Science and Technology Publishing (SCI & TECH)
ISSN: 2632-1017
Vol. 9 Issue 12, December - 2025

Data Augmentation Of Crude Oil Storage Tank Volume Calibration Datasets Using Gaussian Noise Method

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 study, data augmentation of crude oil storage tank volume calibration datasets using Gaussian noise method is presented. The essence of the work is to address the problem of scarcity of required data for machine learning model training and validation regarding crude oil storage tank calibration. As such, data augmentation is performed whereby additional data records are generate to augment the available empirical dataset which is grossly inadequate for application in machine learning model training. The case study crude oil storage tank calibration dataset is obtained through the Manual Strapping Method (MSM) method and it has about 30 data points. The study generated additional data points that sum up to 1500 data point in the augmented dataset. The mean of original dataset with 30 data points is 1,321,851.5, bbls, the standard deviation (S) is 796.62 bbls and the Interquartile range is 2,406,419. The synthetically generated dataset using the Gaussian noise method has mean of 1,321,875.2, bbls, standard deviation (S) of 802.03 bbls and Interquartile range of 2,406,537. The confidence interval for the generated dataset at 95% confidence level shows that there is no significant difference between the mean of the original dataset and the synthetically augmented dataset. Hence, the augmented dataset is a good replica of the original; dataset and it can be used for machine learning model training instead of the original dataset which is grossly inadequate for machine learning model training.**

Keywords—Crude Oil Storage Tank Calibration, Data Augmentation, Gaussian Noise Method, Machine Learning Model, Synthetically Generated Dataset

1. Introduction

Precise volume calibration of crude oil storage tanks is fundamental to inventory management, custody transfer, and loss control within the petroleum industry [1,2,3]. Traditionally, this process relies on empirical measurements obtained through the Manual Strapping Method (MSM) or the more modern Electro Optical Distance Ranging (EODR) technique [4,5]. While both methods provide the necessary depth-to-volume relationships, the Manual Strapping Method (MSM) is often constrained by high operational costs and time requirements, resulting in remarkably small datasets [6,7,8].

This scarcity of data poses a significant challenge for the integration of Machine Learning (ML) models, which require substantial volumes of information to achieve high predictive accuracy and robustness [9.10,11]. Small sample sizes often lead to overfitting, where a model captures noise rather than the underlying physical geometry of the tank [12,13,14]. To bridge this gap, data augmentation offers a viable solution to synthetically expand empirical datasets without losing the integrity of the original measurements [15,16].

This study explores the application of the Gaussian Noise method to augment MSM and EODR datasets. By injecting controlled stochastic variations into the original volume-depth measurements, the Gaussian Noise method can generate a more dense data structure that preserves the essential statistical properties and linear/non-linear trends of the physical tanks. This approach aims to create a more resilient foundation for computational modeling, ensuring that the distinct volumetric differences between MSM and EODR methods are accounted for in a data-rich environment.

2. Methodology

The primary objective of this research is to address the data scarcity issue in crude oil storage tank volume calibration by increasing the size of the available empirical datasets. The small sample size (30 data points per method) is insufficient for robust machine learning model training. Data augmentation using Gaussian noise is employed to generate synthetic data points while maintaining the statistical properties, trends, and relationships present in the original Manual Strapping Method (MSM) and Electro Optical Distance Ranging (EODR) datasets. The case studies are two crude oil storage tank calibration datasets, the first being the calibration dataset obtained through the Manual Strapping Method (MSM) method and the second being the calibration dataset obtained through the Electro Optical Distance Ranging (EODR) method. The data are the volume recorded at different tank depth and the two dataset are plotted in Figure 1 and Figure 2.

