Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 12 Issue 12, December - 2025

Crude Oil Storage Tank Volume Calibration Data Prediction Using Random Forest Regression Model

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*—In this work, crude oil storage tank volume prediction using Random Forest Regression (RFR) model is presented. Two set of the tank volume datasets acquired through the Manual Strapping Method (MSM) and also through the Electro Optical Distance Ranging (EODR) method are matched by the tank depth. About 30 of such paired data items in the case study datasets are then used as input for data augmentation where the resulting augmented dataset has a total of 1560 paired data records. The augmented dataset is used for the training and validation of the RFR model for prediction of the crude oil storage tank volume that is equivalent to the MSM measured volume for any given EODR measured crude oil storage tank volume and tank depth. The RFR model prediction performance results show Mean Square Error (MSE) of 0.1588 bbls, Root Mean Square Error (RMSE) of 0.3986 bbls, Mean Absolute Error (MAE) of 0.3434 bbls and prediction accuracy of 0.9999%. Also, the RFR model has minimum absolute prediction error of 0.0032 bbls at tank depth of 8210.0 mm and the maximum absolute error of 0.6898 bbls at tank depth of 5310.0 mm. The ideas presented in this work are useful for the oil industry where alternative to the manual calibration is highly preferred.**

***Keywords*—Crude Oil Storage Tank Volume Calibration, Random Forest Regression Model, Ground Truthing, Manual Strapping Method, Electro Optical Distance Ranging (EODR) method**

1. Introduction

In the petrochemical industry, the precise calibration of crude oil storage tanks is vital for managing inventory, facilitating custody transfers, and tracking product loss [1,2]. Because calibration accuracy directly affects financial reporting and operational success, tanks must be calibrated before use and re-evaluated every five years to meet international standards [3,4]. Historically, the Manual Strapping Method (MSM) has served as the standard industry practice for measuring tank capacity [5].

Despite being recognized as highly accurate, the MSM method involves physically measuring the tank circumference, which is notoriously tedious, time-consuming, and labor-intensive [6]. Furthermore, the manual approach presents significant safety hazards to personnel, who must operate in confined spaces, and carries inherent risks of misreading the strapping tape or manual errors, making it increasingly discouraged in modern, automated, and high-efficiency operations [7].

Conversely, Electro-Optical Distance Ranging (EODR) offers a modern, high-speed alternative, using electronic sensors and laser technology to generate a 3D model of the tank, significantly reducing both measurement time and human safety risks [8,9,10]. While EODR is efficient and generally meets the required accuracy standards, it is often viewed as slightly less accurate in direct comparison to the "ground truth" provided by rigorous manual methods [11].

Accordingly, this study introduces a Random Forest Regression (RFR) model to combine the speed of EODR with the precision of the traditional MSM method [12,13]. By training on historical, paired datasets, the RFR model learns to accurately predict MSM volumes directly from EODR data. This automated, predictive approach removes the need for manual, time-consuming calibration, offering a high-

precision, efficient, and reliable solution for tank volume estimation.

2. Methodology

This study employs a Random Forest Regression (RFR) model to predict crude oil storage tank volume calibration data typically obtained via the labor-intensive Manual Strapping Method (MSM), using data acquired through the more efficient Electro Optical Distance Ranging (EODR) method **[15].** The MSM data is treated as the "ground truth" due to its accepted high accuracy, while the EODR data serves as the input features for the predictive model. The overarching goal is to eliminate the need for manual calibration by leveraging the easily acquired EODR measurements to accurately generate the corresponding MSM volume data.

The research procedure is structured into three primary sections: data acquisition, data pre-processing, and the training and evaluation of the Random Forest Regression model.

