Journal of Multidisciplinary Engineering Science Studies (JMESS)
ISSN: 2458-925X
Vol. 11 Issue 12, December - 2025

Xgboost Model For Ground Truthing Of Electro-Optical Distance Ranging Crude Oil Storage Tank Volume Calibration Dataset

Emmanuel Udama Odeh[1]
Department of Mechanical and Aerospace Engineering,
University of Uyo, Uyo Akwa Ibom State Nigeria
emmanuelodeh@uniuyo.edu.ng

John Dennis Urua[2]
Department of Mechanical and Aerospace Engineering,
University of Uyo, Uyo Akwa Ibom State Nigeria
johnurua1@gmail.com

Uwah Etebom Francis[3]
Department of Mechanical and Aerospace Engineering,
University of Uyo, Uyo Akwa Ibom State Nigeria
francis.nerster@gmail.com

***Abstract*—Accurate calibration of crude oil storage tank volumes is essential for inventory management and custody transfer, traditionally requiring laborious Manual Strapping Methods (MSM). While Electro-Optical Distance Ranging (EODR) offers a faster, non-intrusive alternative, it often suffers from lower precision compared to ground-truth manual methods. This study develops an XGBoost machine learning regression model to calibrate and enhance the accuracy of EODR datasets by training them against paired, high-accuracy MSM ground-truth data. Preprocessed EODR, tank depth, and structural parameters were utilized to predict MSM-level volumes, training a regression model to correct for systematic EODR measurement deviations. The model's performance was analyzed across multiple epochs, showing significant improvement in precision through progressive optimization, reducing MSE by approximately 40% at the final epoch compared to initial stages. The optimized XGBoost model (Epoch 5) achieved a 0.929% accuracy (1-MAPE), with a Mean Squared Error (MSE) of 0.0910, Root Mean Square Error (RMSE) of 0.3016, and Mean Absolute Error (MAE) of 0.2474. Residual analysis indicated high reliability, with errors ranging from a minimum of -0.036 to a maximum of 0.5898 across varied tank depths. The results demonstrate that the proposed XGBoost model efficiently converges to deliver superior accuracy, serving as a robust tool for bridging the gap between rapid EODR data acquisition and high-precision manual calibration standards.**

Keywords*—XGBoost Model, Manual Strapping Method (MSM), Ground Truthing, Electro-Optical Distance Ranging (EODR), Crude Oil Storage Tank Calibration

1. Introduction

The precise determination of crude oil storage tank capacity, commonly known as tank calibration or strapping, is a critical process in the petroleum industry [1,2]. It ensures the accurate inventory management, custody transfer, and financial settlement of crude oil, as a highly accurate calibration chart (tank table) allows for the calculation of exact liquid volumes based on measured levels. Traditionally, the Manual Strapping Method (MSM) has been the established, high-accuracy technique for creating these tables [3,4]. However, the MSM is time-consuming, labor-intensive, and presents safety risks due to the need for manual, physical measurements on tank shells [5].

As an alternative, Electro-Optical Distance Ranging (EODR) techniques, such as 3D laser scanning and triangulation, have emerged to provide faster and more efficient data collection [6,7,8]. While EODR offers high-speed, non-contact data acquisition, it can sometimes be susceptible to measurement noise or less accurate in capturing the full, complex physical deformities of older, massive tanks compared to the traditional manual method [9,10].

Recent advancements in industrial metrology and "Industry 4.0" practices emphasize the integration of Machine Learning (ML) to enhance measurement reliability [11,12]. Machine learning models, particularly ensemble methods like XGBoost (Extreme Gradient Boosting), have shown superior performance in prediction and regression tasks, outperforming traditional algorithms, especially when dealing with complex, non-linear relationships in data [13].

Therefore, there is a need for a data-driven approach that leverages the speed of EODR while guaranteeing the accuracy of the manual method.

This study aims to bridge this gap by employing an XGBoost machine learning model to train on EODR data and "ground-truth" it against the traditional manual strapping measurements. By training a regression model on paired EODR/MSM datasets, this research seeks to improve the accuracy of EODR-based calibration, culminating in an efficient, intelligent system that reduces the errors often associated with solely automated or manual methods.