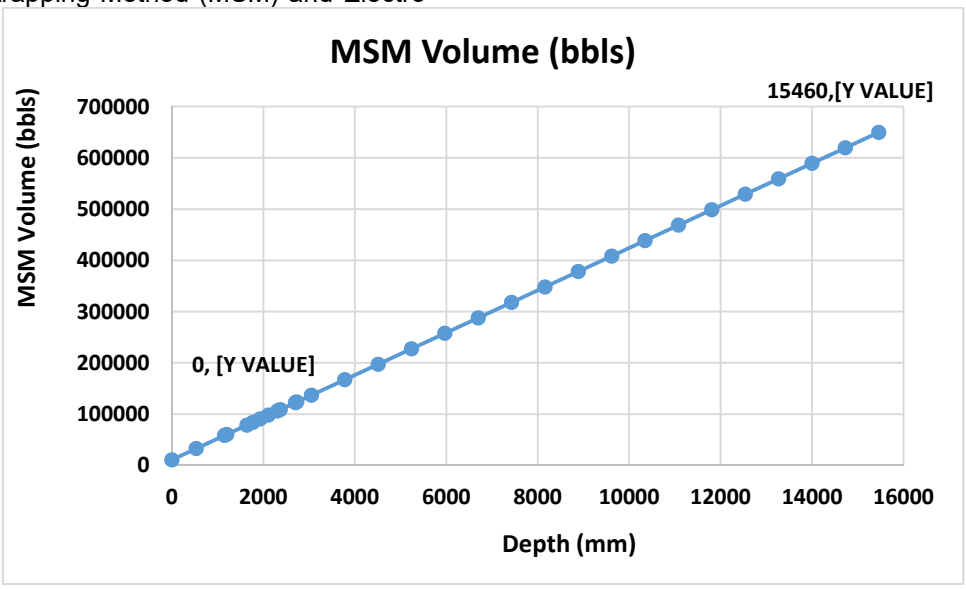

Figure 1 The graph of the crude oil storage tank calibration dataset obtained through the Manual Strapping Method (MSM) method

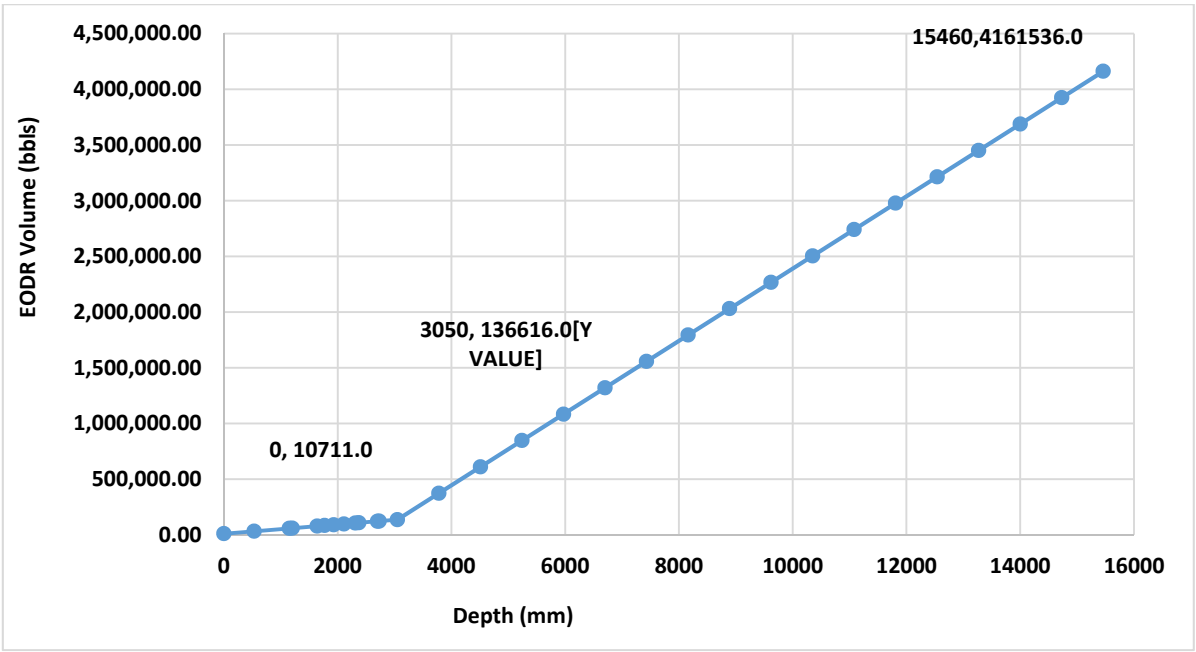

Figure 2 The graph of the crude oil storage tank calibration dataset obtained through the Electro Optical Distance Ranging (EODR) method

It can be seen that the two datasets have significant differences in the recorded crude oil volumes for different tank depths. Also, each of the two datasets has 30 data points. The essence of the data augmentation is to generate additional data point that will increase the number of data points while maintaining the statistical features of the original datasets. In this work, the Gaussian Noise model is used for the data augmentation.

