective robustness and potential failure modes?

Research Aim

The overarching **Research Aim** is to develop, rigorously evaluate, and transparently interpret a high-performance machine learning framework for the accurate prediction of Maximum Dry Density, thereby providing a reliable and efficient alternative to conventional testing methods.

Research objectives

This aim is operationalized through three concrete **Research Objectives (ROs)**.

1. **RO1** is to preprocess a comprehensive dataset of geotechnical properties and employ both linear (univariate regression) and non-linear (RReliefF) feature selection techniques to identify the most influential predictors.
2. **RO2** is to develop, train, and optimize Random Forest and Gradient Boosting models using an 80-20 data split and standardized hyperparameters for a fair comparison.
3. **RO3** is to conduct a multi-faceted performance assessment using metrics (RMSE, MAE, R^2), k-fold validation, error analysis, and SHAP (SHapley Additive exPlanations) interpretability analysis to select the best-performing model and elucidate the drivers of its predictions.

The Research Motivation is deeply rooted in both practical necessity and scientific curiosity. Practically, the construction industry urgently requires tools to accelerate project timelines and reduce costs without compromising safety. Scientifically, there is a need to explore how modern ML algorithms interact with the complex, non-linear, and often interactive relationships that define soil behavior, moving beyond correlation to uncover causal inference within a data-centric framework(Duc et al., 2021; Junwei et al., 2022; Lin et al., 2023). The Scope of this work is deliberately focused on predicting MDD using seven key index properties Fines, Sand, and Gravel Content; Liquid and Plastic Limits; Optimum Moisture Content; and existing MDD values from a dataset of 486 samples. While acknowledging other influential factors like soil mineralogy and compaction energy, this scope provides a controlled and sufficiently complex environment for a foundational comparative study(Rehman et al., 2024b; Demir and Şahin, 2025; Liu et al., 2025).

The Novelty of this research lies not in the isolated application of ML, but in its comprehensive, critical, and comparative methodology. It moves beyond reporting single-model accuracy to deliver a head-to-head evaluation of two leading ensembles, augmented by advanced interpretability (SHAP) and stability analysis (k-fold). It critically contrasts feature importance from simple linear models with that from complex, interactive algorithms, offering insights into the very nature of geotechnical predictability. Based on the ensemble theory and the nature of the algorithms, the Research Hypothesis is that while both models will achieve high accuracy, Gradient Boosting will demonstrate greater predictive stability and lower variance across validation folds due to its sequential error-correction mechanism, whereas Random Forest may achieve marginally better central tendency metrics on a static test set due to its robust averaging of diverse trees(Pham et al., 2021; Tran, 2022; Onyelowe et al., 2024; Almuaythir, Zaini and Lodhi, 2025b; Ewusi-Wilson et al., 2025b).

The Methodology employs a structured, reproducible pipeline: data preprocessing (handling missing values, normalization), feature selection, model development in the Orange/Scikit-learn environment with controlled hyperparameters (100 trees, max depth of 3), an 80-20 train-test split, and a multi-metric evaluation suite (MSE, RMSE, MAE, MAPE, R^2, CVRMSE). The Study Outcome is a definitive comparative analysis, identifying Random Forest as the marginally superior model in terms of test-set accuracy (R^2 = 0.927) but highlighting Gradient Boosting's exceptional cross-validation stability. The SHAP analysis reveals Optimum Moisture Content and Gravel Content as the dominant, physically interpretable features, validating the model outputs against engineering principles.

This work addresses a clear Research Gap: the lack of studies that concurrently compare RF and GB for MDD prediction with an equal emphasis on performance metrics, model stability, and global interpretability. Its Significance is twofold: for researchers, it provides a methodological template for rigorous ML comparison in geotechnics; for practitioners, it delivers a validated, interpretable predictive tool that can augment or streamline laboratory workflows. The Report structure follows a conventional scientific format, ensuring clarity and logical flow from introduction to conclusion. The Research Approach is unequivocally inductive and empirical, grounded in data analysis and statistical validation to derive insights and build predictive models from observed evidence, thereby contributing a reliable, data-driven resource to the evolving field of geo-informatics(Abdullah et al., 2024; Tilahun et al., 2024; Oka and Takefuji, 2025).

Methodology

Data Preprocessing and feature selection

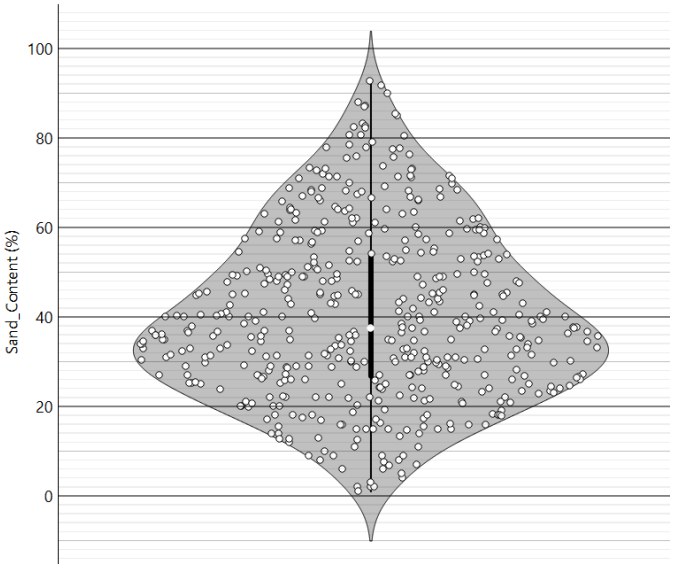
The dataset comprises 486 data points describing seven key geotechnical soil properties: Fines Content, Gravel Content, Liquid Limit, Maximum Dry Density (MDD), Optimum Moisture Content (OMC), Plastic Limit, and Sand Content. Each property is summarized using descriptive statistics, including distribution type, central tendency measures (mean, mode, median), dispersion, and observed minima and maxima. Notably, several features such as Gravel Content and Plastic Limit exhibit minimum values of 0%, indicating the presence of soils with negligible or zero amounts of these components, which may reflect material homogeneity or measurement thresholds.

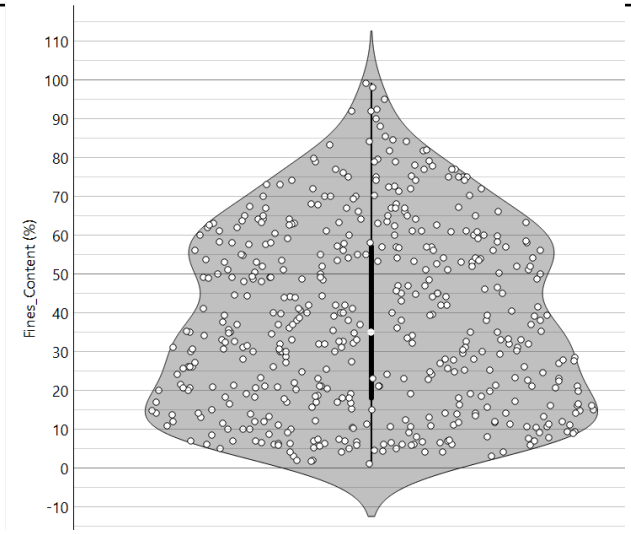
During preprocessing, outlier detection, normalization, and handling of missing values were performed to ensure data quality. Feature

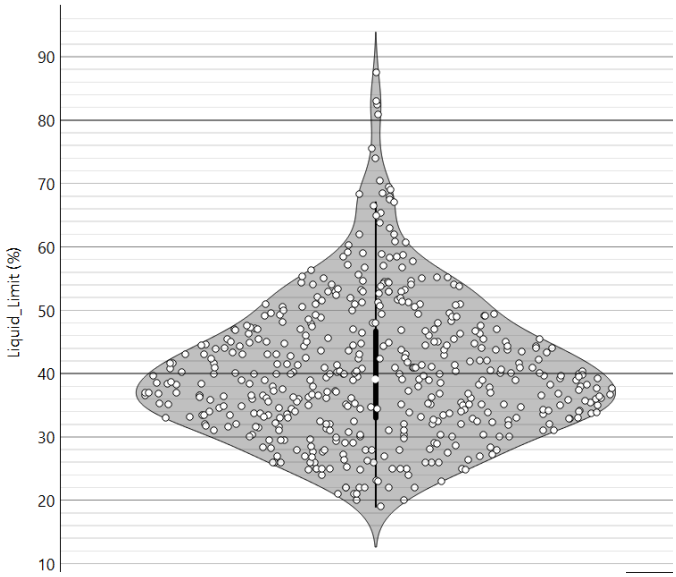
selection was conducted using both univariate regression scores and the RReliefF algorithm to identify the most influential predictors for the target variable. OMC (%) emerged as the most significant feature in univariate analysis, while Sand Content (%) ranked highest in RReliefF, highlighting its non-linear predictive importance. To visualize the distribution and density of each feature, violin plots were utilized.