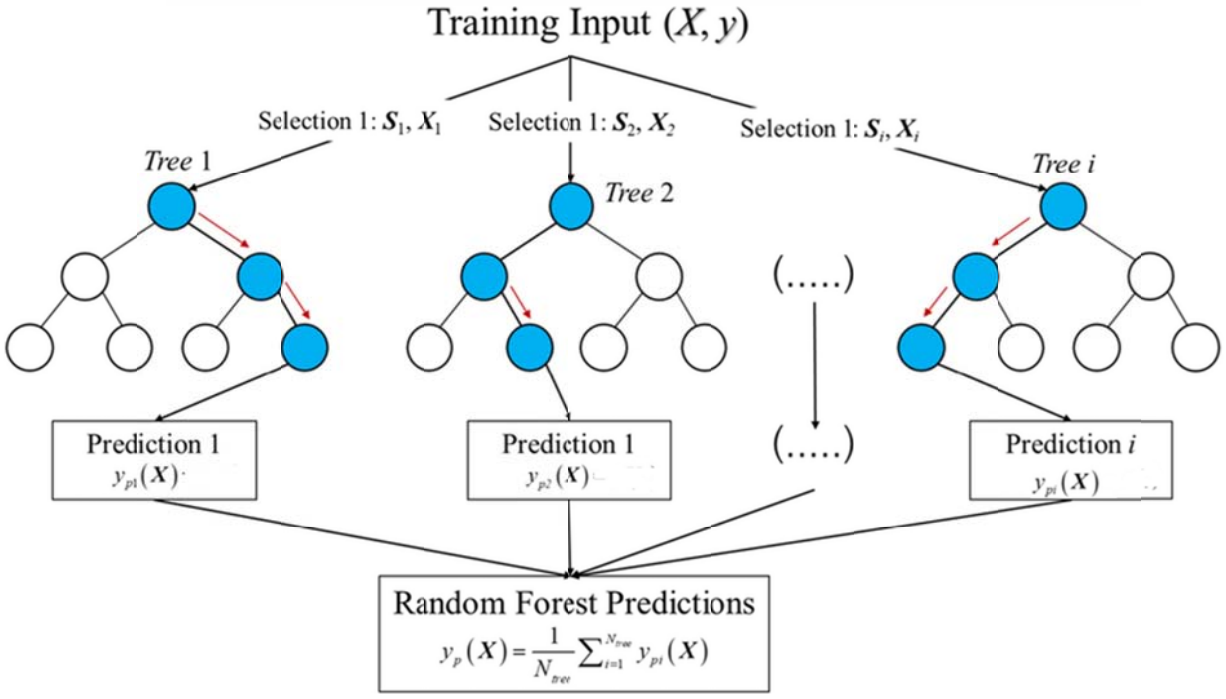

Figure 1 The architecture of the Random Forest Model

2.1 Acquisition of the Case Study Dataset [14]

Case study datasets were acquired utilizing two distinct methodologies; the Manual Strapping Method (MSM) and the Electro Optical Distance Ranging (EODR) Method. The Manual Strapping Method (MSM) is the traditional method which is considered highly accurate method that involves physical, manual measurements of the tank's dimensions and is considered the ground truth dataset **[15].** This process is acknowledged as tedious and time-consuming. On the other hand, the Electro Optical Distance Ranging (EODR) Method is the modern approach which employs electronic sensors and devices to gather data efficiently **[5].** This dataset is assumed to be less accurate than the MSM dataset but is far simpler and faster to acquire in practical applications. In this research, the datasets comprises of paired measurements of tank depth (input feature X) and the corresponding crude oil volume (output feature Y) as determined by each method.

2.2 Pre-processing of the Case Study Dataset

The raw datasets from both acquisition methods are subjected to a number of pre-processing steps to ensure data quality and suitability for the RFR model. The pre-processing steps implemented are as follows:

Step 1: Data Cleaning: Identification and handling of missing values, outliers, or erroneous entries within both EODR and MSM datasets.

Step 2: Data Alignment: Ensuring that the EODR depth and volume measurements are precisely aligned with the corresponding MSM ground truth measurements for the same tank depths.

Step 3: Feature Selection: The EODR measured volume and associated tank depth will be designated as input features (X) for the model.

Step 4: Data Splitting: The combined, pre-processed dataset will be randomly partitioned into two subsets: a training set (e.g., 80%) used to build the model, and a testing set (e.g., 20%) reserved for evaluating the model's performance on unseen data.