2. Methodology

This study utilizes an XGBoost machine learning model to improve the accuracy of Electro-Optical Distance Ranging (EODR) crude oil storage tank calibration by training it against ground-truth Manual Strapping Method (MSM) data. The methodology involves acquiring paired EODR/MSM datasets, preprocessing data for consistency, and training a regression model to predict high-accuracy MSM volumes from easy-to-acquire EODR measurements. The XGBoost Model Architecture is as shown in Figure 1.

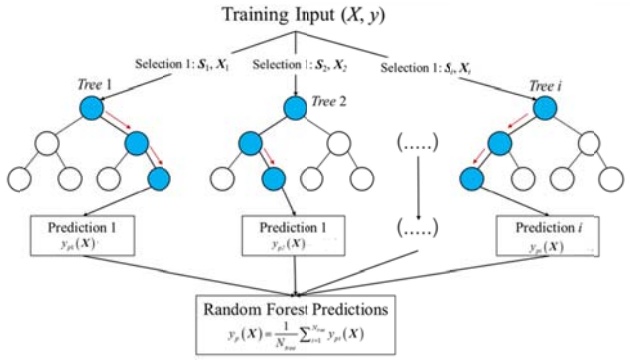

Figure The XGBoost Model Architecture [14]

2.1 Data Acquisition and Description

The dataset used for this research was obtained by carrying out calibration procedures on industrial crude oil storage tanks, utilizing two distinct methods: the traditional Manual Strapping Method (MSM) and the automated Electro-Optical Distance Ranging (EODR) method. In the Manual Strapping Method (MSM) data (which is the ground truth data) the physical measurements of the tank circumference, shell course heights, and plate thicknesses were taken manually using calibrated steel tapes. This data was processed to generate a tank capacity chart, defining the precise volume of oil at specific depth intervals. Due to its high accuracy, the MSM dataset is adopted as the ground truth. In the Electro-Optical Distance Ranging (EODR) data, a Total Station (electronic survey device) was used to internally scan at least eight or more points around the circumference of each course of the tank to determine the internal diameters. This electronic approach is fast and provides a detailed dataset, although it is assumed to have slightly different accuracy levels compared to the manual method.

The acquired data includes paired records of tank levels/depths, EODR measured volume, and corresponding MSM measured volume, creating a mapping of EODR-measured volumes to ground truth volumes. As such, the consolidated dataset for the model training consists of these experimental records gathered from industrial vertical cylindrical storage tanks, focusing on the relationship between tank depth and volume. The input feature (or the independent variable) consist of the Tank Liquid Depth (m) and EODR-calculated Volume (barrels/liters). The target label (or the dependent variable is the MSM-calculated Volume (barrels/liters), representing the ground truth data. The dataset consists of high-density calibration points mapping volume to liquid levels up to the maximum capacity of the tanks.

2.2 Data Pre-processing

Data preprocessing was carried out to mitigate acquisition errors and enhance XGBoost learning, featuring the following steps:

i. **Data Cleaning:** The raw datasets from both MSM and EODR were inspected for inconsistencies, and erroneous readings caused by environmental factors or human error during manual measurement were removed.

ii. **Missing Value Imputation/Removal:** Observations with missing values in either the EODR or MSM records were removed to maintain data integrity.

iii. **Data Alignment (Alignment of Volume-Height Data):** Since EODR and MSM might not have been taken at identical, granular depths, interpolation techniques (e.g., linear interpolation) were used to match the EODR measured volumes to the precise corresponding depths in the MSM table, creating a 1:1 mapping for training.

iv. **Feature Selection:** The input features for the model were finalized to include Tank Level and EODR Volume to predict the target label (MSM Volume).

v. **Normalization/Scaling:** While XGBoost is tree-based and does not strictly require normalization, standard scaling was applied to ensure that the volume and depth parameters (which have different units and scales) were treated appropriately to improve training stability.

vi. **Dataset Splitting:** The refined dataset was split into training and testing sets (e.g., 80% for training and 20% for testing) to evaluate the model's predictive ability on unseen data.