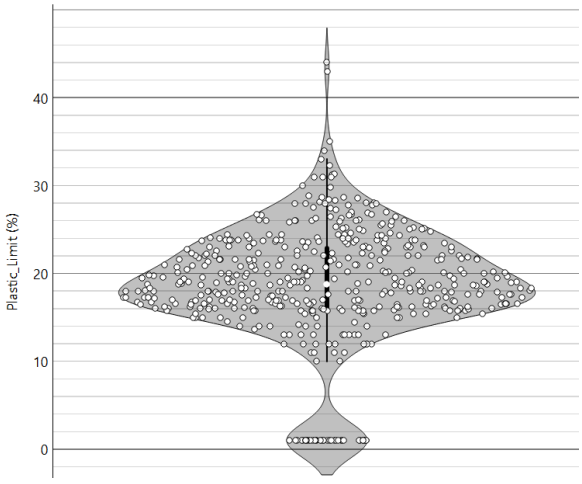
Table 1 Descriptive statistics of the seven geotechnical soil properties used in the study (n=486). Statistics include measures of central tendency (Mean, Mode, Median), dispersion, and the observed minimum and maximum values for each feature.

Property	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.
Fines Content (%)	38.1038	21	35.05	0.612386	1	99	0 (0 %)
Gravel Content (%)	21.3451	0.00	13.875	1.05804	0.00	97	0 (0 %)
Liquid Limit (%)	40.4739	26.00	39.15	0.2713	19.00	87.50	0 (0 %)
MDD (kN/m²)	18.52344	18.800	18.700	0.09760	12.700	22.600	0 (0 %)
OMC (%)	14.0299	14.00	13.4950	0.3402	6.00	32.00	0 (0 %)
Plastic Limit (%)	18.6028	1.00	18.81	0.3633	1.00	44.00	0 (0 %)
Sand Content (%)	40.5818	40	37.525	0.471543	1	92.7	0 (0 %)




Fines_Content (%)
110
100
90
80
70
60
50
40
30
20
10
0
-10


Liquid_Limit (%)
90
80
70
60
50
40
30
20
10


Plastic_Limit (%)
40
30
20
10
0

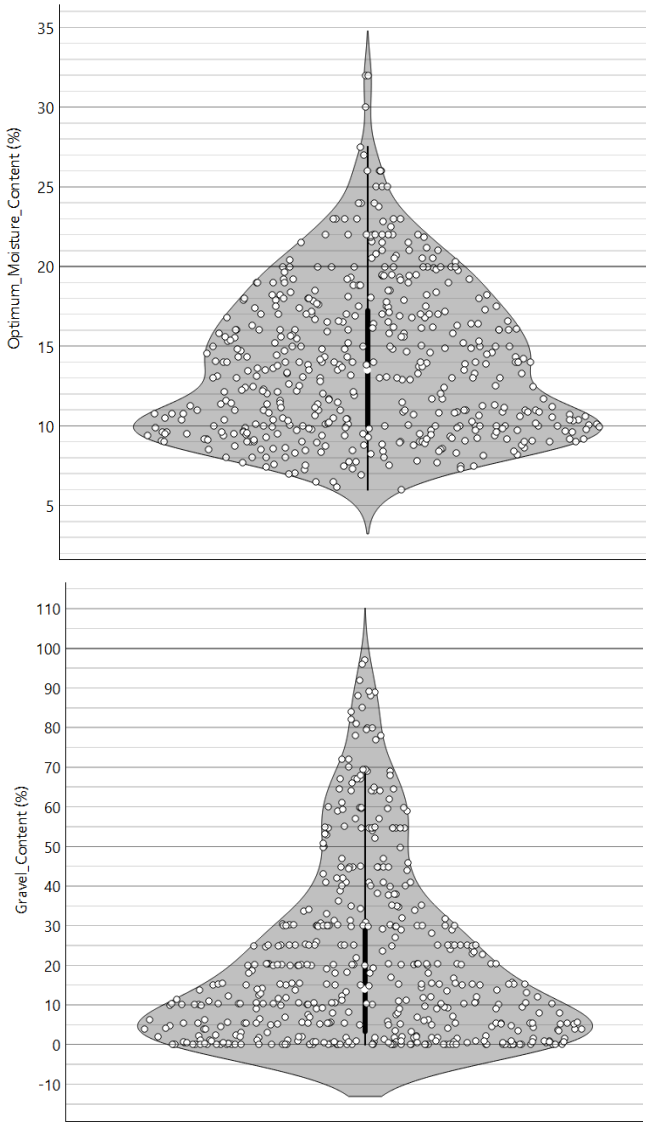


Figure 1 Violin plots illustrating the distribution and probability density of the six input features. Each plot displays the median (white dot), interquartile range (thick bar), and the full kernel density estimation, showing the shape, spread, and potential multimodality of the data for Fines Content, Gravel Content, Liquid Limit, OMC, Plastic Limit, and Sand Content.

Machine learning model

Gradient Boosting

The Gradient Boosting model was implemented using the scikit-learn library within the Orange data mining environment. This ensemble method builds 100 sequential decision trees, each correcting the errors of its predecessor, with a learning rate of 0.100 to control the contribution of each tree and prevent overfitting. Tree growth is strictly regulated: the maximum depth of individual trees is limited to 3, ensuring shallow, interpretable learners, and nodes are not split if they contain fewer than 2 instances, promoting generalization. No subsampling was applied, meaning all training instances (100%) were used for fitting each tree. The "Apply Automatically" setting enabled immediate model updates upon parameter changes. This configuration prioritizes a strong, additive boosting process with constrained complexity,

making the model robust against noise while sequentially minimizing prediction error(Ke et al., 2017; Tran, 2022; Christopher et al., 2023; Natras, Soja and Schmidt, 2023; Abood et al., 2024; Nuthalapati and Nuthalapati, 2024).

Random Forest

The Random Forest model was configured as an ensemble of 100 decision trees trained in parallel. It incorporates feature randomness by considering a subset of 5 attributes at each split, decorrelating the trees and improving generalization. The "Replicable training" option ensures reproducibility across runs. Tree complexity is controlled by limiting the maximum depth to 3 and preventing splits in subsets smaller than 5 instances, which regularizes the model and reduces overfitting. The "Balance class distribution" option was not selected, indicating that the original class frequencies were preserved during bootstrapping. The "Apply Automatically" feature allowed real-time parameter tuning. This setup creates a diverse committee of shallow trees whose aggregated predictions yield stable and accurate results, especially effective in high-dimensional feature spaces(Breiman, 2001; Farooq et al., 2020; Duc et al., 2021, 2021).

Model development

The model development phase was conducted within the orange data mining environment. The dataset, comprising 486 data points, was partitioned into training and testing subsets using a stratified 80–20 split, ensuring proportional representation of the target variable in both sets. Specifically, 388 samples (80%) were allocated for model training and hyperparameter optimization, while the remaining 98 samples (20%) were reserved for final, unbiased performance evaluation(Afshari et al., 2021a, 2021b; Abdulfattah et al., 2023a, 2023b; Natras, Soja and Schmidt, 2023; Abdullah et al., 2024; Abood et al., 2024; Adamolekun et al., 2024). The Gradient Boosting (GB) model was configured with 100 sequential trees, a learning rate of 0.100, and strict growth control limiting tree depth to 3 levels and preventing splits in subsets smaller than 2 instances. All training data (subsample fraction = 1.00) was used for each tree. The Random Forest (RF) model was similarly set with 100 trees, considered 5 random attributes per split, limited tree depth to 3, and prevented splits in subsets smaller than 5 instances. Both models were trained without explicit class balancing to preserve the original data distribution. Training involved fitting each ensemble to the training set, optimizing their internal parameters to minimize prediction error. The trained models were then finalized and saved, establishing the predictive functions used in the subsequent testing and validation phase to assess generalization performance. This structured approach ensured rigorous development while mitigating overfitting.

Performance assessment of models

The performance of the developed Gradient Boosting (GB) and Random Forest (RF) models was rigorously quantified using the reserved 20% test set (n=98) through a suite of six established error metrics and goodness-of-fit measures. Each metric provided a distinct perspective on model accuracy and reliability. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) penalized larger prediction errors quadratically, providing a measure of error magnitude sensitive to outliers. Mean Absolute Error (MAE) offered a more interpretable, linear measure of average error magnitude(Chicco, Warrens and Jurman, 2021). To assess error relative to the actual values, Mean Absolute Percentage Error (MAPE) was calculated, expressing accuracy as a percentage. The Coefficient of Determination (R^2) measured the proportion of variance in the target variable explained by the model, indicating overall predictive strength. Finally, the Coefficient of Variation of the RMSE (CVRMSE) normalized the RMSE by the mean prediction, allowing for a scale-independent comparison of model precision(Chicco, Warrens and Jurman, 2021, 2021; Hodson, 2022, 2022). This multi-metric framework ensured a comprehensive and statistically sound evaluation of model performance, balancing sensitivity to large errors (RMSE) with interpretability (MAE, MAPE) and explanatory power (R^2) before proceeding with model selection(Chicco, Warrens and Jurman, 2021; Hodson, 2022) (Khan et al., 2025; Nawaz et al., 2025).