2.3 Development, Training, and Evaluation of the Random Forest Regression (RFR) Model

The Random Forest Regression model (Figure 1) is meant to model the non-linear relationship between the tank depth and volume data **[16].** The model aims

to read in the EODR measured volume at any given tank depth and predict the expected volume value that the MSM dataset would have provided at that exact depth. Specifically, the Random Forest Regression (RFR) characteries the relationship between depth (input feature$X$) and volume (output feature Y). Given a dataset:

$$\boldsymbol{D} = \{(\boldsymbol{X_1}, \boldsymbol{Y_1}), (\boldsymbol{X_2}, \boldsymbol{Y_2}), .., (\boldsymbol{X_n}, \boldsymbol{Y_n})\} \textbf{(1)}$$

Where, X represents depth in mm , and Y represents volume in $bbls$. The Random Forest Regression consists of multiple Decision Trees, where each tree is trained on a random subset of the dataset. The final prediction is obtained by averaging the outputs of all the trees. Each tree T_i is trained on a randomly sampled subset of data:

$$\boldsymbol{D_i} = \{(\boldsymbol{X_{i1}}, \boldsymbol{Y_{i1}}), (\boldsymbol{X_{i2}}, \boldsymbol{Y_{i2}}), .., (\boldsymbol{X_{im}}, \boldsymbol{Y_{im}})\}, \boldsymbol{where\ m} < \boldsymbol{n} \textbf{(2)}$$

This ensures diversity in training and reduces overfitting. Each Decision Tree is built by recursively splitting the dataset based on the feature X using a splitting criterion. The splitting criterion for regression trees is to minimize the Mean Squared Error (MSE) at each node:

$$\boldsymbol{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{Y_i} - \widehat{\boldsymbol{Y}}\right)^2 \textbf{(3)}$$

Where $\hat{Y}$ is the mean target value of samples in the node. Each split is chosen to minimize the weighted sum of MSE for the left and right child nodes:

$$\boldsymbol{MSE_{split}} = \frac{N_L}{N}\boldsymbol{MSE_L} + \frac{N_R}{N}\boldsymbol{MSE_R} \textbf{(4)}$$

Where, N_L and N_R denote number of samples in the left and right nodes, respectively, then, MSE_L and MSE_R denote the MSE values for the left and right child nodes.

The tree grows until a stopping condition is met, which include: Minimum number of samples in a leaf node or Maximum tree depth. Once all trees T_1, T_2, ., T_B (where B is the number of trees) are trained, the final prediction for a new input X' is obtained by averaging the outputs of all trees.

$$\widehat{\boldsymbol{Y}} = \frac{1}{B}\sum_{i=1}^{B}\boldsymbol{T_i}(\boldsymbol{X'}) \textbf{(5)}$$

Where, $T_i(X')$ is the predicted value for the i^{th} tree. The hyperparameters used in the implementation of the Random Forest Regression (RFR) model are presented in Table 1.

Table 1 The hyperparameters of the Random Forest Regression (RFR) **Model**

Hyperparameter	Value
Number of trees	100
Max-depth	None (decided by the model)
Min samples per leaf	1
N estimator	100
Random state	42

2.3.1 Model Training

The RFR model is trained on the designated training dataset. The training process involves building an ensemble of decision trees, where each tree is constructed using a random subset of the training data and features. The final prediction from the RFR model is the average of the predictions from all individual trees. The objective function during training will be to minimize the difference between the model's predictions and the actual MSM ground truth values in the training set.

2.3.2 Model Evaluation

The performance of the trained RFR model is rigorously evaluated using the held-out testing dataset. Key regression metrics are employed to quantify the model's accuracy and predictive capability are presented in Table 2.

Table 2 The regression metrics employed to quantify the model's accuracy and predictive capability

Metric Name	Abbreviation	Formula
Mean Squared Error	MSE	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
Root Mean Squared Error	RMSE	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$
Mean Absolute Error	MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$

R-squared (Coefficient of Determination)	R^2	$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Mean Absolute Percentage Error	MAPE	$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left\|\frac{y_t - \hat{y}_t}{y_t}\right\| \times 100\%$
Percentage Error	PE	$\frac{y_i - \hat{y}_i}{y_i} \times 100\%$
Min Error	MinE	$\min_{i=1}^{n}(y_i - \hat{y}_i)$
Max Error	MaxE	$\max_{i=1}^{n}(y_i - \hat{y}_i)$

