

Improving the prediction of soil compaction parameters using machine learning models

Z. Matougui
A. Medjnoun
L. Djerbal
R. Bahar

Department of Civil Engineering, Laboratory of LEEGO, University of Sciences and Technology Houari Boumerdiene (USTHB), Algeria

ABSTRACT: This study aims to enhance the predictive accuracy of fundamental soil compaction parameters, specifically maximum dry density and optimum moisture content, through the application of machine learning models. To achieve this, Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Gradient Boosting (GB) algorithms were employed, coupled with advanced validation techniques using cross-validation approach, utilizing a pre-existing dataset from the literature. This investigation highlights the inherent biases associated with the validation methods employed in the baseline study. Furthermore, the findings demonstrate a notable enhancement in both the accuracy and reliability of predictions, highlighting the efficacy of the proposed methodology.

1 INTRODUCTION

Estimation studies, often used in project management, offer numerous advantages for effective planning, decision-making, and successful project execution. In geotechnical engineering, estimation studies play a important role in evaluating various parameters related to soil and rock mechanics, foundation design, and earthworks.

Soil compaction, a process involving the application of pressure to soil to reduce its volume and increase its density, holds significant applications across various fields, particularly in geotechnical engineering. This physical process of densifying soil contributes to its bearing capacity, enhancing shear strength, as well as reducing permeability and compressibility. Essential compaction parameters, namely maximum dry density (MDD) and optimum moisture content (OMC), are derived from compaction test results. These parameters are crucial in geotechnical engineering for structures like earth dams, motorway embankments, bridge abutments, and fills behind retaining walls.

This study aims to refine the estimation of compaction parameters using modern predictive tools such as machine learning. A prior investigation by (Günaydın 2009) relied primarily on simple artificial neural networks (ANN) and regression analysis, with a focus on ANN models and statistical methods (Alawi and Rajab 2013; González Farias, Araujo, and Ruiz 2018; Ul Rehman 2017). This research examines the efficacy of various machine learning algorithms beyond ANNs, including Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Gradient Boosting (GB). The goal is to identify the optimal algorithm or combination of algorithms for predicting soil compaction parameters (MDD and OMC) with enhanced accuracy and reliability. This is achieved through rigorous evaluation of different machine learning models, employing a comprehensive data science approach and rigorous validation procedures, including cross-validation techniques to assess model generalization and performance. Additionally, factor selection prioritizes non-collinear parameters to ensure the robustness and accuracy of predictive models. Advanced parameter selection techniques are utilized to pinpoint the most informative predictors for model training. By addressing potential biases and limitations observed in previous methods, our study aims to provide more robust and generalizable predictions of soil compaction parameters through a systematic and data-driven approach.

2 MATERIAL

The dataset of this study consists of 127 samples, and various properties are measured and reported, including percentages of fines (FG), sand, gravel, and specific density (Gs), as well as liquid limit (w_L), plastic limit (w_p), plasticity index (IP), and soil type (ST). Additionally, the compaction parameters MDD and OMC are included. Table 1 provides a summary of the soil composition and compaction parameters for fine-grained soils collected from dams in the Nigde region of Turkey mentioned in the article of Günaydın (Günaydın 2009).

Table 1 Summary statistics of soil composition and compaction parameters

	FG	Sand	Gravel	Gs	w_L	w_p	ST	Ip	OMC	MDD
mean	48.55	37.60	13.85	2.73	39.73	21.61	4.64	18.12	16.09	17.65
std	14.41	11.86	12.93	0.04	7.16	4.03	2.65	4.56	2.88	0.98
min	13.00	15.49	0.05	2.66	25.00	14.21	1.00	7.88	7.60	16.08
25%	38.00	28.70	4.16	2.70	34.10	18.72	2.00	14.95	14.30	16.93
50%	49.00	35.62	10.11	2.73	39.00	21.78	4.00	17.63	16.10	17.64
75%	60.00	44.95	22.77	2.76	46.05	24.90	8.00	20.94	18.00	18.12
max	83.30	71.26	67.10	2.85	55.00	29.84	8.00	29.17	23.75	20.51

According to Table 1, the fines content (FG) exhibits a mean of 48.55%, with a wide range from 13.00% to 83.30%, indicating significant variability among samples. Similarly, the mean sand content is 37.60%, with a standard deviation of 11.86%, reflecting considerable variability.