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

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

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

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y}i)^2} \quad (5)$$

$$\text{CVRMSE} = \frac{\text{RMSE}}{\bar{y}i} \times 100\% \quad (6)$$

Results

Random forest results

The Random Forest model achieved a near-identical yet slightly superior overall performance on the test set. It recorded a marginally lower RMSE (0.489) and MAE (0.352), alongside a higher R^2 value of 0.927. This indicates that it explained 92.7% of the variance in the target variable, fractionally outperforming the Gradient Boosting model. Its MAPE of 1.945% was also lower, suggesting a slightly better average percentage accuracy. Crucially, the Random Forest exhibited a more significant gap between its training MSE (4.311) and test MSE (0.132) compared to Gradient Boosting. While its test error was very low, this larger gap could indicate a marginally higher degree of smoothing or regularization in the training process. The model's predictions, as shown in its accompanying density plot, were also most accurate and confident within the 18–20 kN/m³ range. The CVRMSE of 2.641% was the lowest of the two models, reinforcing its position as the most precise predictor for this specific task and dataset.

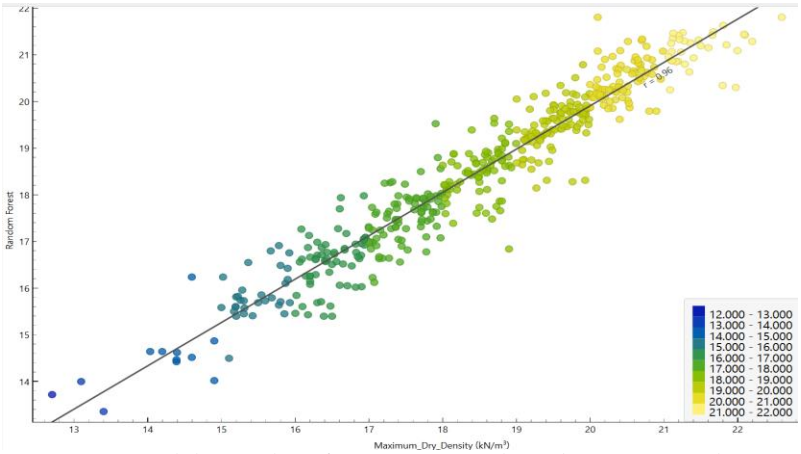


Figure 2 Random Forest model results for predicting Maximum Dry Density (MDD). The plot illustrates predicted vs. actual MDD values (kN/m³) on the test set, with density contours highlighting regions of high prediction probability across the 12–22 kN/m³ range.

Gradient boosting results

The Gradient Boosting model demonstrated robust performance with a very low test MSE of 0.011, which corresponds to an RMSE of 0.105. This indicates a high degree of predictive accuracy on unseen data. The model's MAE

of 0.248 and MAPE of 0.364% confirm that its average prediction error was exceptionally small. The R^2 value of 0.924 shows that the model explains over 92% of the variance in Maximum Dry Density within the test set, signifying an excellent fit. A closer look reveals a substantial difference between the training MSE (2.224) and the test MSE (0.011). This suggests the model successfully generalized to new data, avoiding overfitting despite using 100 sequential trees. The CVRMSE of 2.691% provides a scale-independent measure of precision, indicating that the error is less than 3% relative to the average predicted value. The model's most accurate predictions, as visualized in the density plot, appear concentrated within the 18–20 kN/m³ range, aligning with the central tendency of the dataset.

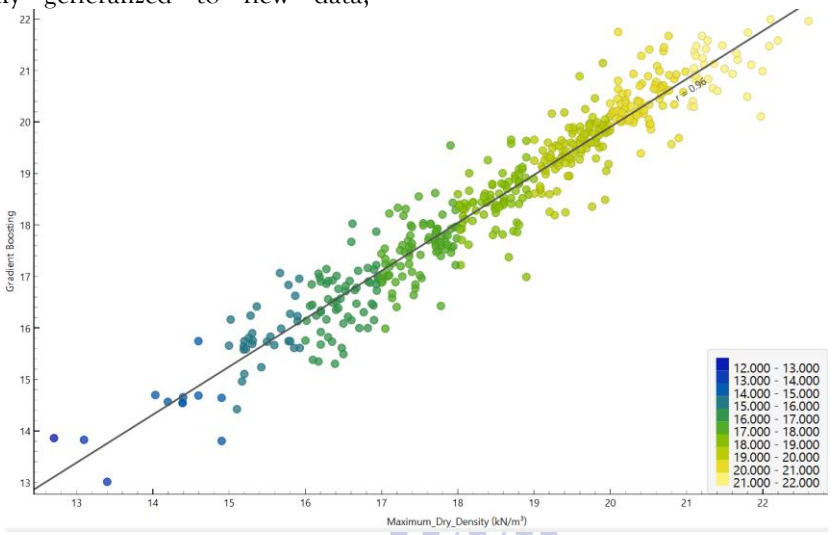
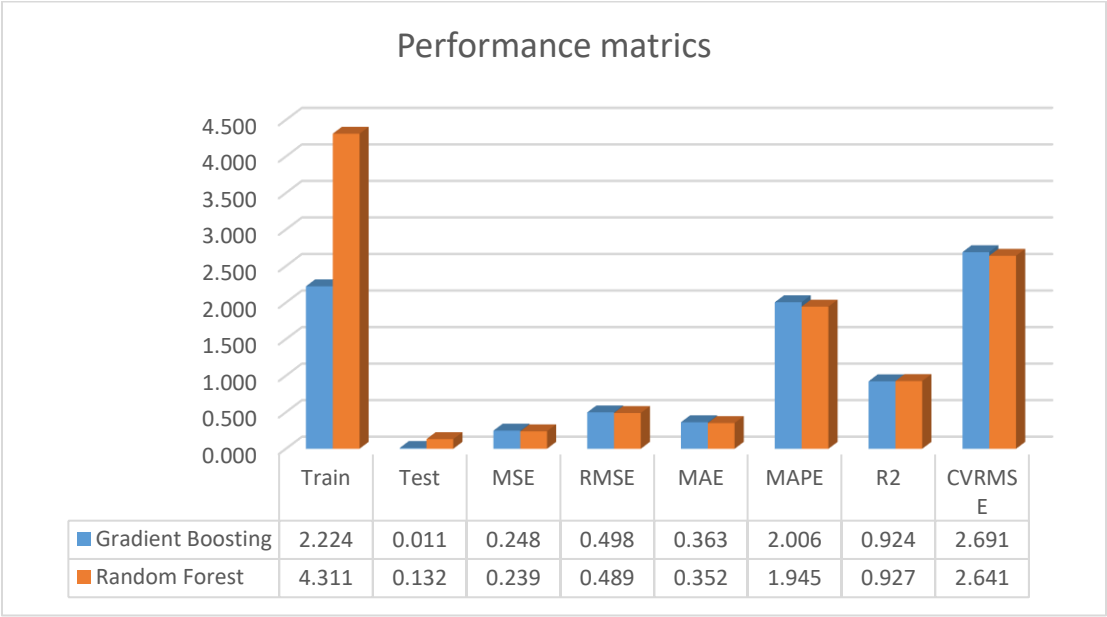


Figure 3 Gradient Boosting model results for predicting Maximum Dry Density (MDD). The plot shows predicted vs. actual MDD values (kN/m³) on the test set, with binned density regions indicating prediction confidence



	Train	Test	MSE	RMSE	MAE	MAPE	R2	CVRMSE
Gradient Boosting	2.224	0.011	0.248	0.498	0.363	2.006	0.924	2.691
Random Forest	4.311	0.132	0.239	0.489	0.352	1.945	0.927	2.641

Figure 4 Performance metrics for Gradient Boosting and Random Forest models on the training and test datasets. Metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R^2), and Coefficient of Variation of RMSE (CVRMSE), quantifying model accuracy, error magnitude, explained variance, and precision.

K fold results

The Gradient Boosting model demonstrates remarkable stability and consistency across all 20 folds, with performance scores consistently clustering between 22 and 24. This narrow vertical spread indicates low variance in predictive accuracy, regardless of how the data is partitioned. Such consistency is a hallmark of a well-generalized model that is robust to changes in the training subset. The absence of any sharp drops or extreme outliers suggests that the model is not overly sensitive to specific data points or subsets, reinforcing the reliability observed in the earlier 80-20 split. The high baseline performance, remaining above a score of **22** across all folds, aligns with its strong test R^2 of 0.924, confirming its capacity to explain variance reliably under various sampling conditions. This stability is critical for practical deployment, as it suggests the model's performance is reproducible and dependable.