Where
n: Number of data points/samples
y_i: Actual value
$\hat{y}_i$: Predicted (forecasted) value
$\bar{y}$: The mean (average) of all actual y values

3. Results and discussion

3.1 The results of Random Forest Regression (RFR) Model

The Random Forest Regression (RFR) model was developed to predict the cumulative volume of crude oil based on tank dip levels (depth). The performance of the model was evaluated using a testing dataset. The results for the Random Forest Regression (RFR) model is presented in Figure 2 for the confusion matrix, Figure 3 for residual plot, Figure 4 for the volume versus tank depth plot, Figure 5 for the plot of MSE, RMSE, MAE and R^2 versus epoch, Table 3 for the error metric scores versus epoch and Table 4 for the overall or steady state metric scores of the RFR model.

The high R^2 value of 0.978 indicates that approximately 97.8% of the variance in the tank volume data is explained by the model, demonstrating excellent goodness-of-fit. The low MAPE (0.031%) and low MAE (0.3434) highlight the model's high precision in forecasting volume, with minimal deviation from actual calibration chart values.

Analysis of prediction errors showed the model performed consistently across different tank depths. The minimum error recorded was 0.0032 at a depth of 8210.0, while the maximum error was 0.6898 at a depth of 5310.0. The very low average percentage error (-0.029%) indicates that the model has negligible bias, neither consistently overestimating nor underestimating the volume significantly.