In contrast, the gravel content is lower on average, with a mean of 13.85% and a standard deviation of 12.93%. The specific density shows relatively consistent values across samples, with a mean of 2.73 g/cm³ and a minimal standard deviation of 0.04 g/cm³. The liquid limit (w_L) and plastic limit (w_p) exhibit more variability, with mean values of 39.73% and 21.61%, respectively, and standard deviations of 7.16% and 4.03%. The plasticity index (I_p) has a mean value of 4.64, indicating the range of moisture contents over which the soil exhibits plastic behaviour. The OMC and maximum dry density (MDD) present values of 16.09% and 17.65%, respectively, with standard deviations of 2.88% and 0.98%. These results suggest relatively small but significant variations in compaction behaviour among the samples. Overall, these statistics provide valuable insights into the soil characteristics and compaction parameters, essential for geotechnical engineering applications.

3 METHOD

The methodology depicted in Figure 1 outlines the process followed in this paper. Data acquisition from literature archive search initiated the study, after which a rigorous approach was adopted for data preprocessing, multicollinearity analysis, splitting, and model calibration. A 70/30 proportion was utilised, allocating 70% of the data to the training set and 30% to the testing set, ensuring sufficient data for model calibration while retaining a suitable portion for evaluation. To mitigate the risk of overfitting and ensure the robustness of the predictive models, cross-validation techniques were employed during model calibration. By partitioning the training data into subsets for training and validation and averaging the model performance across these subsets, more reliable estimates of model performance were obtained, reducing the likelihood of overfitting. Additionally, the parameters selected for the models were informed by a comprehensive collinearity study and feature selection algorithm. This approach facilitated the identification and prioritisation of the most relevant features for predicting OMC and MDD, based on their correlations with the target variables and their potential to enhance model performance.

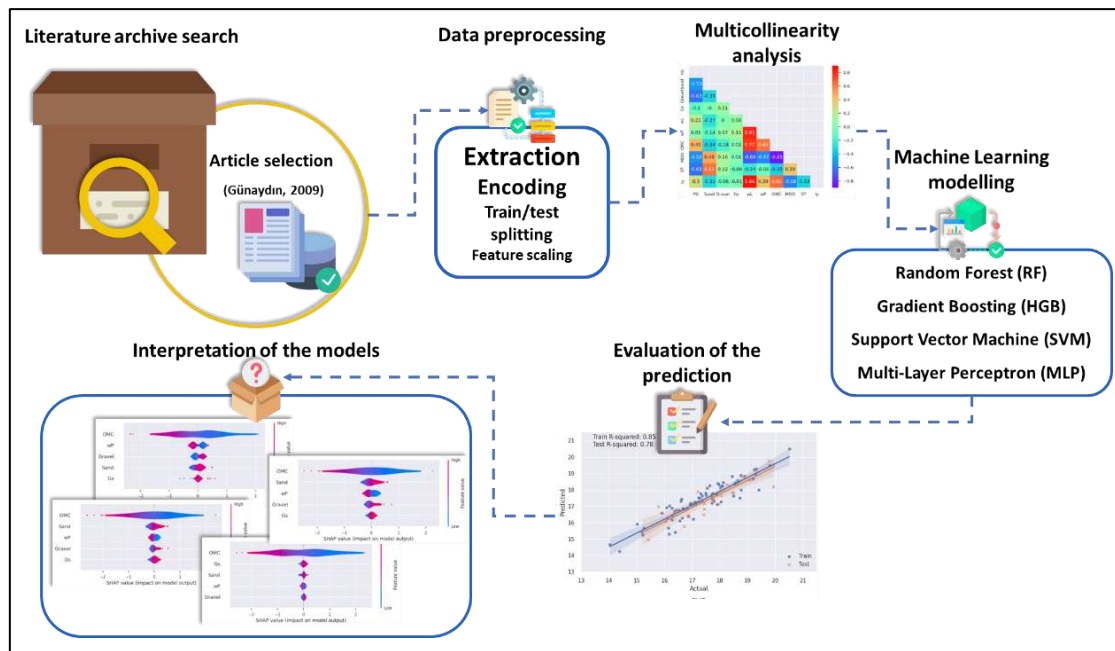


Figure 1 Flowchart of the methodology employed in this study

In summary, this study improved upon previous research by implementing a well-defined data splitting procedure, employing cross-validation techniques for model calibration, and adopting an informed approach to parameter selection based on rigorous analyses of feature correlations and selection algorithms. These methodological refinements enhance the reliability and validity of our predictive models for estimating OMC and MDD in soil samples.