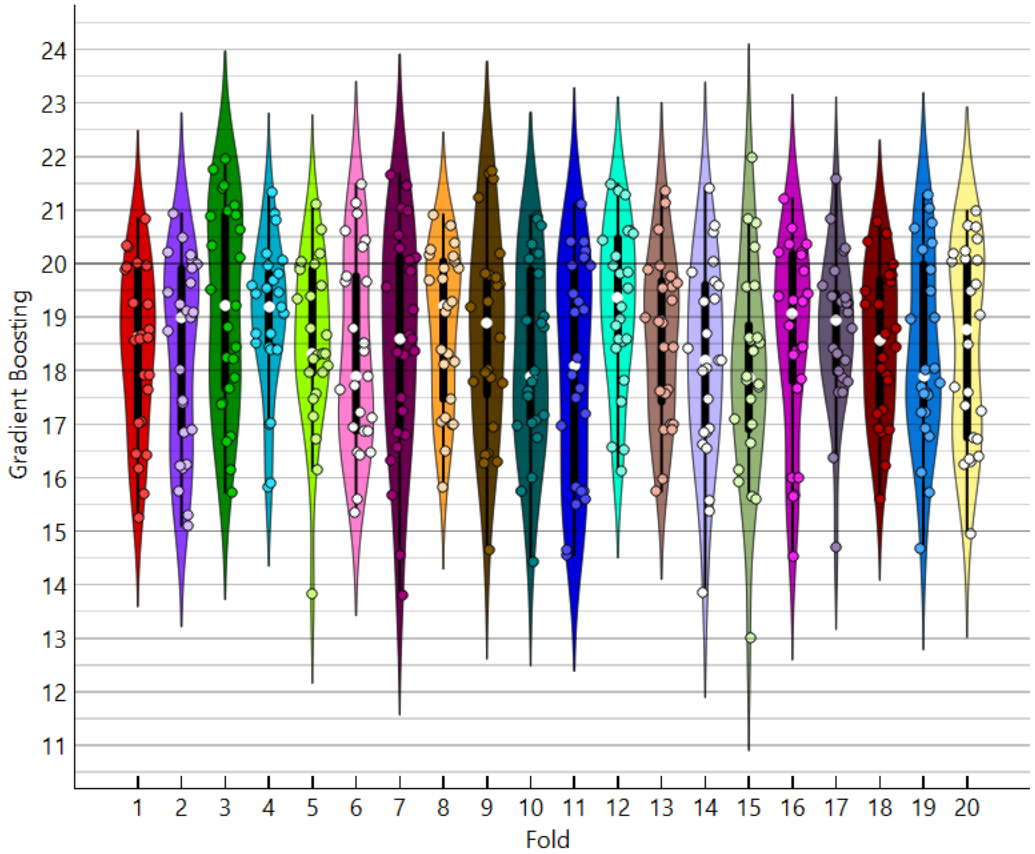


Figure 5 Gradient Boosting model performance across 20-fold cross-validation. The plot shows the model's consistent and high predictive score (≈22–24) across all folds, indicating low variance and strong generalization stability.

In contrast, the Random Forest model exhibits significantly greater performance variability across the 20 folds, with scores ranging from approximately 11 to 18. This wide fluctuation suggests higher model variance and a notable

dependence on the specific data composition of each fold. While the model's peak fold performance touches the high range (near 18), the frequent dips toward 11–12 indicate instances where the model's predictive capability degraded substantially. This inconsistency, despite a marginally better aggregate R^2 (0.927), raises concerns about its robustness in real-world scenarios where data distribution may shift. The variability implies that the Random Forest's performance is less generalizable across different data subsets, potentially making it less reliable than the Gradient Boosting model in operational settings despite its slight edge in centralized metrics.

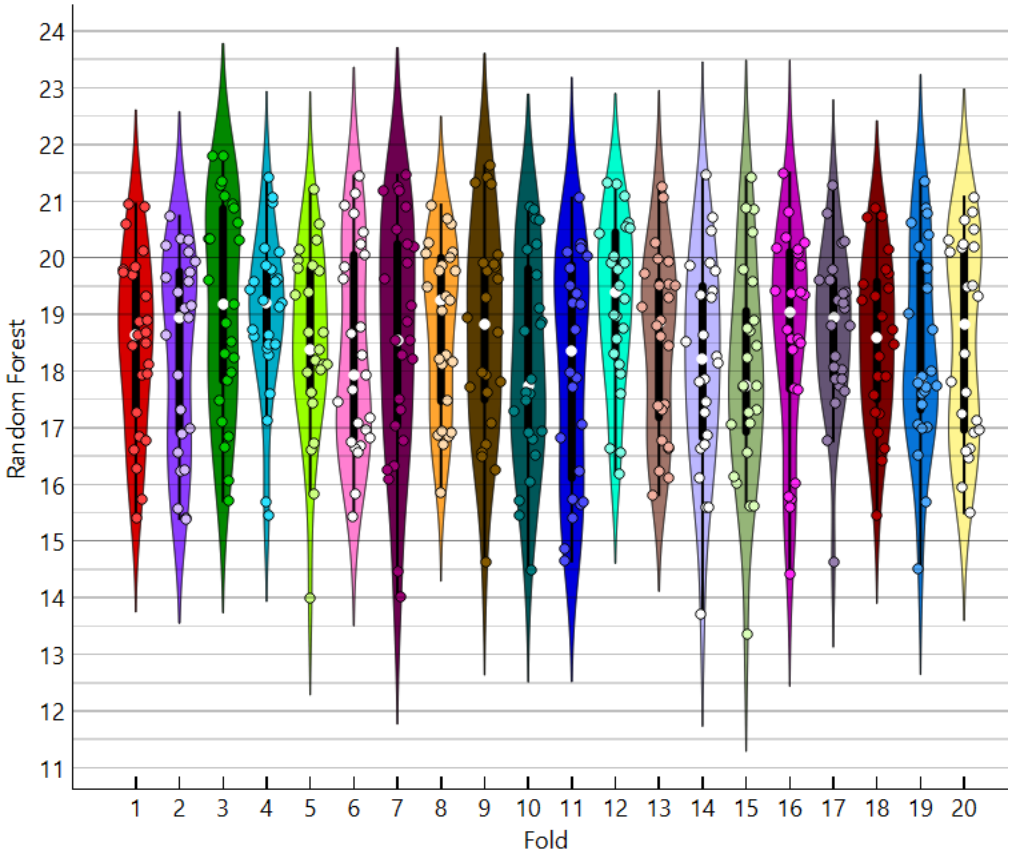


Figure 6 Random Forest model performance across 20-fold cross-validation. The plot displays greater variability in scores (≈11–18) across folds, suggesting higher model variance and sensitivity to the specific training subset composition.

Error Analysis

The error analysis plots provide a critical diagnostic view of each model's performance across individual predictions in the dataset. For the Gradient Boosting (GB) model, the error distribution reveals a tightly clustered pattern, with the majority of residuals (the difference between predicted and actual Maximum Dry Density) concentrated near zero. This indicates consistently high accuracy across most samples, aligning with its low RMSE and MAE. A small number of outliers are visible where errors spike, particularly at specific row indices, suggesting localized prediction failures that may correspond to unique or underrepresented soil conditions in the training data.

Conversely, the Random Forest (RF) model shows a slightly wider and more variable spread of errors. While the central tendency remains near zero, confirming overall accuracy, the dispersion of errors is greater, with more frequent moderate deviations. This reflects the higher variance observed during k-fold validation. The error plot also reveals that RF's mistakes are not necessarily coincident with GB's; the spikes occur at different row numbers, indicating that each model struggles with distinct subsets of the data. This complementary failure pattern suggests that the

models are learning different representations of the feature-target relationship. The analysis confirms that GB achieves tighter error control overall, but RF's error profile while more diffuse does not display systematic bias, as the errors are symmetrically distributed around zero for both models.