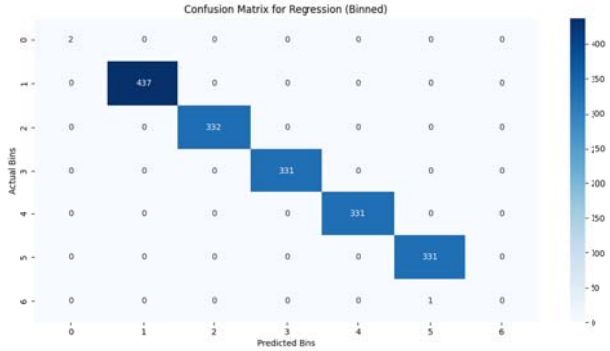

Figure 2 The Confusion matrix for RFR model

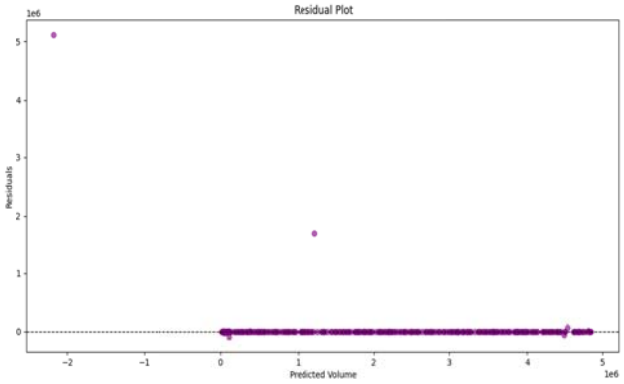

Figure 3: Residual plot for **RFR model**

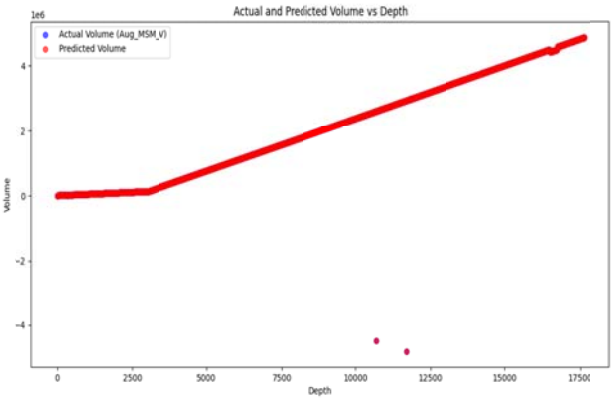

Figure 4: Volume vs Tank Depth for **RFR model**

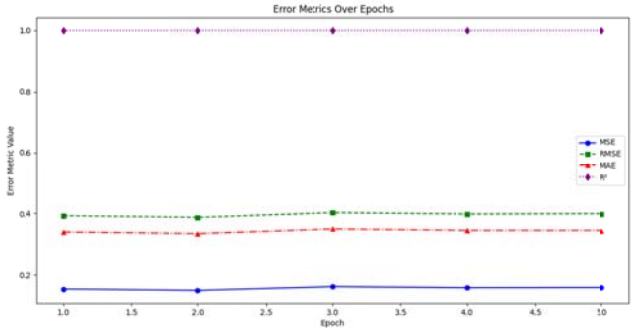

Figure 5: Error metrics for **RFR model**

Table 3: Error metrics score versus epoch for random RFR model

Epoch	Score					
	MSE	RMSE	MAE	R^2	MAPE (%)	% error
1	0.1536	0.3919	0.3384	1.0000	0.0001%	0.0001%
2	0.1493	0.3863	0.3330	1.0000	0.0001%	0.0001%
3	0.1616	0.4020	0.3482	1.0000	0.0001%	-0.0001%
4	0.1580	0.3974	0.3435	1.0000	0.0001%	0.0001%
5	0.1588	0.3986	0.3434	1.0000	0.0001%	-0.0001%

Table 4 Error metrics at the end of Epoch for RFR model

Metric	Value obtained by the Random Forest Regression (RFR) model
MSE	0.1588
RMSE	0.3986
MAE	0.3434
R^2	0.978
MAPE	0.031%
Percentage error	-0.029%
Min Error	0.0032 at Depth 8210.0
Max Error	0.6898 at Depth 5310.0

3.2 Discussion

The results demonstrate that the Random Forest Regression (RFR) model is a highly effective tool for automating the prediction of crude oil storage tank calibration data. The high accuracy, reflected in the R^2 of 0.978 and a remarkably low MAPE of 0.031%, confirms that the ensemble learning approach successfully captured the non-linear relationship between tank depth and volume.

The RFR model's superior performance can be attributed to its ability to handle non-linear data and reduce overfitting, common issues in tank calibration, where irregular tank deformation or sensor inaccuracies can create noise in the data. The low RMSE (0.3986) in comparison to the scale of total volume indicates that the model's predictions are highly reliable for inventory accounting.

While the maximum error was 0.6898 (found at depth 5310.0), it is important to note that this error is relatively small, representing a negligible fraction of the total capacity of a typical crude oil storage tank. The fact that the maximum error is quite low, combined with the low overall MAPE, suggests the model is robust and not significantly affected by outliers at specific, complex, or partially full, parts of the tank.

The percentage error of -0.029% indicates a negligible underestimation of volume on average, which is highly favorable for safety-critical stock accounting, ensuring accurate inventory management and minimizing product loss tracking discrepancies. Compared to traditional methods or simpler regression, the RFR model provides superior accuracy by identifying patterns, which is critical for non-uniform, large storage tanks.

4. Conclusion

Application of Random Forest Regression (RFR) model to automate crude oil storage tank volume calibration is presented. By bridging the gap between the labor-intensive Manual Strapping Method (MSM) and the high-speed Electro-Optical Distance Ranging (EODR) technique, the study achieved its

primary objective, notably, eliminating the need for manual, physical calibration without compromising precision.

The application of the Random Forest Regression (RFR) model for predicting crude oil storage tank volume calibration data, using EODR inputs as predictors and MSM data as ground truth, has proven to be highly effective and accurate. With a high coefficient of determination and low error metrics, the model demonstrates an exceptional ability to map the non-linear relationship between EODR-measured tank depths and cumulative volume. The results, as illustrated in the plotted performance graphs and summarized in the evaluation tables, indicate that the model achieves high precision with minimal deviation from the actual manual strapping method (MSM) calibration chart values.

In any case, while this study establishes a robust foundation for ML-driven calibration, future work should focus on:

i. Dynamic and Real-Time Monitoring: Integrating real-time sensor streams to account for environmental factors like thermal expansion (governed by ISO 8222) and pressure variations.

ii. Structural Defect Identification: Leveraging RFR or deep learning to automatically detect and compensate for tank tilt, longitudinal deflection, or "dead volume" (internal structures) as defined in ISO 7507-1.

iii. Cross-Geometry Generalization: Extending the model's application to complex geometries, such as horizontal cylindrical tanks (API 2.2F/ISO 12917) and spherical vessels.

iv. Digital Metrology Frameworks: Aligning predictive models with Industry 4.0 "digital twins" to create fully automated, standards-compliant metrological systems for global petroleum asset management.