The Gaussian noise augmentation is a technique used to generate synthetic data points by adding random noise sampled from a Gaussian (normal) distribution to existing data. This helps in increasing the dataset size and improving model generalization. Gaussian noise follows a normal distribution which is mathematically expressed as:

$$\boldsymbol{N(\mu, \sigma^2)} \textbf{ (1)}$$

Where, μ is the mean (expected value) of the distribution, σ^2 is the variance (spread) of the distribution, and σ is the standard deviation which controls how much variation is added. For a single EODR volume reading, a noisy version is generated as follows:

$$\boldsymbol{V_{(EODR,new)} = V_{(EODR,original)} + \epsilon} \textbf{ (2)}$$

Where ϵ is the Gaussian noise term, given as follows:

$$\boldsymbol{\epsilon \sim N(0, \sigma^2)} \textbf{ (3)}$$

In this case, ϵ is sampled from a normal distribution with mean 0, and σ is chosen based on the expected sensor noise level (e.g., 1% to 5% of the measurement). For each EODR volume reading V_{EODR}, multiple noise version is generated as follows:

$$\boldsymbol{V_{(EODR,augmented)} = V_{EODR} + \epsilon} \textbf{ (4)}$$

For multiple augmentations, the process is repeated with different noise samples:

$$\boldsymbol{V^{(i)}_{(EODR,augmented)} = V_{EODR} + \epsilon^{(i)} \; for \; i = 1, 2, ., k} \textbf{(5)}$$

Where, k is the number of augmented samples per original data point. To determine the noise level, the standard deviation is set as follows:

$$\boldsymbol{\sigma = \alpha \cdot \bar{V}_{EODR}} \textbf{ (6)}$$

For instance, given EODR reading of $500\, bbls$, if the noise level is selected as $\alpha = 0.02$ (2% noise), then the standard deviation can be computed as follows:

$$\sigma = 0.02 \times 500 = 10$$

Then, to generating 5 augmented samples using the Gaussian Noise model, the Gaussian noise is sampled from $N(0, 10^2)$, which then gives the new data as presented in Table 1. So, from the output, instead of just one EODR reading (500 bbls), about 5 variations of the EODR data records are generated which can help improve the model training.

Table 1: Sample data augmentation output using the Gaussian Noise model with noise data generated for the case of $N\left(0, 10^2\right)$

Sample	Noise ϵ	Augmented volume
1	$+3.5$	503.5
2	-7.8	492.2
3	$+12.1$	512.1
4	-4.3	495.7
5	$+9.2$	509.2

The data generated using the Gaussian Noise model is compared with the actual (original) using the an online Means Difference Confidence Interval calculator by statskingdom.com.

The data augmentation process is implemented through the following steps:

i. **Normalization:** The original 30-point datasets are normalized to a consistent scale to ensure numerical stability during the noise addition process.

ii. **Noise Application:** For each of the 30 data points in both the MSM and EODR datasets, random noise ε is generated and added multiple times (e.g., creating 50 or 100 variations per data point) to increase the dataset size by a factor of k.

iii. **Variance Selection:** The standard deviation (σ) of the Gaussian noise is carefully selected to reflect the expected experimental error in tank calibration (e.g., 0.5%–1% of the original value) to ensure the generated data remains physically realistic.

iv. **Integration:** The generated synthetic data points are combined with the original 30 data points to form a larger training dataset.

v. **Validation:** The statistical distribution (mean, standard deviation, and correlation) of the augmented dataset is compared with the original dataset to ensure that the fundamental relationship between depth and volume is maintained, while the density of the data has increased.