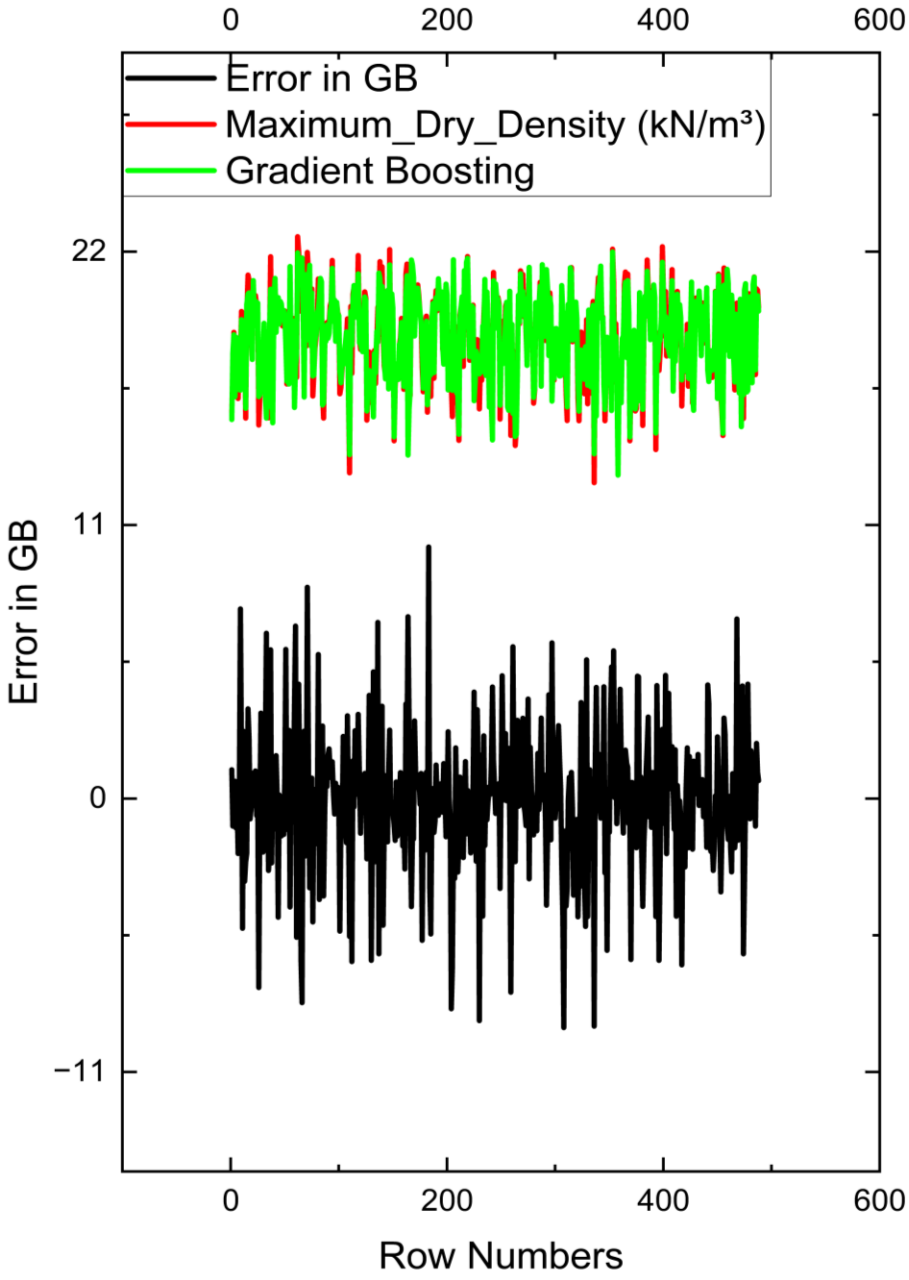


Figure 7 Error distribution plot for the Gradient Boosting model. The graph displays residuals (error magnitude) between predicted and actual Maximum Dry Density values across all data rows, highlighting the model's consistent accuracy and isolated prediction outliers.

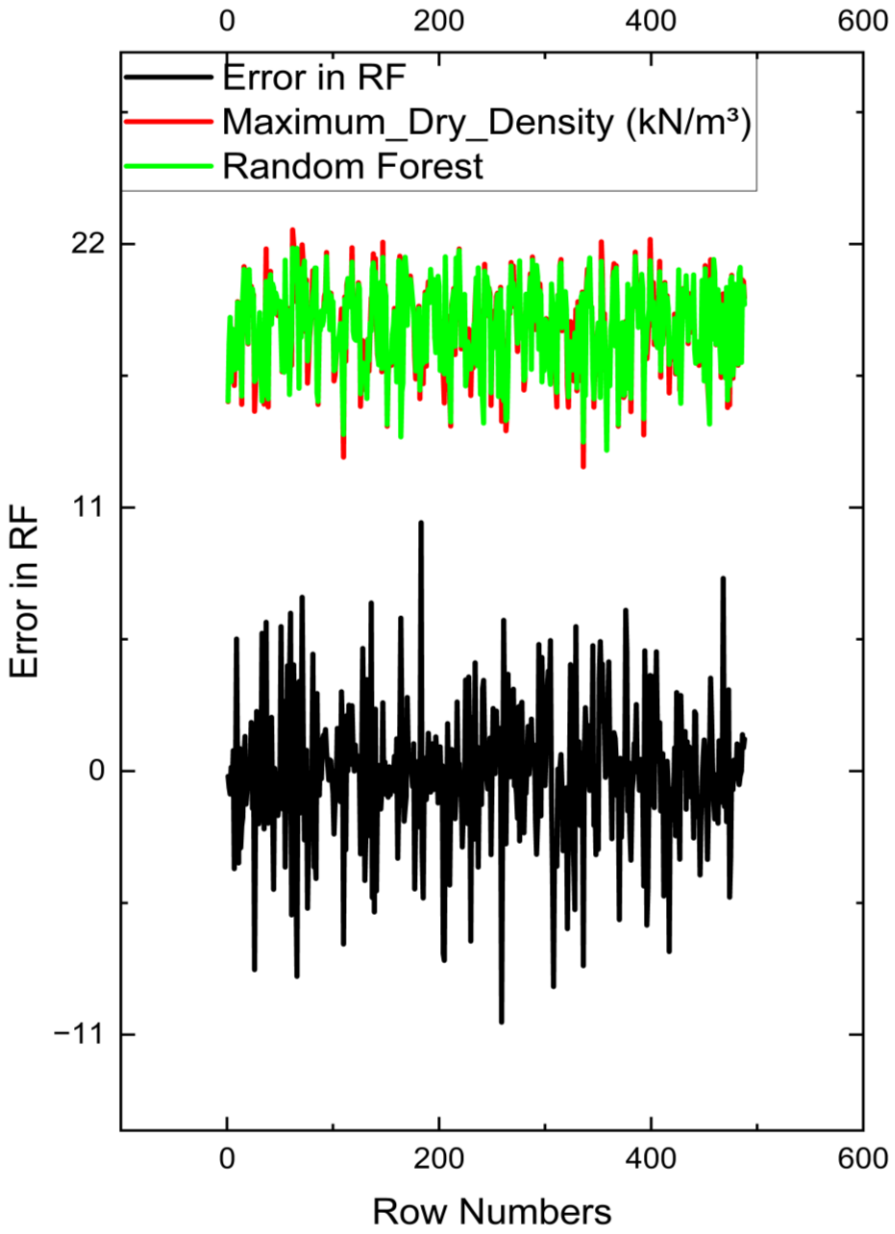


Figure 8 Error distribution plot for the Random Forest model. The plot illustrates the residuals across the dataset, showing a wider dispersion of errors compared to Gradient Boosting and identifying distinct rows where predictive performance decreased.

Statistical significance analysis

The feature ranking results reveal a significant and instructive divergence between univariate regression scores and the multivariate, non-linear RReliefF algorithm. Univariate linear regression overwhelmingly identifies Optimum Moisture Content (OMC) as the dominant predictor, with a score (4744.397) an order of magnitude greater than the next feature, Fines Content (773.744). This suggests that OMC has the strongest individual linear relationship with the target variable, Maximum Dry Density (MDD). However, the RReliefF scores, which assess feature importance in the context of interactions and non-linear dependencies, present a markedly different hierarchy. Here, Sand Content emerges as the most important feature (0.119), followed by Plastic Limit (0.099). OMC, despite its high univariate score, ranks near the bottom in RReliefF (0.070).

This critical discrepancy highlights that while OMC may be a powerful standalone linear indicator, its predictive utility is substantially diminished when considered within the complex, interactive feature space of the geotechnical system. Sand Content and Plastic Limit, conversely, provide more unique and non-redundant information that is crucial for the ensemble tree models (Random Forest and Gradient Boosting) to make accurate predictions. This underscores the necessity of employing multi-method feature selection to avoid over-relying on simple linear correlations and to identify features that offer genuine, conditional predictive power in machine learning contexts.

Table 2Feature importance ranking based on univariate regression and RReliefF scores. The table lists the six geotechnical features in descending order of univariate score, highlighting the discrepancy in feature significance between linear (Univariate Regression) and non-linear, multivariate (RReliefF) selection methods.

#	Feature Name	Univar. reg. Score	RReliefF Score
1	OMC (%)	4744.397	0.070
2	Fines Content (%)	773.744	0.088
3	Plastic Limit (%)	369.929	0.099
4	Liquid Limit (%)	301.381	0.075
5	Gravel Content (%)	178.857	0.073
6	Sand Content (%)	65.050	0.119

Shapely additive analysis

The SHAP (SHapley Additive exPlanations) analysis provides a critical, model-agnostic interpretation of feature importance and influence for both the Random Forest (RF) and Gradient Boosting (GB) regressors, moving beyond aggregate metrics to reveal the directional impact of each feature. The summary plots illustrate how each variable pushes the model's prediction of Maximum Dry Density (MDD) higher or lower. In both models, Optimum Moisture Content (OMC) consistently exhibits the widest dispersion of SHAP values spanning both positive and negative impacts confirming its role as the most influential yet complex feature. High OMC values generally correspond to negative SHAP values (reducing predicted MDD), which aligns with geotechnical principles where increased moisture typically reduces dry density beyond the optimum point. Notably, the SHAP analysis clarifies discrepancies from earlier feature selection results. While Sand Content ranked highest in RReliefF, its SHAP values in both models are concentrated near zero, indicating a more muted and consistent effect. Conversely, Gravel Content shows a strong positive impact in both model's higher gravel content consistently increases predicted MDD reinforcing its practical engineering significance. The SHAP plots also reveal subtle model-specific behaviors: in Gradient Boosting, Plastic Limit shows a tighter, more negative impact cluster, whereas in Random Forest its influence is more diffuse. The traditional feature importance bar charts (based on mean absolute SHAP values) uniformly rank OMC highest, followed by Gravel Content, largely mirroring the models' internal Gini or gain-based rankings. However, these bar charts lack the nuanced directional insight of the summary plots. Critically, SHAP confirms that both ensemble models rely on a similar hierarchy of features but utilize them differently: GB exhibits sharper value thresholds (steeper colour gradients), while RF shows smoother, more distributed impacts. This validates the models' structural differences GB's sequential error correction versus RF's independent tree averaging while affirming that OMC and Gravel Content are the dominant, physically interpretable drivers of MDD prediction in both frameworks.