In all, this research concludes that the RFR model serves as a robust alternative to conventional calibration, offering the petroleum industry a reliable, efficient, and accurate method to generate tank capacity tables without relying solely on labor-intensive, time-consuming manual processes. In addition, the RFR model provides a high-fidelity, efficient alternative to traditional calibration, reducing safety risks and labor costs while ensuring the precision required for modern petroleum logistics.

References

1. Shunashu, I. L., & Casmir, R. (2020). Assessing the impact of measurement uncertainty in custody transfer to the development of oil & gas industry in Tanzania. *Business Education Journal*, *6*(2).

2. Daher, E., & Schoeib, S. (2024, September). Revolutionizing Tool Management in Oil and Gas Facilities. In *SPE International Conference and Exhibition on Health, Safety, Environment, and Sustainability?* (p. D031S032R006). SPE.

3. World Health Organization. (2022). *WHO operational handbook on tuberculosis. Module 3: diagnosis. Tests for TB infection*. World Health Organization.

4. WHO Expert Committee on Biological Standardization. Meeting, & World Health Organization. (2007). *WHO expert committee on biological standardization: fifty-sixth report* (Vol. 941). World Health Organization.

5. 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.

6. Chen, G., Wan, Y., Lin, H., Hu, H., Liu, G., & Peng, Y. (2021). Vertical tank capacity measurement based on Monte Carlo method. *PLoS One*, *16*(4), e0250207.

7. OLUWASEYI, A. A., ENOBONG, H., CHUKWUEBUKA, N., & ANDREW, E. E. (2024). Improving worker safety in confined space entry and hot work operations: Best practices for high-risk industries. *GLOBAL JOURNAL OF ADVANCED RESEARCH AND REVIEWS Учредители: Global Scholar Publications*, *2*(2), 031-039.

8. National Research Council, Division on Engineering, Physical Sciences, & Committee on Review of Advancements in Active Electro-Optical Systems to Avoid Technological Surprise Adverse to US National Security. (2014). *Laser radar: progress and opportunities in active electro-optical sensing*. National Academies Press.

9. Cavanaugh, R. (2024). *Electro-Optical and Infrared Design for Uncrewed Aerial System Collision Avoidance* (Master's thesis, The University of Arizona).

10. Wang, J. T., Liu, Z. Y., Zhang, L., Guo, L. G., Bao, X. S., & Tong, L. (2010). Automatic measurement system for vertical tank volume by electro-optical distance-ranging method. *Applied Mechanics and Materials*, *26*, 416-421.

11. Yesudasu, S. (2024). *Enhancing Logistics Automation with AI: Application of Dual-arm Humanoid Torso for AI-powered Depalletizingand Package Handling* (Doctoral dissertation, Normandie Université).

12. Lukita, C., Lutfiani, N., Panjaitan, A. R. S., Rahardja, U., & Huzaifah, M. L. (2023, December). Harnessing the power of random forest in predicting startup partnership success. In *2023 eighth international conference on informatics and computing (ICIC)* (pp. 1-6). IEEE.

13. Krzemińska, A., Miller, T., Kozlovska, P., & Lewita, K. (2023). Harnessing the power of random forest machine learning in global agriculture innovation. *Collection of scientific papers «SCIENTIA»*, (July 28, 2023; Tel Aviv, Israel), 59-65.

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.

15. Firmansyah, V., Nugroho, P., Prihensa, H. Y., & Muslim, A. (2020). Comparison Study of Vertical Cylinder Tank Diameter Measurement Between Strapping and Optical Method. *Spektra: Jurnal Fisika dan Aplikasinya*, *5*(3), 231-238.

16. Chen, G., Wan, Y., Lin, H., Hu, H., Liu, G., & Peng, Y. (2021). Vertical tank capacity measurement based on Monte Carlo method. *PLoS One*, *16*(4), e0250207.