3.1 Machine learning algorithms

Several machine learning algorithms are employed to predict the OMC and MDD of soil samples. The algorithms used include:

- **MLP Regressor (Multi-Layer Perceptron Regressor):** This algorithm is a type of feedforward artificial neural network that uses multiple layers to perform regression tasks. It learns from the dataset by adjusting the weights between neurons to minimize the difference between predicted and actual values.
- **Random Forest Regressor:** This algorithm is an ensemble learning method that constructs a multitude of decision trees during training and outputs the average prediction of individual trees. It is robust to overfitting and can handle large datasets with high dimensionality (Breiman 2001).
- **SVR (Support Vector Regressor):** SVR is a type of Support Vector Machine (SVM) that is adapted for regression tasks. It works by mapping input data into a high-dimensional feature space and finding the hyperplane that best separates the data points while maximizing the margin (Vanneschi and Silva 2023).
- **Gradient Boosting Regressor:** This algorithm builds an ensemble of decision trees sequentially, where each tree corrects the errors of its predecessor. It minimizes a loss function by adding decision trees in a forward stage-wise manner (Friedman 2002).

These algorithms are popular for regression and classification cases and have proven their effectiveness (Eyo et al. 2022; Matougui et al. 2023a; Pham et al. 2019). By utilising these machine learning algorithms, we aim to develop predictive models that can more accurately estimate the OMC and MDD of soil samples.

3.2 Evaluation of the models

In order to evaluate the performance of our machine learning models for predicting the OMC and MDD of soil samples, we employed the following evaluation metrics:

- **R-squared (R2):** R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (target) that is explained by the independent variables (features) in the model. It ranges from 0 to 1, where 1 indicates a perfect fit. A higher R2 value indicates better model performance.
- **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted values and actual values. It provides a straightforward interpretation of the average magnitude of errors in the predictions. Lower MAE values indicate better model accuracy.
- **Root Mean Squared Error (RMSE):** RMSE is a measure of the average magnitude of the errors between predicted and actual values, with the errors being squared before averaging. RMSE provides a measure of how spread out the errors are, with lower values indicating better model performance.

4 RESULTS AND DISCUSSION

4.1 Multicollinearity result

From the correlation matrix presented in Figure 2, several notable correlations emerge. FG (fines content) demonstrates moderate negative correlations with sand (-0.53) and gravel (-0.63), alongside a moderate positive correlation with OMC (0.45). Sand exhibits moderate positive correlations with gravel (0.33) and MDD (0.49), while gravel displays a weak positive correlation with MDD (0.16). Liquid limit (w_L) and plastic limit (w_p) demonstrate strong positive correlations with each other (0.81) and with plasticity index (IP) (0.81 and 0.81, respectively). OMC shows robust positive correlations with w_L (0.77), w_p (0.63), and IP (0.65), alongside a strong negative correlation with MDD (-0.85). MDD exhibits strong negative correlations with w_L (-0.69), w_p (-0.57), and IP (-0.58), along with a strong positive correlation with ST (soil type) (0.39).

Given the collinearity within the dataset, it is crucial to consider variables that are not highly correlated with each other when predicting OMC and MDD, to avoid issues of multicollinearity. Potential predictors for OMC, based on their lower collinearity with other factors, include Gs, ST, w_L , sand, and possibly gravel content. For MDD, variables such as ST, Gs, sand, gravel, OMC, and w_p could serve as effective predictors. These variables demonstrate relatively weaker correlations with other factors in the dataset, underscoring their potential utility in predicting OMC and MDD. Nonetheless, further analysis, such as feature selection techniques, would be indispensable for determining the most effective predictors.