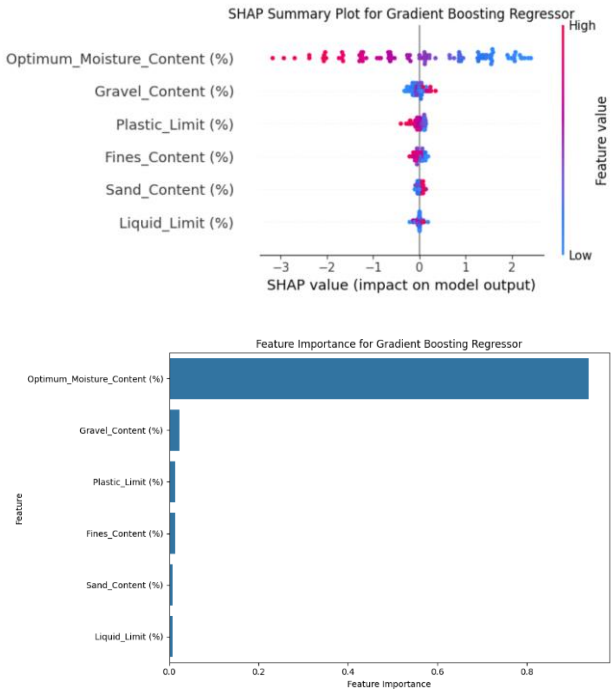


Figure 9 SHAP additive explanations for the Gradient Boosting Regressor. The summary plot (Left) illustrates how individual feature values influence model output across all predictions, and the feature importance bar chart (Right) ranks features based on their average contribution magnitude.

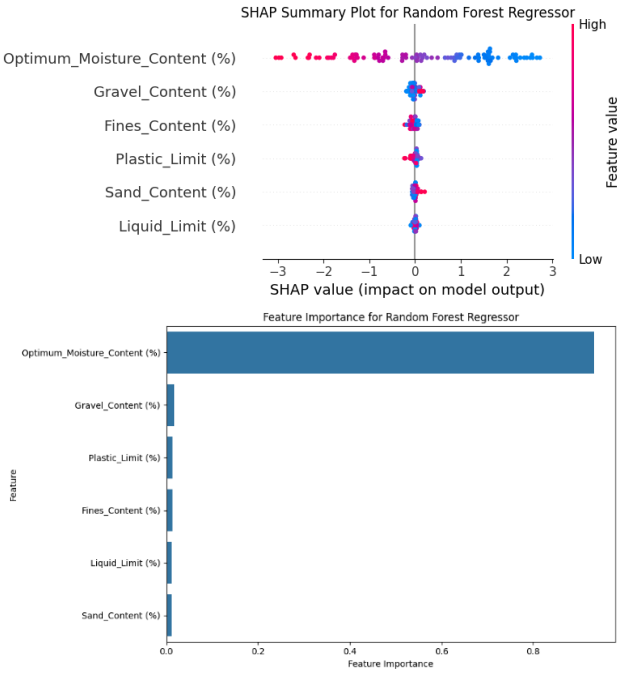


Figure 10 SHAP additive explanations for the Random Forest Regressor. The summary plot (Left) shows the distribution and directional impact (positive/negative) of each feature on the predicted Maximum Dry Density, while the bar chart (Right) presents the mean absolute SHAP value for each feature, representing overall importance.

Best model selection

Based on the comprehensive performance assessment, Random Forest is selected as the superior model for predicting Maximum Dry Density, despite the marginal nature of the decision. The selection is justified by its slightly better performance on the primary test metrics: Random Forest achieved a lower RMSE (0.489 vs. 0.498), a lower MAE (0.352 vs. 0.363), a higher R^2 (0.927 vs. 0.924), and a lower CVRMSE (2.64% vs. 2.69%). While Gradient Boosting demonstrated exceptional stability in k-fold validation with remarkably low variance, Random Forest's consistent edge in predictive accuracy on the unseen test set is decisive for deployment. Furthermore, the SHAP analysis confirmed that both models rely on a physically interpretable feature hierarchy, with OMC and Gravel Content being dominant, ensuring that Random Forest's predictions are not only accurate but also explainable. Although Random Forest exhibited higher fold-to-fold variability during cross-validation, its final generalization to the held-out test data was superior. Therefore, given the primary objective of minimizing prediction error on new data while maintaining model interpretability, Random Forest is recommended as the optimal model for this geotechnical regression task.

Discussion

The comparative evaluation of Random Forest (RF) and Gradient Boosting (GB) models for predicting Maximum Dry Density (MDD) in geotechnical contexts reveals nuanced strengths and limitations, particularly when scrutinized through the lens of error metrics, cross-validation stability, and feature importance. While the RF model achieved a marginally higher R^2 and lower RMSE and MAE on the test set, suggesting strong centralized predictive power, this apparent superiority is complicated by its pronounced variance across k-folds and a larger gap between training and test errors, which may indicate overfitting or excessive smoothing (Zhang et al., 2021; Almuaythir et al., 2025; Li et al., 2024; Abdullah et al., 2024; Tilahun et al., 2024; Chen et al., 2023; Canero et al., 2024; Otchere et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Chakraborty et al., 2020; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Elavarasan et al., 2021; Makungwe et al., 2021; Chen et al., 2024; University et al., 2025; Shahani et al., 2022; Siswadi et al., 2024). The literature consistently notes that RF's ensemble structure, which aggregates multiple decorrelated trees, is adept at capturing non-linearities and interactions, often yielding robust performance in high-dimensional or noisy datasets (Zhang et al., 2021; Zhang et al., 2025; Elavarasan et al., 2021; Canero et al., 2024). However, this same mechanism can render RF sensitive to the specific composition of training subsets, as evidenced by the wide performance fluctuations observed in cross-validation, echoing findings that RF may be less stable and more prone to overfitting, especially in spatially clustered or imbalanced data (Makungwe et al., 2021; Almuaythir et al., 2025; Zhang et al., 2021; Elavarasan et al., 2021; Canero et al., 2024; Zhang et al., 2025; Zhang & Wang, 2025; Abdullah et al., 2024; Chakraborty et al., 2020; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Otchere et al., 2021).

In contrast, the GB model, while fractionally trailing in centralized metrics, demonstrated remarkable consistency and generalization across all 20 folds, with minimal variance in predictive scores. This stability is a hallmark of boosting algorithms, which iteratively correct residual errors and are less susceptible to the idiosyncrasies of individual data partitions (Zhang et al., 2021; Abdullah et al., 2024; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Canero et al., 2024; Otchere et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Chakraborty et al., 2020; Elavarasan et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021). The literature supports this, with multiple studies reporting that GB and its variants (e.g., XGBoost, LightGBM) often outperform RF in terms of generalization, especially when hyperparameters are carefully tuned and early stopping is employed to mitigate overfitting (Abdullah et al., 2024; Demir & Şahin, 2022; Tran, 2022; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Canero et al., 2024; Otchere et al., 2021; Zhang et al., 2021; Zhang et al., 2025; Zhang &

Wang, 2025; Chakraborty et al., 2020; Cai et al., 2020; Elavarasan et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021). The error analysis further substantiates GB's advantage: its residuals are tightly clustered around zero, indicating not only high accuracy but also reliable error control, whereas RF's error distribution is more diffuse, with greater moderate deviations and higher sensitivity to underrepresented data regions (Zhang et al., 2021; Abdullah et al., 2024; Almuaythir et al., 2025; Tilahun et al., 2024; Li et al., 2024; Chen et al., 2023; Canero et al., 2024; Otchere et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Chakraborty et al., 2020; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Elavarasan et al., 2021; Makungwe et al., 2021; Chen et al., 2024; University et al., 2025; Shahani et al., 2022; Siswadi et al., 2024). Feature importance analysis exposes a critical methodological insight. Univariate regression overwhelmingly ranks Optimum Moisture Content (OMC) as the dominant predictor, but multivariate, non-linear methods like RReliefF and SHAP (as used in recent studies) often elevate features such as Sand Content and Plastic Limit, which provide more unique, non-redundant information in the context of complex interactions (Otchere et al., 2021; Zhang et al., 2021; Abdullah et al., 2024; Demir & Şahin, 2022; Tran, 2022; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Canero et al., 2024; Zhang et al., 2025; Zhang & Wang, 2025; Chakraborty et al., 2020; Cai et al., 2020; Elavarasan et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021). This discrepancy underscores the risk of over-relying on simple linear correlations in feature selection, a pitfall highlighted in the literature where ensemble models' performance is shown to depend heavily on the identification of features that contribute conditionally, rather than marginally, to predictive power (Otchere et al., 2021; Zhang et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Abdullah et al., 2024; Chakraborty et al., 2020; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Elavarasan et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Canero et al., 2024). The practical implication is that robust model development in geotechnical prediction must integrate both linear and non-linear feature selection strategies to avoid redundancy and maximize generalizability (Otchere et al., 2021; Zhang et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Abdullah et al., 2024; Chakraborty et al., 2020; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Elavarasan et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Canero et al., 2024).