2.3 XGBoost Model Formulation

The XGBoost (Extreme Gradient Boosting) is based on an ensemble of decision trees and uses a gradient boosting framework to minimize the objective function. The model iteratively bu,ilds trees to correct the residual errors of the previous trees. Consider a training dataset, D , which is defined analytically as follows;

$$\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{n}} \quad \textbf{(1)}$$

Where, x_i is the input feature, and y_i is the output feature. The XGBoost builds trees on D using the objective function:

$$\mathcal{L}(\theta) = \sum_{i=2}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \quad (2)$$

Where, $l(y_i, \hat{y}_i)$ is the loss function (MSE)

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad (3)$$

Where $\Omega(f_t)$ is the regularization term that controls model complexity:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum \omega_j^2 \quad (4)$$

Where, γ is the is the regularization parameter for the number of trees T, and λ is the regularization parameter for leaf weights ω_j. At each iteration, the model improves the prediction by adding a new tree, $f_t(x)$, which minimizes the loss function as follows;

$$\tilde{g}_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i} \quad (5)$$

$$\tilde{h}_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^2} \quad (6)$$

$$f_t(x) = -\frac{\sum_i \tilde{g}_i}{\sum_i \tilde{h}_i + \lambda} \quad (7)$$

The hyperparameters used in the implementation of the XGBoost (Extreme Gradient Boosting) model are presented in Table 1.

Table 1 The hyperparameters of the XGBoost (Extreme Gradient Boosting) Model

S/N	Hyperparameter	Value
1	Number of trees (n estimator)	100
2	Learning rate	0.1
3	Max_depth	6
4	Min_child_weight	1
5	Sub_sample	1.0
6	Gamma	0
7	Random state	42

3. Results and discussion

The XGBoost (Extreme Gradient Boosting) model performance is visualized in Figures 2–7, which include the confusion matrix, residual plot, volume versus tank depth plot, and metrics (MSE, RMSE, MAE, R^2) versus epoch, respectively. Detailed error metrics over epochs are provided in Table 2, while the steady-state results are summarized in Table 3. According to Table 3, the model achieved a 0.929% accuracy (1-MAPE), with an MSE of 0.0910, RMSE of 0.3016, and MAE of 0.2474. Analysis of the errors showed a minimum of -0.036 at a tank depth of 6290.0 and a maximum of 0.5898 at a tank depth of 5310.0.

Also, as shown in Table 2, Figure 6 and Figure 7, the XGBoost model demonstrates progressive optimization, culminating in superior performance at Epoch 5. While epochs 2-4 show minor fluctuations in error (MSE/RMSE/MAE) and a slow, steady increase in R^2, the final epoch shows a significant improvement in accuracy, reducing MSE by $\approx 40\%$ compared to the, suggesting efficient convergence.

In the initial phases (Epochs 1-4): The model shows instability, with errors increasing slightly at Epoch 3 (0.1614 MSE) before dropping, indicating a need for more iterations to stabilize, a common trait in boosting as it corrects previous errors. In the final optimal phase (Epoch 5) a significant reduction in error metrics (MSE: 0.0910, RMSE: 0; 3017 and MAE: 0.2474) occurs here, indicating the model has effectively captured the underlying data patterns. The R^2 value improves from 0.96751 to 0,9683, indicating strong, consistent predictive power throughout the training process. The MAPE (%) shows the most consistent downward trend, dropping from 0.1541% to 0.0710%. This suggests the model became significantly more precise regarding the relative percentage of its errors as training progressed.

In all, the model successfully converged by the 5th epoch. The sharp improvement between Epoch 4 and 5 suggests that the model likely found a better local minimum or settled into a more optimal weight configuration toward the end of this run.