This method allows for the creation of a synthetic dataset that is significantly larger than the initial 30 points, improving the generalization ability of the downstream machine learning models used to predict volume from tank depth.

3. Results and discussion

3.1 Results of the Data Augmentation using the Gaussian Noise Method

The implementation of the Gaussian Noise Method for augmenting the crude oil storage tank volume calibration dataset yielded a significantly expanded sample size while maintaining the underlying physical characteristics of the original measurements. The volume distribution before and after Gaussian noise augmentation is presented in Figure 3 while the volume distribution before and after Gaussian noise augmentation is presented in Figure 4. The augmented data closely mirrors the original MSM (Master Standard Meter) and EODR (Electro-Optical Distance Ranging) distributions, suggesting that the Gaussian method successfully introduced variability without distorting the fundamental statistical properties of the tank calibration curves.

The scatter plot in Figure 5 demonstrates a strong linear correlation between the original and augmented datasets. The overlap between the MSM and EODR volumes indicates that the noise injection remained within acceptable physical tolerances for storage tank environments. Prediction Accuracy: As shown in the line chart (Figure 6), the predicted MSM volumes closely track the actual volumes across the entire calibration range. However, Figure 7 highlights the residual errors, showing a maximum absolute error of 777.668 bbls.

The results in Table 2 provides a quantitative breakdown of the Gaussian method's performance. The model achieved a high Coefficient of Determination (R^2) of 0.9911, indicating that 99.11% of the variance in the volume data is captured. The Mean Absolute Percentage Error (MAPE) is exceptionally low at 0.0039%, translating to a 1-MAPE accuracy of 99.68%.