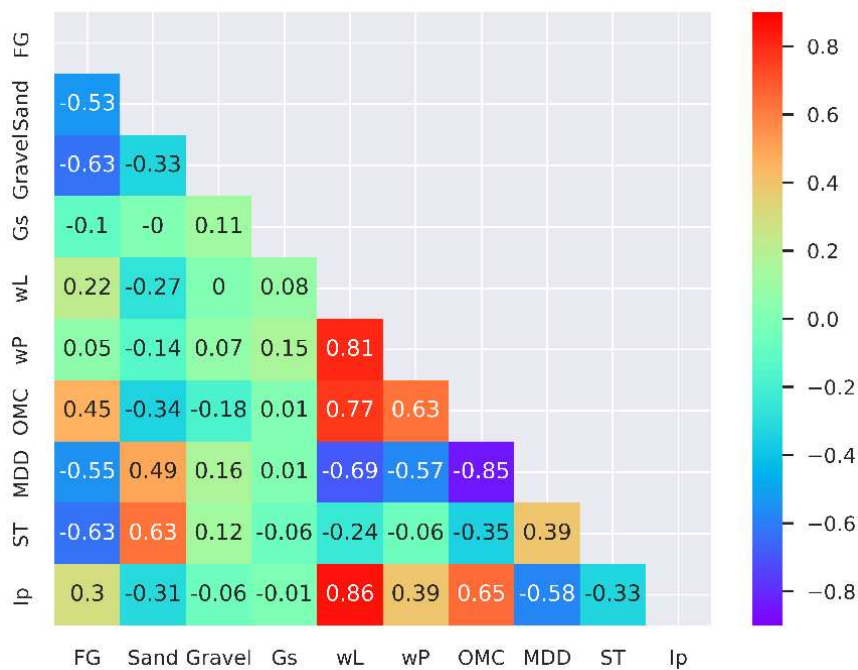


Figure 2 Correlation of parameters

4.2 Performance evaluation

After extensive data preprocessing, parameter selection, model calibration, and training, the performance of various machine learning algorithms were evaluated in predicting OMC and MDD. The results are presented in Table 2. For OMC estimation, the RF slightly outperformed the MLP with an R2 value of 0.77 compared to 0.75. The Random Forest model also showed lower MAE and RMSE values (0.727 and 0.931, respectively) than the MLP. Other models, including SVR and Gradient Boosting, also demonstrated competitive performance with R2 values around 0.73-0.75. In predicting MDD, the MLP performed the best with an R2 of 0.78, closely followed by the SVR model with an R2 of 0.75. Both models showed low MAE and RMSE values. The RF also performed well, albeit slightly lower than the MLP and SVR.

Overall, these results highlight the effectiveness of machine learning algorithms in predicting OMC and MDD. The RF and MLP emerged as top performers, demonstrating strong correlations and relatively low errors in both predictions.

Table 2 Performance of proposed models for test and train datasets

	Models	R ²	MAE	RMSE
OMC	MLP Regressor	0.75	1.237	1.568
	Random Forest Regressor	0.77	0.727	0.931
	SVR	0.73	1.330	1.669
	Gradient Boosting Regressor	0.75	0.820	1.083
	MLP Regressor	0.78	0.373	0.499
MDD	Random Forest Regressor	0.71	0.298	0.420
	SVR	0.75	0.380	0.516
	Gradient Boosting Regressor	0.73	0.246	0.373

Figure 3 illustrates the actual and predicted values of OMC for both the training and testing sets. Similarly, Figure 4 presents the actual and predicted values of MDD for the training and testing sets. In both figures, the MLP and SVR models demonstrate consistency in their predictions, showing minimal discrepancy between the training and testing sets. Conversely, the