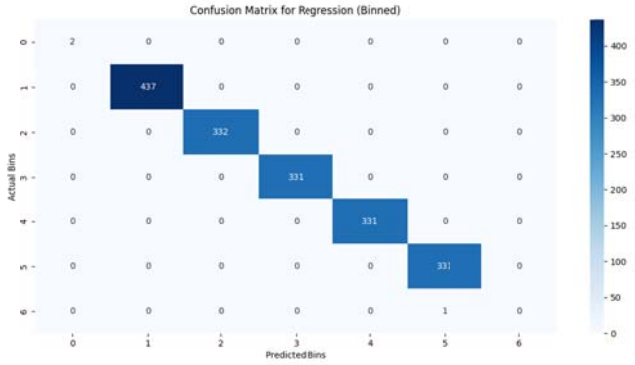

Figure 2 The Confusion matrix for XGBoost **model**

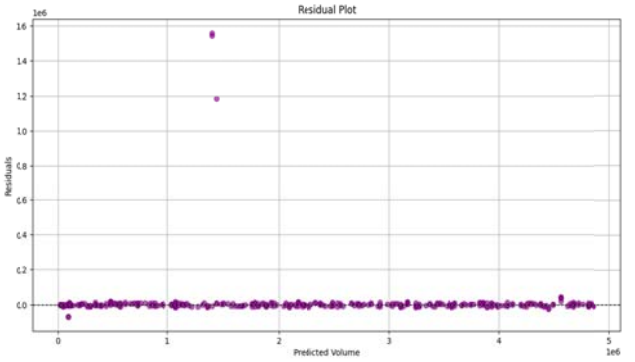

Figure 3: Residual plot for XGBoost **model**

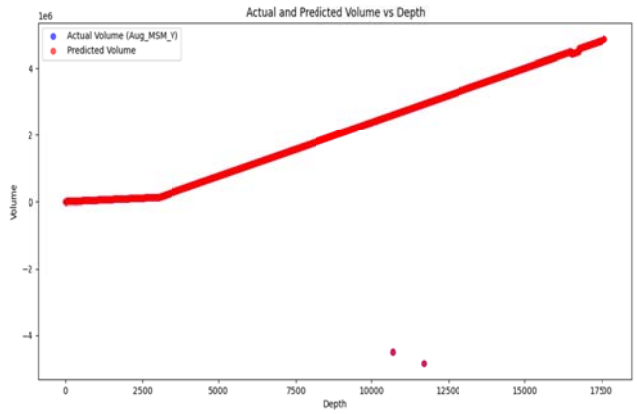

Figure 4: Volume vs Tank Depth for XGBoost **model**

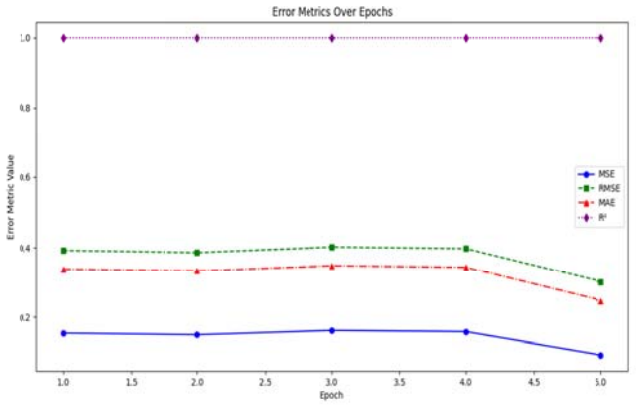

Figure 5: Error metrics for XGBoost **model**

Table 2: Error metrics score versus epoch for random XGBoost model

Score					
Epoch	MSE	RMSE	MAE	R^2	MAPE (%)
1	0.1536	0.3919	0.3383	0.96751	0.1541%
2	0.1489	0.3859	0.3330	0.968	0.10301%
3	0.1614	0.4017	0.3481	0.9681	0.097%
4	0.1579	0.3973	0.3435	0.9682	0.0881%
5	0.0910	0.3017	0.2474	0.9683	0.0710%