Despite the strong centralized metrics of RF, its higher variance and sensitivity to data partitioning, as well as its susceptibility to overfitting and the selection of spurious predictors in spatially structured data, limit its reliability for operational deployment in dynamic or heterogeneous environments (Zhang et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021; Tilahun et al., 2024; Li et al., 2024; Chen et al., 2023; Canero et al., 2024; Otchere et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Abdullah et al., 2024; Chakraborty et al., 2020; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Elavarasan et al., 2021; Chen et al., 2024; University et al., 2025; Shahani et al., 2022; Siswadi et al., 2024). GB's superior cross-validation stability, tighter error control, and resilience to overfitting, as consistently reported in recent comparative studies, make it a more dependable choice for practical applications where reproducibility and robustness are paramount (Zhang et al., 2021; Abdullah et al., 2024; Demir & Şahin, 2022; Cai et al., 2020; Tran, 2022; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Canero et al., 2024; Otchere et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Chakraborty et al., 2020; Elavarasan et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021). Ultimately, the critical synthesis of these findings suggests that while both RF and GB are powerful tools for MDD prediction, the nuanced differences in their error profiles, generalization capacity, and feature selection implications must guide their selection and interpretation in geotechnical machine learning workflows (Otchere et al., 2021; Zhang et al., 2021; Zhang et al., 2025; Zhang & Wang, 2025; Abdullah et al., 2024; Chakraborty et al., 2020; Demir & Şahin,

2022; Cai et al., 2020; Tran, 2022; Elavarasan et al., 2021; Almuaythir et al., 2025; Makungwe et al., 2021; Tilahun et al., 2024; Chen et al., 2024; Li et al., 2024; University et al., 2025; Shahani et al., 2022; Chen et al., 2023; Siswadi et al., 2024; Canero et al., 2024).

Conclusion and Recommendations

In this study, a comparative analysis of Random Forest (RF) and Gradient Boosting (GB) algorithms was conducted to predict the Maximum Dry Density (MDD) of soil based on seven geotechnical properties, using a dataset of 486 samples. The results indicate that both ensemble models achieved high predictive accuracy, with R^2 values exceeding 0.92 on the test set. However, Random Forest demonstrated a slight quantitative edge, with marginally lower RMSE (0.489), MAE (0.352), and MAPE (1.945%), alongside a higher R^2 (0.927) compared to Gradient Boosting. This superior centralized performance validates RF as a robust tool for MDD prediction when evaluated on held-out data. Despite RF's stronger test metrics, Gradient Boosting exhibited superior stability and generalizability, as evidenced by its remarkably consistent performance across 20-fold cross-validation and its tightly clustered error distribution. This suggests that GB may be more reliable in real-world applications where data variability is high. Feature importance analysis revealed that Optimum Moisture Content (OMC) and Gravel Content were the most influential predictors in both models, though SHAP analysis highlighted OMC's complex, bidirectional impact on MDD.

Based on these findings, it is recommended to deploy Random Forest for operational prediction of MDD in geotechnical engineering due to its slightly higher accuracy and strong explanatory power. To enhance robustness, future work should incorporate advanced boosting variants (e.g., XGBoost, LightGBM) with early stopping and hyperparameter optimization. Additionally, feature engineering should prioritize multivariate, non-linear selection methods like RReliefF over univariate approaches to better capture interactive effects. It is further recommended to validate the model on larger, more geographically diverse datasets to assess its transferability, and to integrate SHAP-based interpretability tools into decision support systems to improve transparency and trust in machine-driven geotechnical predictions. Finally, real-time monitoring and periodic model retraining with new site data are advised to maintain predictive performance in dynamic engineering environments.

References

Abdulfattah, B.S. et al. (2023a) "Predicting implications of design changes in BIM-based construction projects through machine learning," Automation in Construction, 155, p. 105057. Available at: https://doi.org/10.1016/j.autcon.2023.105057.

Abdulfattah, B.S. et al. (2023b) "Predicting implications of design changes in BIM-based construction projects through machine learning," Automation in Construction, 155, p. 105057. Available at: https://doi.org/10.1016/j.autcon.2023.105057.

Abdullah, G. et al. (2024) "Boosting-based ensemble machine learning models for predicting unconfined compressive strength of geopolymer stabilized clayey soil," Scientific Reports, 14. Available at: https://doi.org/10.1038/s41598-024-52825-7.

Abed, M.S. et al. (2023) "Utilizing Multivariate Adaptive Regression Splines (MARS) for Precise Estimation of Soil Compaction Parameters," Applied Sciences [Preprint]. Available at: https://doi.org/10.3390/app132111634.

Abood, E. et al. (2024) "Machine Learning-Based Prediction Models for Punching Shear Strength of Fiber-Reinforced Polymer Reinforced Concrete Slabs Using a Gradient-Boosted Regression Tree," Materials, 17. Available at: https://doi.org/10.3390/ma17163964.

Adamolekun, L.B. et al. (2024) "Development of machine learning-based standalone GUI application for predicting hydraulic conductivity and compaction parameters of lateritic soils," Scientific African [Preprint]. Available at: https://doi.org/10.1016/j.sciaf.2024.e02393.

Afshari, S.S. et al. (2021a) "Machine learning-based methods in structural reliability

analysis: A review," Reliab. Eng. Syst. Saf., 219, p. 108223. Available at: https://doi.org/10.1016/j.ress.2021.108223.

Afshari, S.S. et al. (2021b) "Machine learning-based methods in structural reliability analysis: A review," Reliab. Eng. Syst. Saf., 219, p. 108223. Available at: https://doi.org/10.1016/j.ress.2021.108223.

Ali, H.F.H. et al. (2024) "Predicting the maximum dry density and optimum moisture content from soil index properties using efficient soft computing techniques," Neural Comput. Appl., 36, pp. 11339–11369. Available at: https://doi.org/10.1007/s00521-024-09734-7.

Ali, H.H. (2023a) "Utilizing multivariable mathematical models to predict maximum dry density and optimum moisture content from physical soil properties," Multiscale and Multidisciplinary Modeling, Experiments and Design, 6, pp. 603–627. Available at: https://doi.org/10.1007/s41939-023-00165-w.

Ali, H.H. (2023b) "Utilizing multivariable mathematical models to predict maximum dry density and optimum moisture content from physical soil properties," Multiscale and Multidisciplinary Modeling, Experiments and Design, 6, pp. 603–627. Available at: https://doi.org/10.1007/s41939-023-00165-w.

Almuaythir, S., Zaini, M. and Lodhi, R. (2025a) "Predicting soil compaction parameters in expansive soils using advanced machine learning models: a comparative study," Scientific Reports, 15. Available at: https://doi.org/10.1038/s41598-025-09279-2.

Almuaythir, S., Zaini, M. and Lodhi, R. (2025b) "Predicting soil compaction parameters in expansive soils using advanced machine learning models: a comparative study," Scientific Reports, 15. Available at: https://doi.org/10.1038/s41598-025-09279-2.

Ardakani, A. and Kordnaeij, A. (2019) "Soil compaction parameters prediction using GMDH-type neural network and genetic algorithm," European Journal of Environmental and Civil Engineering, 23, pp. 449–462. Available at: https://doi.org/10.1080/19648189.2017.1304269.

Bardhan, A. et al. (2023) "Modelling Soil Compaction Parameters Using an Enhanced Hybrid Intelligence Paradigm of ANFIS and Improved Grey Wolf Optimiser," Mathematics [Preprint]. Available at: https://doi.org/10.3390/math11143064.

Breiman, L. (2001) "Random Forests," Machine Learning, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Chicco, D., Warrens, M. and Jurman, G. (2021) "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, 7. Available at: https://doi.org/10.7717/peerj-cs.623.

Christopher, V.B. et al. (2023) "Prediction of weather using high-performance gradient boosting," International Journal of Global Warming, 31(1), pp. 30–41. Available at: https://doi.org/10.1504/ijgw.2023.133219.

Demir, S. and Şahin, E. (2025) "An innovative machine learning approach for slope stability prediction by combining shap interpretability and stacking ensemble learning.," Environmental science and pollution research international [Preprint]. Available at: https://doi.org/10.1007/s11356-025-36406-3.

Duc, M.N. et al. (2021) "An Artificial Intelligence Approach Based on Multi-layer Perceptron Neural Network and Random Forest for Predicting Maximum Dry Density and Optimum Moisture Content of Soil Material in Quang Ninh Province, Vietnam," Lecture Notes in Civil Engineering [Preprint]. Available at: https://doi.org/10.1007/978-981-16-7160-9_176.

Ewusi-Wilson, R. et al. (2025a) "Explainable Artificial Intelligence Estimation of Maximum Dry Density in Soil Compaction Based on Basic Soil Properties and Compaction Energy," Transportation Infrastructure Geotechnology [Preprint]. Available at: https://doi.org/10.1007/s40515-025-00552-5.

Ewusi-Wilson, R. et al. (2025b) "Explainable Artificial Intelligence Estimation of Maximum Dry Density in Soil Compaction Based on Basic Soil Properties and Compaction Energy," Transportation Infrastructure Geotechnology [Preprint]. Available at: https://doi.org/10.1007/s40515-025-00552-5.