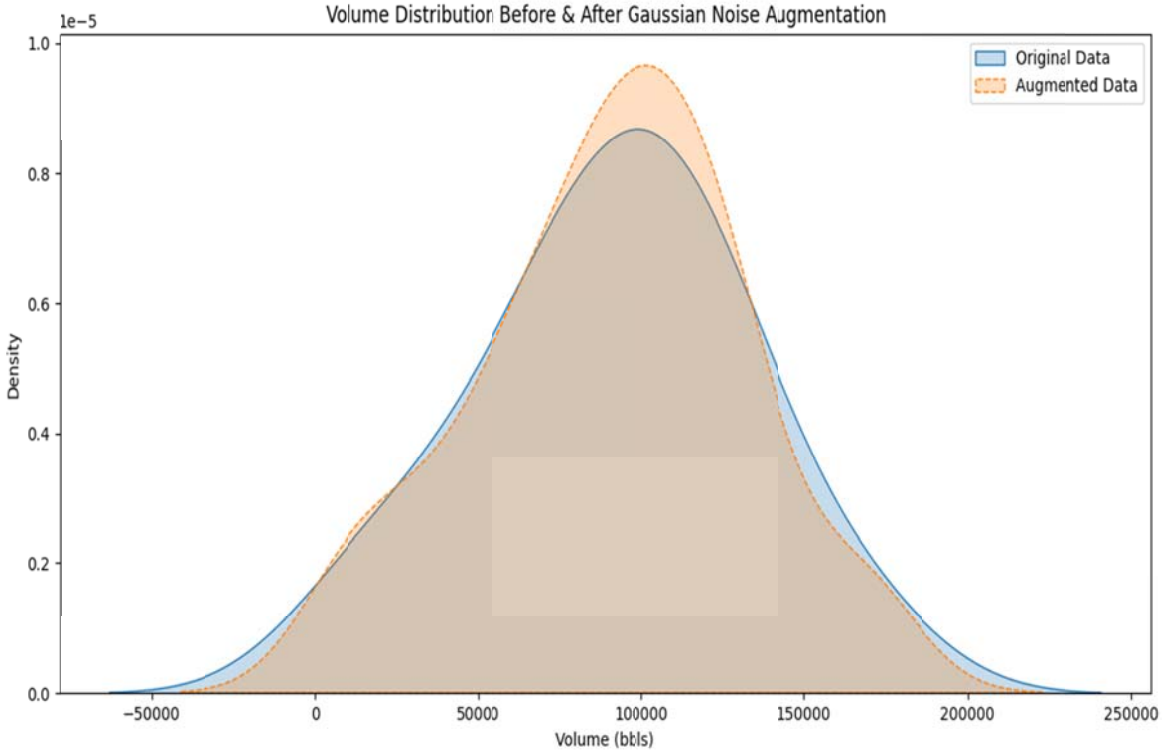

Figure 3 Volume distribution before and after Gaussian noise augmentation

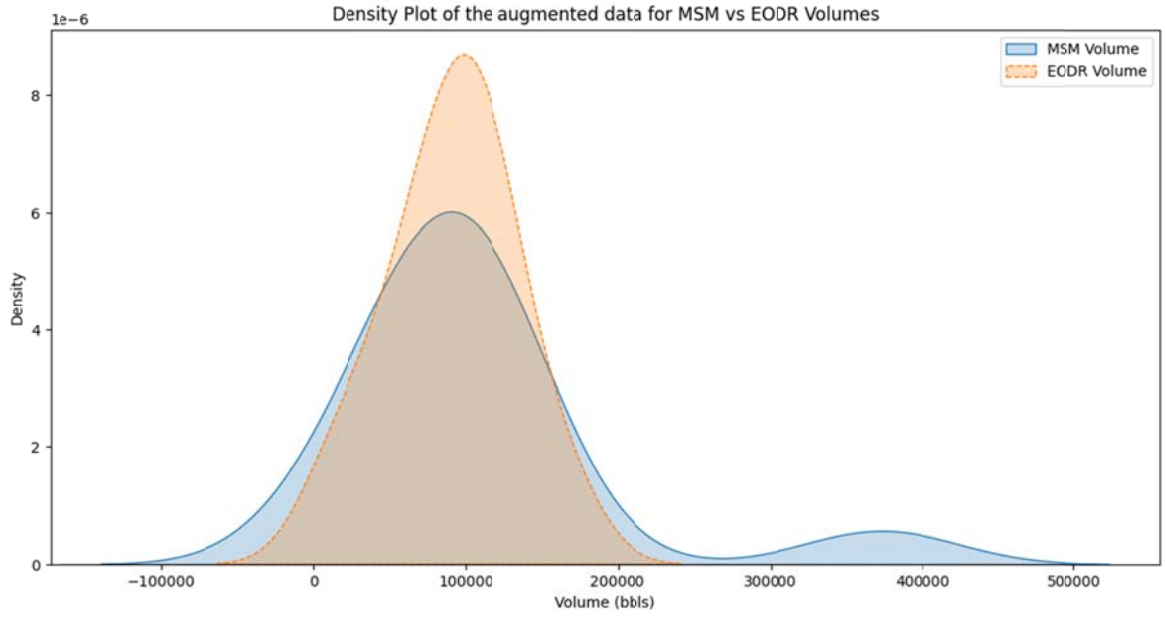

Figure 4: Density plot of the augmented data for MSM vs EODR volumes

Source: Researcher

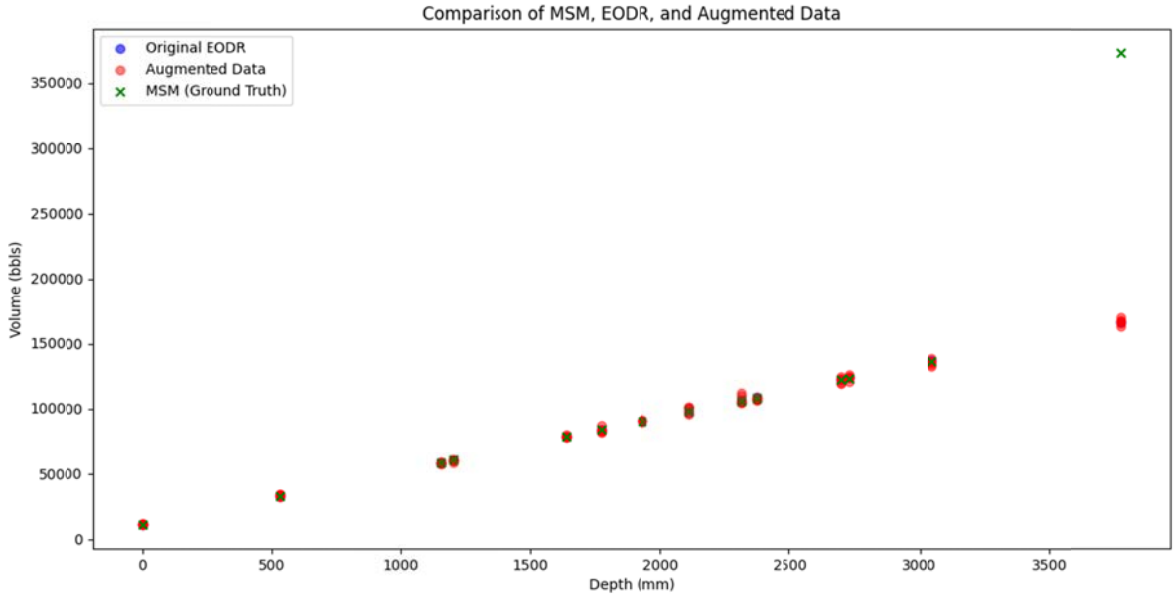

Figure 5: Comparison of MSM, EODR, and augmented data using a scatter plot