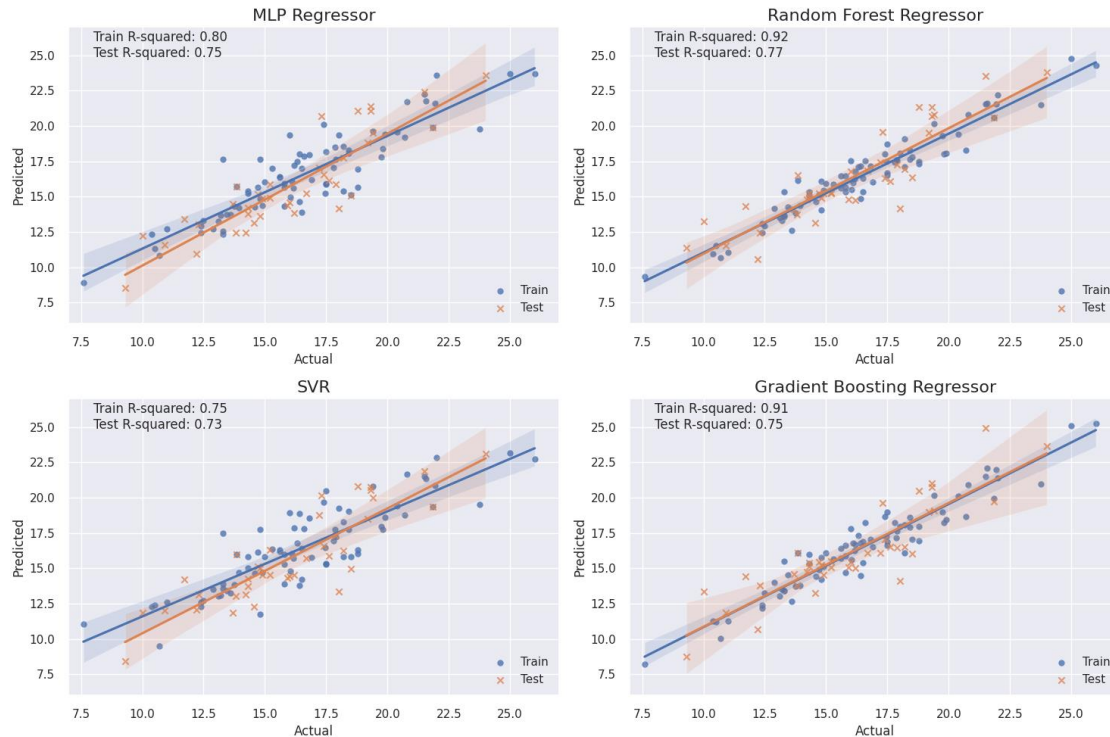


Figure 3 Actual and predicted values of OMC for the training and testing sets

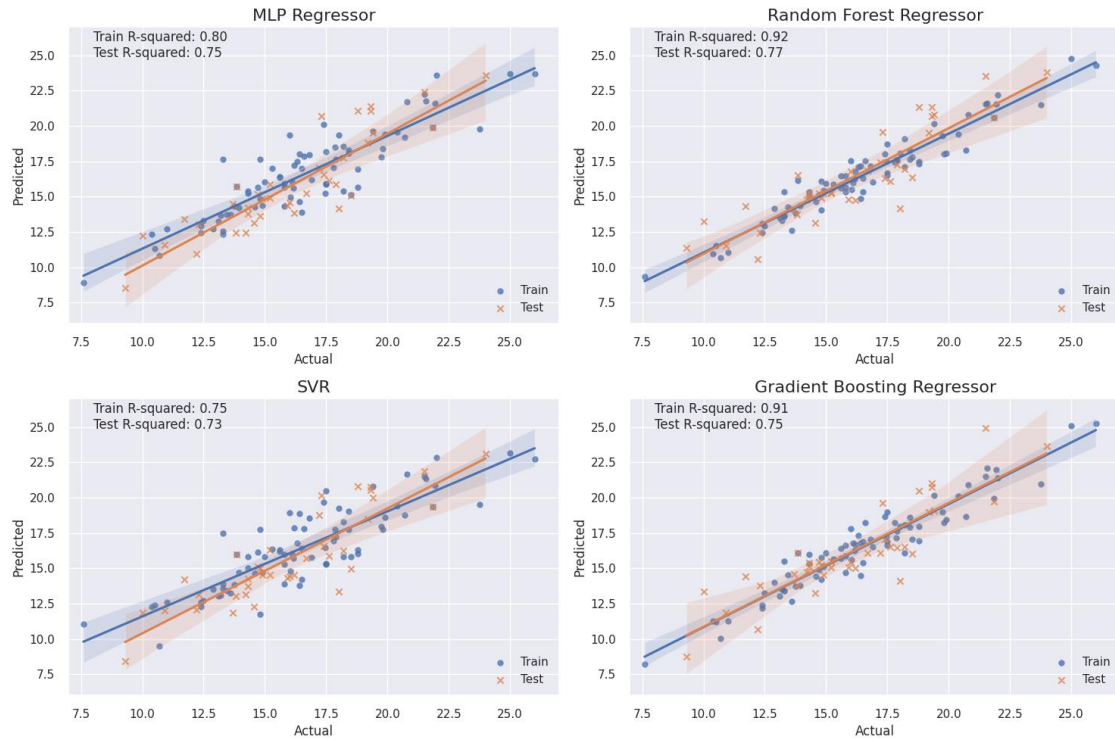


Figure 4 Actual and predicted values of MDD for the training and testing sets

decision tree-based models, such as RF and Gradient Boosting, exhibit a larger gap between the predicted values of the training and testing sets for both OMC and MDD models.