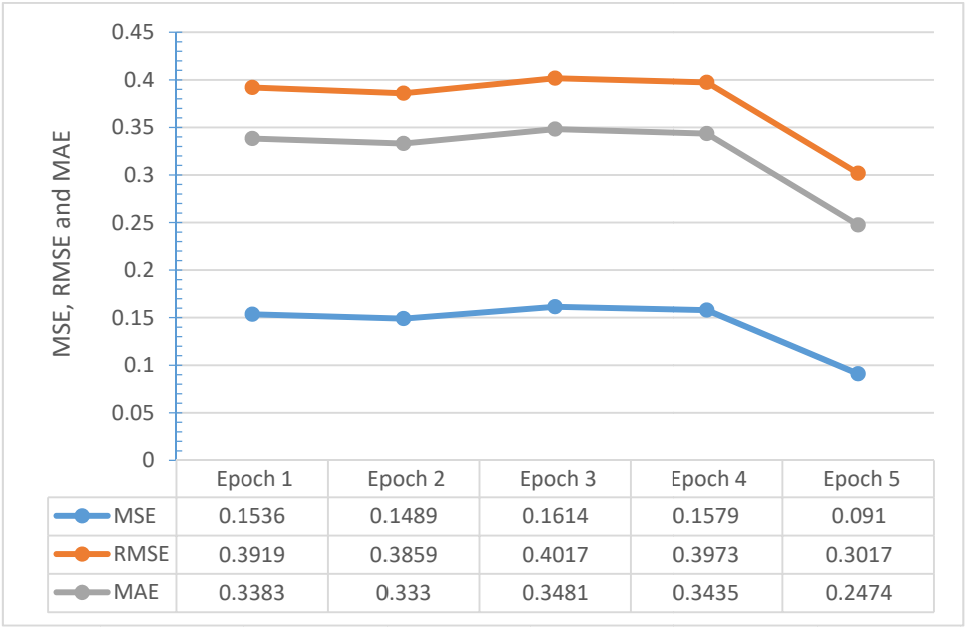

Figure 6 MSE, RMSE and MAE versus epoch

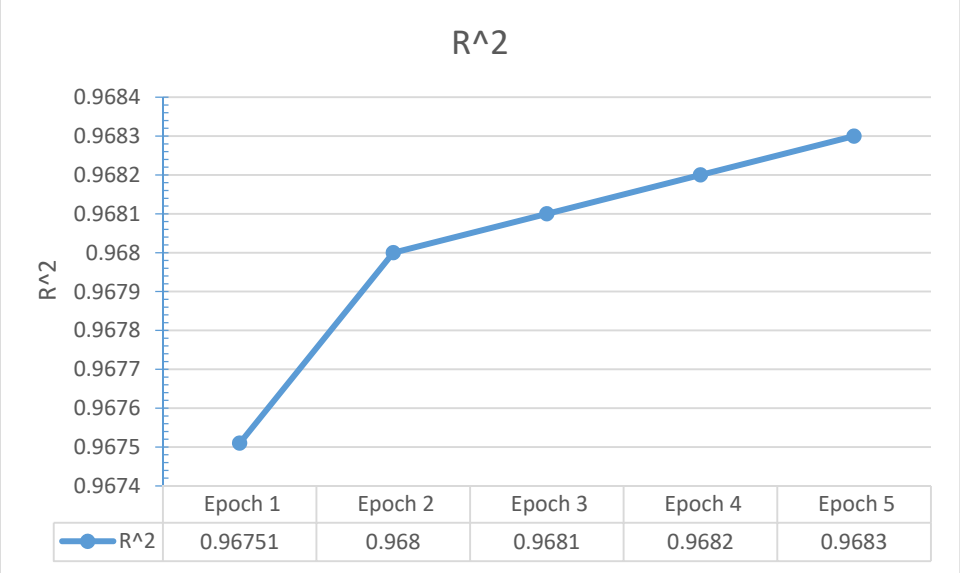

Figure 7 R^2 versus epoch

Table 3 Error metrics at the end of Epoch for XGBoost

Metric	Score XGBoost
MSE	0.0910
RMSE	0.3016
MAE	0.2474
R^2.	0.9683
1-MAPE	0.929%
Percentage error	-0.046%
Min Error	-0.036 at Depth 6290.0
Max Error	0.5898 at Depth 5310.0

4. Conclusion

This research introduces a validated XGBoost approach to enhance EODR precision in tank calibration, using MSM data as ground truth. The results reveal the model's ability to map intricate, nonlinear patterns, yielding strong predictive accuracy (consistent R^2 of 0.9683).

Key findings indicate that the model achieves optimal performance by the 5th epoch, showing efficient convergence and superior accuracy. Specifically, the final model, when compared to initial training phases, significantly reduces the error metrics (MSE, RMSE, MAE) and achieves a 40% reduction in mean squared error, indicating high precision. The steady, downward trend in the Mean Absolute Percentage Error (MAPE) also confirms the model's reliability in reducing measurement errors. Furthermore, the sharp performance improvement observed between epochs 4 and 5 indicates that the model effectively minimizes the residuals of previous trees, correcting for initial instabilities. Ultimately, this study provides a robust, efficient, and cost-effective machine-learning-based tool for tank calibration, offering a superior alternative to traditional, time-intensive methods.