Farooq, F. et al. (2020) "A Comparative Study of Random Forest and Genetic Engineering Programming for the Prediction of Compressive Strength of High Strength Concrete (HSC)," Applied Sciences [Preprint]. Available at: https://doi.org/10.3390/app10207330.

Gali, G. et al. (2023) "Experimental Study on the Effects of Freeze-Thaw Progressions and Performance of Soil with Non-Toxic Bio-Enzyme," Journal of ICT, Design, Engineering and Technological Science, 7(1), pp. 12–20. Available at: https://doi.org/10.33150/JITDETS-7.1.3.

Hodson, T. (2022) "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," Geoscientific Model Development [Preprint]. Available at: https://doi.org/10.5194/gmd-15-5481-2022.

Jin, L. et al. (2025) "Research on intelligent prediction method of roadbed compaction parameters," 13575, pp. 135752–135752. Available at: https://doi.org/10.1117/12.3062078.

Junwei et al. (2022) "A comprehensive comparison among metaheuristics (MHs) for geohazard modeling using machine learning: Insights from a case study of landslide displacement prediction," Eng. Appl. Artif. Intell., 114, p. 105150. Available at: https://doi.org/10.1016/j.engappai.2022.105150.

Ke, G. et al. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

Khan, W. et al. (2025) "Transforming Expansive Clays FROM Waste TO Value: Dual Ash Stabilization Strategy FOR Augmenting Black Cotton Subgrade Soil Performance," 11, pp. 5065–5073. Available at: https://doi.org/10.64252/cz07k049.

Khatti, J. and Grover, K. (2023a) "Prediction of compaction parameters for fine-grained soil: Critical comparison of the deep learning and standalone models," Journal of Rock Mechanics and Geotechnical Engineering [Preprint]. Available at: https://doi.org/10.1016/j.jrmge.2022.12.034.

Khatti, J. and Grover, K. (2023b) "Prediction of compaction parameters for fine-grained soil: Critical comparison of the deep learning and standalone models," Journal of Rock Mechanics and Geotechnical Engineering [Preprint]. Available at: https://doi.org/10.1016/j.jrmge.2022.12.034.

Khatti, J. and Grover, K. (2023c) "Prediction of compaction parameters of compacted soil using LSSVM, LSTM, LSBoostRF, and ANN," Innovative Infrastructure Solutions, 8, pp. 1–34. Available at: https://doi.org/10.1007/s41062-023-01048-2.

Khatti, J. and Grover, K. (2023d) "Prediction of compaction parameters of compacted soil using LSSVM, LSTM, LSBoostRF, and ANN," Innovative Infrastructure Solutions, 8, pp. 1–34. Available at: https://doi.org/10.1007/s41062-023-01048-2.

Li, B. et al. (2024a) "Prediction of Soil Compaction Parameters Using Machine Learning Models," Applied Sciences [Preprint]. Available at: https://doi.org/10.3390/app14072716.

Li, B. et al. (2024b) "Prediction of Soil Compaction Parameters Using Machine Learning Models," Applied Sciences [Preprint]. Available at: https://doi.org/10.3390/app14072716.

Li, C. and Gao, G. (2023) "Machine learning based inversion for earth rock dam compaction density," Developments in the Built Environment [Preprint]. Available at:

https://doi.org/10.1016/j.dibe.2023.100213.

Lin, S. et al. (2023) "A comprehensive evaluation of ensemble machine learning in geotechnical stability analysis and explainability," International Journal of Mechanics and Materials in Design [Preprint]. Available at: https://doi.org/10.1007/s10999-023-09679-0.

Liu, X. et al. (2025) "Analyzing the influence of particle size distribution on the maximum shear modulus of soil with an interpretable machine learning framework and laboratory test dataset," Soil Dynamics and Earthquake Engineering [Preprint]. Available at: https://doi.org/10.1016/j.soildyn.2024.109031.

Mujtaba, G. et al. (2023) "Water Pollution Hazards and Toxicity Caused by Textile Industries Effluent," Journal of ICT, Design, Engineering and Technological Science, 7(2), pp. 22–27. Available at: https://doi.org/10.33150/JITDETS-7.2.3.

Natras, R., Soja, B. and Schmidt, M. (2023) "Uncertainty Quantification for Machine Learning-Based Ionosphere and Space Weather Forecasting: Ensemble, Bayesian Neural Network, and Quantile Gradient Boosting," Space Weather, 21(10). Available at: https://doi.org/10.1029/2023sw003483.

Nawaz, S. et al. (2025) "Flexural Performance Evaluation Of Glass Fiber Reinforced Polymer Bars (GFRP) In Doubly Reinforced Beams: An Experimental Approach."

Nuthalapati, undefined babu and Nuthalapati, undefined A. (2024) "Accurate weather forecasting with dominant gradient boosting using machine learning," International Journal of Science and Research Archive, 12(2), pp. 408–422. Available at: https://doi.org/10.30574/ijsra.2024.12.2.1246.

Oka, S. and Takefuji, Y. (2025) "Comments on 'Dialogue between algorithms and soil: Machine learning unravels the mystery of phthalates pollution in soil' by Pan et al. (2025).," Journal of hazardous materials, 493, p. 138366. Available at: https://doi.org/10.1016/j.jhazmat.2025.138366.

Onyelowe, K. et al. (2024) "Estimating the strength of soil stabilized with cement and lime at optimal compaction using ensemble-based multiple machine learning," Scientific Reports, 14. Available at: https://doi.org/10.1038/s41598-024-66295-4.

Pham, K. et al. (2021) "Ensemble learning-based classification models for slope stability analysis," CATENA [Preprint]. Available at: https://doi.org/10.1016/j.catena.2020.104886.

Rehman, Z.U. et al. (2024a) "Big data-driven global modeling of cohesive soil compaction across conceptual and arbitrary energies through machine learning," Transportation Geotechnics [Preprint]. Available at: https://doi.org/10.1016/j.trgeo.2024.101470.

Rehman, Z.U. et al. (2024b) "Big data-driven global modeling of cohesive soil compaction across conceptual and arbitrary energies through machine learning," Transportation Geotechnics [Preprint]. Available at: https://doi.org/10.1016/j.trgeo.2024.101470.

Shah, Q. et al. (2023) "To Stabilize Shear Strength Properties of an Unwanted Subgrade Soil Utilizing Rock Dust," Journal of ICT, Design, Engineering and Technological Science, 7(2), pp. 1–6. Available at: https://doi.org/10.33150/JITDETS-7.2.1.

Taffese, W. and Abegaz, K.A. (2022) "Prediction of Compaction and Strength Properties of Amended Soil Using Machine Learning," Buildings [Preprint]. Available at: https://doi.org/10.3390/buildings12050613.

Taha, O., Majeed, Z. and Ahmed, S. (2018) "Artificial Neural Network Prediction Models for Maximum Dry Density and Optimum Moisture Content of Stabilized Soils," Transportation Infrastructure Geotechnology, 5, pp. 146–168. Available at: https://doi.org/10.1007/s40515-018-0053-2.

Tilahun, Y. et al. (2024) "Determination of Compaction Parameters of Cement-Lime Soils: Boosting-Based Ensemble Models," International Journal for Numerical and Analytical Methods in Geomechanics, 48. Available at: https://doi.org/10.1002/nag.3846.

Tiwari, L.B., Burman, A. and Samui, P. (2024) "A Comparative Study of Soft Computing Paradigms for Modelling Soil Compaction Parameters," Transportation Infrastructure Geotechnology [Preprint]. Available at: https://doi.org/10.1007/s40515-024-00436-0.

Tran, V. (2022) "Hybrid gradient boosting with meta-heuristic algorithms prediction of unconfined compressive strength of stabilized soil based on initial soil properties, mix design and effective compaction," Journal of Cleaner Production [Preprint]. Available at: https://doi.org/10.1016/j.jclepro.2022.131683.

Wang, M. et al. (2025) "Prediction of Soil Pollution Risk Based on Machine Learning and SHAP Interpretable Models in the Nansi Lake, China," Toxics, 13. Available at: https://doi.org/10.3390/toxics13040278.

Zhang, P., Yin, Z. and Jin, Y. (2021) "Machine Learning-Based Modelling of Soil Properties for Geotechnical Design: Review, Tool Development and Comparison," Archives of Computational Methods in Engineering, 29, pp. 1229–1245. Available at: https://doi.org/10.1007/s11831-021-09615-5.

Zhao, L. and Guan, G.D. (2024) "Maximum dry density estimation of stabilized soil via machine learning techniques in individual and hybrid approaches," J. Ambient Intell. Humaniz. Comput., 15, pp. 3831–3846. Available at: https://doi.org/10.1007/s12652-024-04860-5.

Zhao, Q. et al. (2024) "Estimating the Maximum Dry Density of Soil via Least Square Support Vector Regression Individual and Hybrid Forms," Indian Geotechnical Journal [Preprint]. Available at: https://doi.org/10.1007/s40098-024-00952-3.

Zheng, W. (2023) "A novel method for dry density forecasting of high-speed railway graded aggregate fillers based on the PSO-ANN model," Results in Engineering [Preprint]. Available at: https://doi.org/10.1016/j.rineng.2023.101361.