This discrepancy suggests that MLP and SVR models may have generalized better to unseen data compared to the decision tree-based models, which might be prone to overfitting. The consistent performance of MLP and SVR models across both training and testing sets underscores their reliability and robustness in predicting OMC and MDD values.

For the OMC model, a recursive feature elimination process identified key parameters as w_L , Sand, and Gravel. This selection underscores the significant influence of these factors on predicting the optimum moisture content of the soil, we can inspect the influence of these parameters on the OMC in Figure 5. Interestingly, for the MDD model, the output of the OMC model was incorporated as a parameter alongside w_p , Gs, Sand, and Gravel. This approach highlights the interdependency between the optimum moisture content and maximum dry density of the soil. By incorporating the output of the OMC model, the MDD model leverages the relationship between moisture content and compaction characteristics, enhancing its predictive accuracy. Overall, these parameter selections reflect a nuanced understanding of the underlying factors affecting soil behaviour and compaction properties.

4.3 Interpretation of the models

The models proposed in this study exhibit superior performance compared to previous methods, despite employing a methodology that is both different and more constrained. Even with a more rigorous approach, our models outperform their predecessors. With the current state-of-the-art techniques, we can delve deeper into interpreting the models and alleviate their black-box nature. One such method is the SHAP (SHapley Additive exPlanations) algorithm, which provides insights into the importance of each parameter in the model's predictions. Figure 5 presents the SHAP summary plot for the proposed models, offering a visual representation of the impact of different features on the model's output. This plot enables us to understand the relative contribution of each feature to the model's predictions, thereby enhancing the interpretability of the models.