Source: Researcher

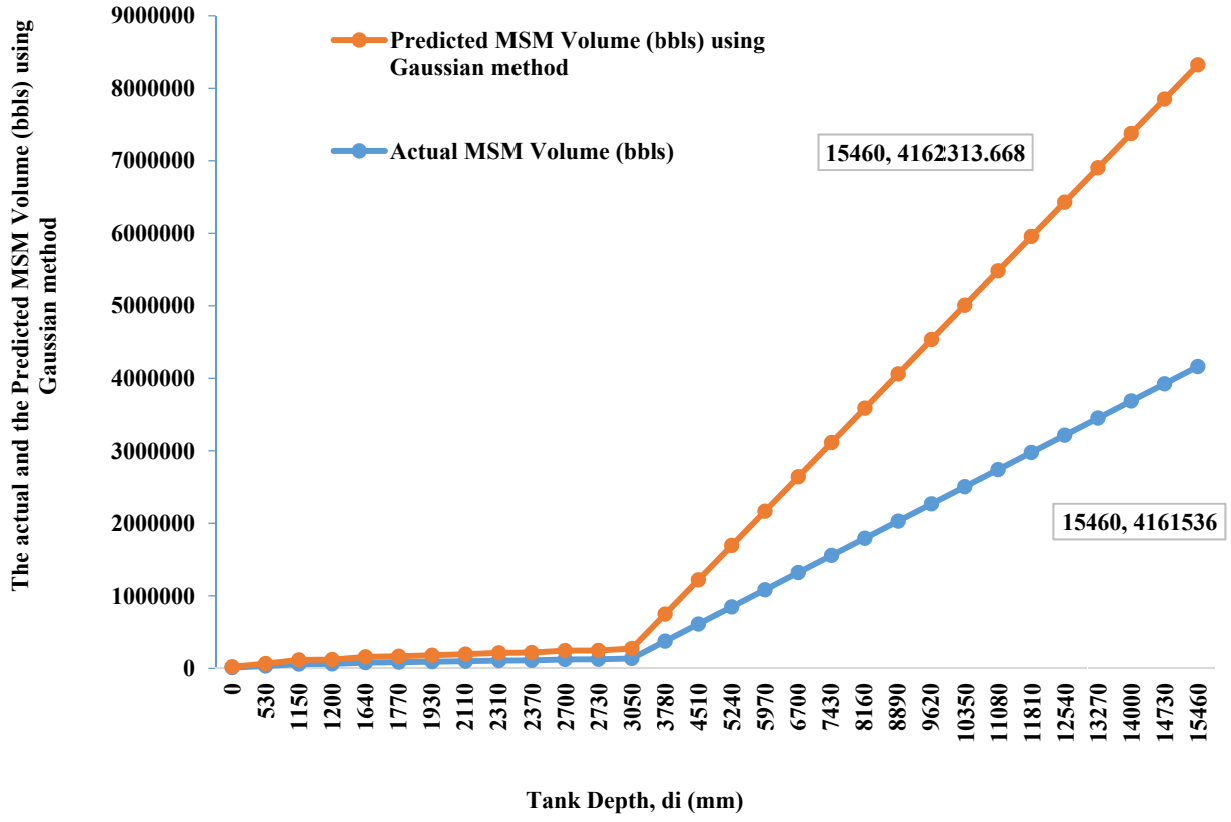

Figure 6: The line chart of the actual and the predicted MSM volume (bbls) using the Gaussian Method

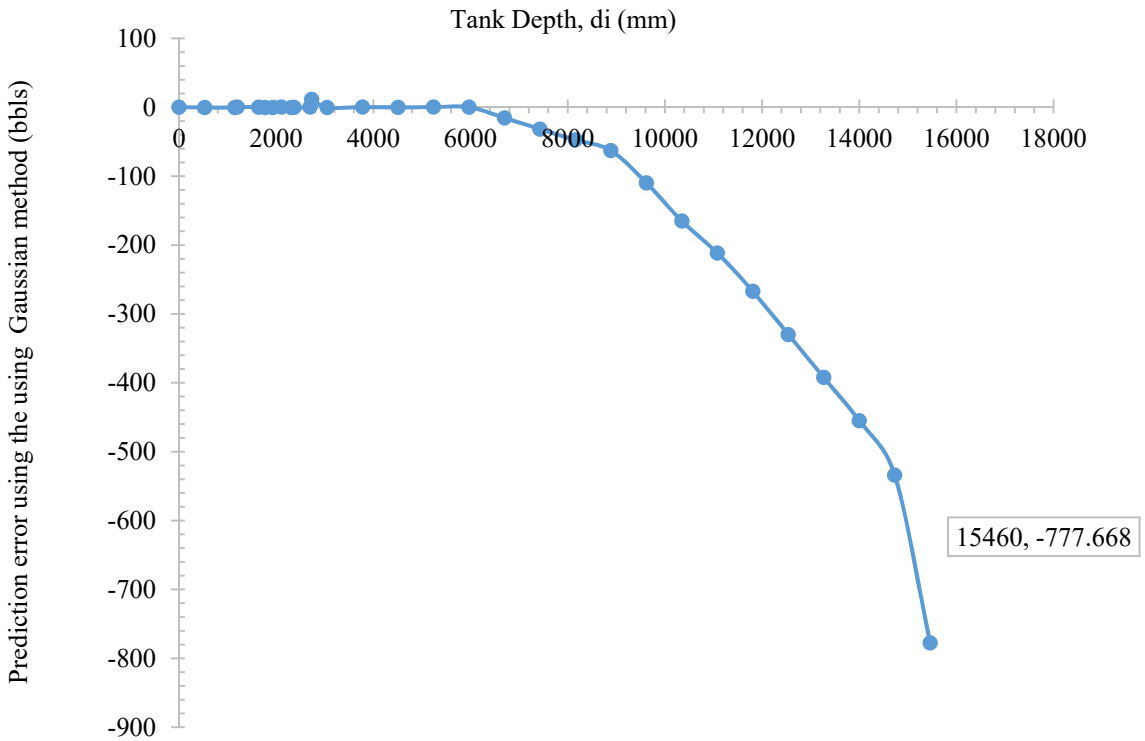

Figure 7: The scatter plot of the prediction error of the MSM volume (bbls) using the Gaussian method

Table 2: Statistical metrics for the augmented dataset using the Gaussian method

Metric	Prediction performance of the Gaussian method
MSE	50772.0495
RMSE	225.32654
MAE	113.884926
R^2	0.99110814
1-MAPE	0.99684333
Mean of Absolute Percentage Error (%)	0.00394215
Minimum Absolute Error	0.09889
Maximum Absolute Error	777.668