Future research could focus on expanding the dataset to include varying tank shapes and sizes to further improve the model's generalization capabilities, and integrating real-time EODR monitoring for dynamic tank capacity management.

References

1. Agboola, O. O., Akinnuli, O. B., Akintunde, A. M., & Kareem, B. (2020). Modelling of cost estimates for the geometrical calibration of upright oil storage tanks. *International Journal of Energy Economics and Policy*, *10*(1), 464-470.

2. Shil, S. K. (2024). Petroleum Storage Tank Design and Inspection Using Finite Element Analysis Model For Ensuring Safety Reliability And Sustainability. *Review of Applied Science and Technology*, *3*(04), 94-127.

3. Wang, P., Zhao, H., & Ren, G. (2022). Development and application of standard device for calibrating steel measuring tape based on machine vision. *Applied Sciences*, *12*(14), 7262.

4. Kyriakou, T., de la Campa Crespo, M. Á., Panayiotou, A., Chrysanthou, Y., Charalambous, P., & Aristidou, A. (2024, May). Virtual instrument performances (vip): A comprehensive review. In *Computer Graphics Forum* (Vol. 43, No. 2, p. e15065).

5. Lautre, N. K. (2024). Current Scenario, Future Scope, and Challenges in Welding. In *Advanced Welding Techniques* (pp. 211-231). CRC Press.

6. Hao, H., Shi, H., Yi, P., Liu, Y., Li, C., & Li, S. (2018, January). Research on volume metrology method of large vertical energy storage tank based on internal electro-optical distance-ranging method. In *2017 International Conference on Optical Instruments and Technology: Optoelectronic Measurement Technology and Systems* (Vol. 10621, pp. 562-567). SPIE.

7. Agboola, O. O., Akinnuli, B. O., Akintunde, M. A., Ikubanni, P. P., & Adeleke, A. A. (2019, December). Comparative analysis of manual

strapping method (MSM) and electro-optical distance ranging (EODR) method of tank calibration. In *Journal of Physics: Conference Series* (Vol. 1378, No. 2, p. 022062). IOP Publishing.

8. Agboola, O. O., Ikubanni, P. P., Ibikunle, R. A., Adediran, A. A., & Ogunsemi, B. T. (2017). Generation of calibration charts for horizontal petroleum storage tanks using microsoft excel. *MAPAN*, *32*(4), 321-327.

9. Hassani, S., & Dackermann, U. (2023). A systematic review of advanced sensor technologies for non-destructive testing and structural health monitoring. *Sensors*, *23*(4), 2204.

10. Sony, S., Laventure, S., & Sadhu, A. Barbosa, C. R. H., Sousa, M. C., Almeida, M. F. L., & Calili, R. F. (2022). Smart manufacturing and digitalization of metrology: a systematic literature review and a research agenda. *Sensors*, *22*(16), 6114.*Structural Control and Health Monitoring*, *26*(3), e2321.

11. Yadav, S., Rab, S., & Wan, M. (2023). Metrology and sustainability in Industry 6.0: Navigating a new paradigm. In *Handbook of quality system, accreditation and conformity assessment* (pp. 1-31). Singapore: Springer Nature Singapore.

12. Adeleke, A. K., Ani, E. C., Olu-lawal, K. A., Olajiga, O. K., & Montero, D. J. P. (2024). Future of precision manufacturing: Integrating advanced metrology and intelligent monitoring for process optimization. *International Journal of Science and Research Archive*, *11*(1), 2346-2355.

13. Gelete, G. (2023). Hybrid extreme gradient boosting and nonlinear ensemble models for suspended sediment load prediction in an agricultural catchment. *Water Resources Management*, *37*(14), 5759-5787.

14. Phulsawat, B., Senjuntichai, A., & Senjuntichai, T. (2024). Prediction of multi-layered pavement moduli based on falling weight deflectometer test using soft computing approaches. *Transportation Infrastructure Geotechnology*, *11*(4), 2348-2381.