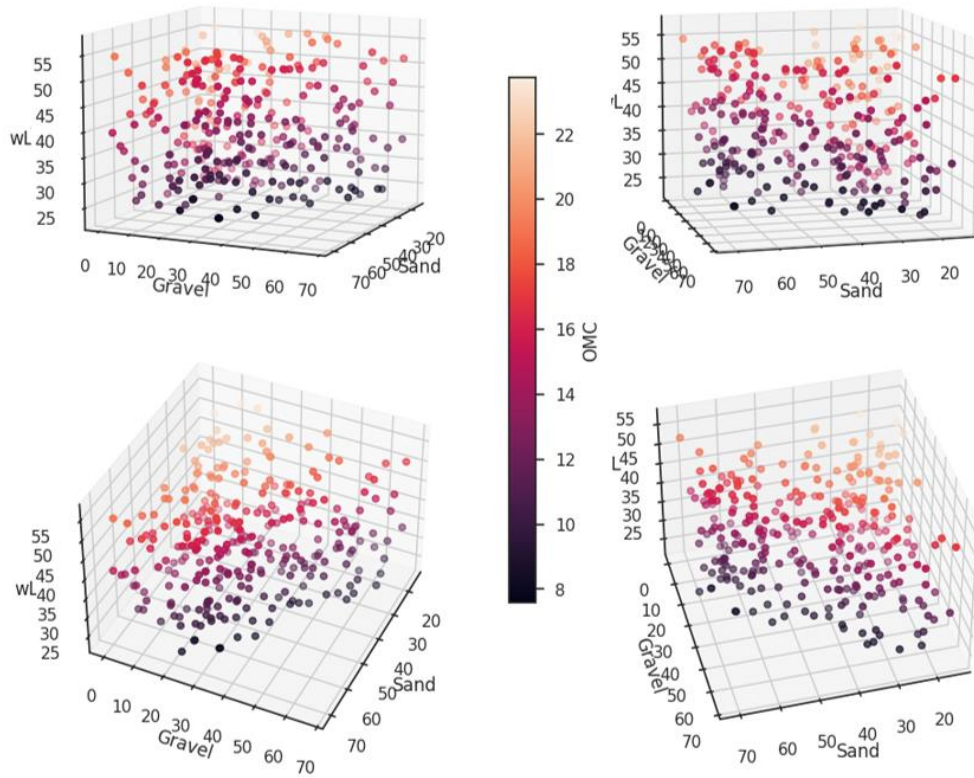


Figure 5 Relationship between the parameters and the OMC

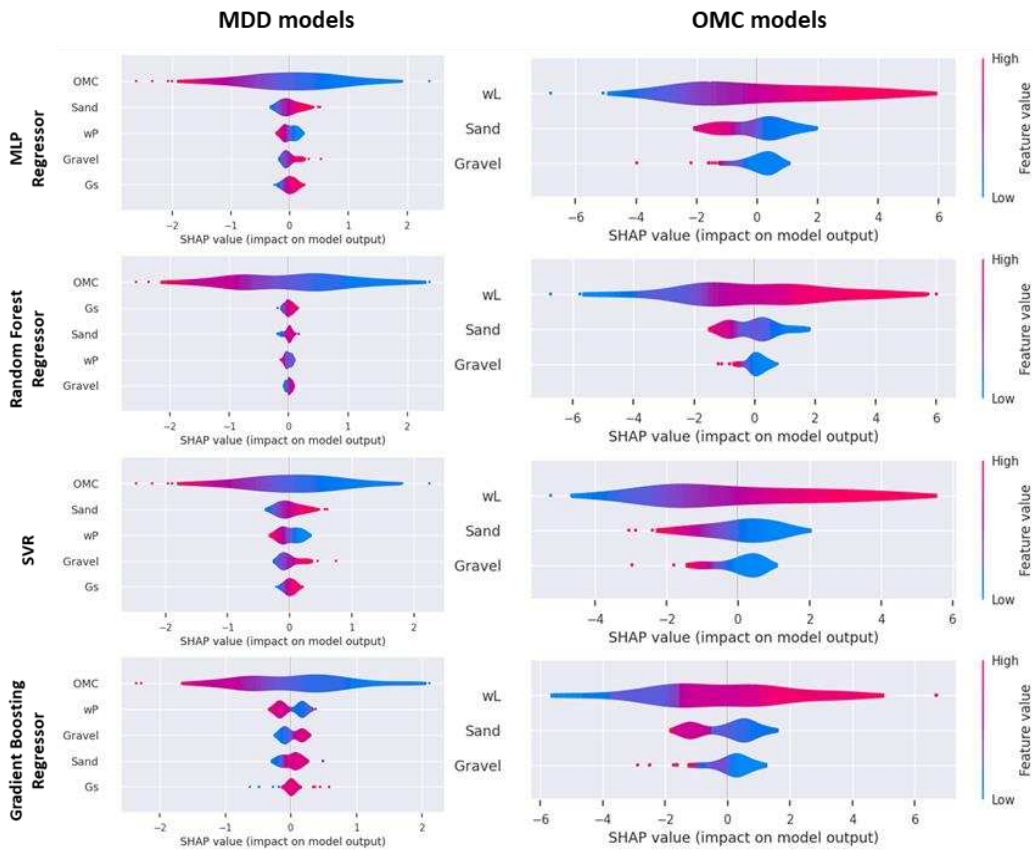


Figure 6 SHAP summary plot for the proposed models

5 CONCLUSION

The findings of this study not only validate but also improve upon previous research by enhancing the accuracy and reliability of predictions. Models developed in this study serve as effective tools for approximating target parameters in preliminary studies while minimizing the number of data points required to ascertain the optimum conditions.

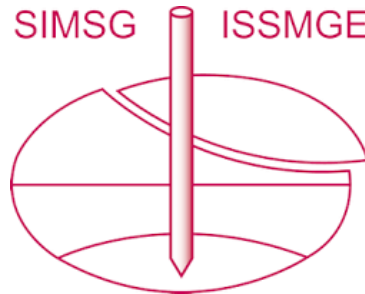
The incorporation of the SHAP interpretation method has notably contributed to a deeper understanding of each parameter's impact within the models. Specifically, it has elucidated that w_L holds the greatest influence on OMC prediction, underscoring the pivotal role of fines content in determining OMC. Additionally, w_L provides insights not only into the proportion of fines but also into their composition, further enhancing the model's predictive capabilities. Conversely, the influence of Sand and Gravel on OMC is comparatively lesser. In the case of MDD prediction, OMC emerges as the most influential parameter across all models. The significance of other parameters varies across algorithms, with their influence being relatively weaker compared to OMC.