3.2 Discussion

The results indicate that the Gaussian Noise Method is an effective technique for addressing the scarcity of high-fidelity calibration data in crude oil storage applications. The high R^2 value and the 99.68% accuracy rate suggest that the augmented data can reliably train predictive models for tank volume estimation. The maximum absolute error of 777.668 bbls, while seemingly high in isolation, must be interpreted relative to the total capacity of the storage tanks. Given that the Mean Absolute Error (MAE) is much lower at 113.88 bbls, the larger errors likely occur at specific points in the tank geometry where Gaussian noise might over-simulate turbulence or measurement fluctuations. Despite this, the RMSE of 225.33 bbls confirms that the majority of the augmented data points remain close to the regression line, making this method a robust choice for enhancing dataset volume for machine learning applications in the oil and gas industry.

4. Conclusion

This research successfully demonstrated that the Gaussian Noise Method is a robust and effective strategy for augmenting crude oil storage tank volume calibration datasets. By expanding the sample size while preserving the physical integrity of the original Master Standard Meter (MSM) and Electro-Optical Distance Ranging (EODR) measurements, the method effectively addressed the challenge of data scarcity in high-fidelity calibration environments.

There are notable findings from this study include, and the first finding is about the statistical fidelity of the data augmentation using the Gaussian Noise

Method. The augmented datasets maintained the fundamental statistical properties and linear correlations of the original measurements, ensuring that the introduced variability remained within acceptable physical tolerances. Again, the model achieved an exceptional R^2 value of 0.9911, indicating that the augmented data captures 99.11% of the variance in tank volume. Also, with a Mean Absolute Percentage Error (MAPE) of only 0.0039%, the method delivered a 1-MAPE accuracy of 99.68%, despite a maximum absolute error of 777.668 bbls.

Ultimately, these results prove that Gaussian noise injection can significantly enhance the training and validation of calibration models without distorting the underlying tank calibration curves. This approach provides a reliable framework for improving the reliability of volumetric measurements in the petroleum industry, offering a cost-effective solution for generating high-quality datasets where physical measurements are limited.

References

1. Patel, S., Parrott, B., Abdellatif, F., & Trigui, H. (2020, November). Custody Transfer Tank Calibration Technology. In *Abu Dhabi International Petroleum Exhibition and Conference* (p. D012S116R012). SPE.

2. 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).

3. 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.

4. Burachek, V., Khomushko, D., Tereshchuk, O., Kryachok, S., & Belenok, V. (2022). Analysis of development tendencies of metrological technologies to control rangefinders of an electronic distance measurement instruments. *Advances in Geodesy and Geoinformation*, e13-e13.

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. 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.

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. 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.

9. Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, *8*, 784455.

10. Hakami, A. (2024). Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance. *Scientific Reports*, *14*(1), 9645.

11. Tsallis, C., Papageorgas, P., Piromalis, D., & Munteanu, R. A. (2025). Application-wise review of Machine Learning-based predictive maintenance: Trends, challenges, and future directions. *Applied Sciences*, *15*(9), 4898.

12. Aliferis, C., & Simon, G. (2024). Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, 477-524.

13. Khayyam, H., Golkarnarenji, G., & Jazar, R. N. (2018). Limited data modelling approaches for engineering applications. In *Nonlinear approaches in engineering applications: energy, vibrations, and modern applications* (pp. 345-379). Cham: Springer International Publishing.

14. Kuhn, M., & Johnson, K. (2013). Over-fitting and model tuning. In *Applied predictive modeling* (pp. 61-92). New York, NY: Springer New York.

15. Kumar, T., Brennan, R., Mileo, A., & Bendechache, M. (2024). Image data augmentation approaches: A comprehensive survey and future directions. *Ieee Access*, *12*, 187536-187571.

16. Gracia Moisés, A., Vitoria Pascual, I., Imas González, J. J., & Ruiz Zamarreño, C. (2023). Data augmentation techniques for machine learning applied to optical spectroscopy datasets in agrifood applications: A comprehensive review. *Sensors*, *23*(20), 8562.