Overall, these insights shed light on the complex interplay between different soil parameters and their impact on soil behaviour. By leveraging advanced interpretation techniques like SHAP, this study not only improves model performance but also provides valuable insights into the underlying mechanisms governing soil properties.

6 REFERENCES

- Alawi, Mohammad H., and Maher I. Rajab. 2013. 'Prediction of California Bearing Ratio of Subbase Layer Using Multiple Linear Regression Models.' *Road Materials and Pavement Design* 14(1):211–19. doi: 10.1080/14680629.2012.757557.
- Breiman, Leo. 2001. 'Random Forests.' *Machine Learning* 45(1):5–32. doi: 10.1023/A:1010933404324.
- Eyo, E. U., S. J. Abbey, T. T. Lawrence, and F. K. Tetteh. 2022. 'Improved Prediction of Clay Soil Expansion Using Machine Learning Algorithms and Meta-Heuristic Dichotomous Ensemble Classifiers.' *Geoscience Frontiers* 13(1):101296. doi: 10.1016/j.gsf.2021.101296.
- Friedman, Jerome H. 2002. 'Stochastic Gradient Boosting.' *Computational Statistics and Data Analysis* 38(4):367–78. doi: 10.1016/S0167-9473(01)00065-2.
- González Farias, Isabel, William Araujo, and Gaby Ruiz. 2018. 'Prediction of California Bearing Ratio from Index Properties of Soils Using Parametric and Non-Parametric Models.' *Geotechnical and Geological Engineering* 36:3485–98. doi: 10.1007/s10706-018-0548-1.
- Günaydın, O. 2009. 'Estimation of Soil Compaction Parameters by Using Statistical Analyses and Artificial Neural Networks.' *Environmental Geology* 57(1):203–15. doi: 10.1007/s00254-008-1300-6.
- Matougui, Zakaria, Lynda Djerbal, and Ramdane Bahar. 2023a. 'A Comparative Study of Heterogeneous and Homogeneous Ensemble Approaches for Landslide Susceptibility Assessment in the Djebahia Region, Algeria.' *Environmental Science and Pollution Research* (0123456789). doi: 10.1007/s11356-023-26247-3.
- Matougui, Zakaria, Lynda Djerbal, and Ramdane Bahar. 2023b. 'Bagging Ensemble Based on Multi-Layer Perceptron Neural Network for Landslide Susceptibility Assessment.' *2023 International Conference on Earth Observation and Geo-Spatial Information, ICEOGI 2023* (MI):1–6. doi: 10.1109/ICEOGI57454.2023.10292962.
- Pham, Binh Thai, Manh Duc Nguyen, Kien Trinh Thi Bui, Indra Prakash, Kamran Chapi, and Dieu Tien Bui. 2019. 'A Novel Artificial Intelligence Approach Based on Multi-Layer Perceptron Neural Network and Biogeography-Based Optimization for Predicting Coefficient of Consolidation of Soil.' *Catena* 173(October 2018):302–11. doi: 10.1016/j.catena.2018.10.004.
- Ul Rehman, A. et al. 2017. 'Prediction of California Bearing Ratio (Cbr) and Compaction Characteris-.' *Acta Geotechnica Slovenica* 1(January):63–72.
- Vanneschi, Leonardo, and Sara Silva. 2023. 'Support Vector Machines.' *Natural Computing Series* 271–81. doi: 10.1007/978-3-031-17922-8_10.

INTERNATIONAL SOCIETY FOR SOIL MECHANICS AND GEOTECHNICAL ENGINEERING



This paper was downloaded from the Online Library of the International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE). The library is available here:

<https://www.issmge.org/publications/online-library>

This is an open-access database that archives thousands of papers published under the Auspices of the ISSMGE and maintained by the Innovation and Development Committee of ISSMGE.

The paper was published in the proceedings of the 18th African Regional Conference on Soil Mechanics and Geotechnical Engineering and was edited by Abdelmalek Bekkouche. The conference was held from October 6th to October 9th 2024 in Algiers, Algeria.