Improving Palm Oil Price Forecasting Using a Gaussian-Enhanced Generative Adversarial Imputation Network

1st Muhammad Zakky Ghufron
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025231022@student.its.ac.id

2nd Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@its.ac.id

3rd Muhammad Suzuri Hitam
Faculty of Computer Science and Mathematics
Universiti Malaysia Terengganu
Terengganu, Malaysia
suzuri@umt.edu.my

4th Agus Tri Haryono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025231020@student.its.ac.id

2025 International Conference on Data Science and Its Applications (ICoDSA) | 979-8-3315-9854-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICoDSA67155.2025.11157606

***Abstract*—Accurate forecasting of palm oil prices plays a vital role in guiding trade policies and maintaining market stability, especially in the growing economic and climate-related uncertainties. However, missing values in palm oil datasets present a major challenge to build reliable forecasting models. This study proposes an improvement to the Generative Adversarial Imputation Network (GAIN) by replacing its original uniform random noise generator with a Gaussian-based generator using the Box-Muller transform to improve the quality and stability of imputing missing values in multivariate datasets. The enhanced GAIN shows better imputation performance, as indicated by lower Root Mean Square Error (RMSE) and Euclidean Distance (ED) compared to the original version. This improvement also translates into better forecasting results when used with Long Short-Term Memory (LSTM), Random Forest (RF), and Linear Regression (LR) models, which achieved RMSEs of 818.45, 932.12, and 858.18, respectively. Further analysis using Pearson's correlation shows that accurately preserving key statistical properties—such as the mean and standard deviation—plays an important role in improving predictive accuracy. Overall, the proposed method offers an effective solution for handling missing data in complex, real-world forecasting tasks. Future work may explore extending this approach to capture cross-sectional patterns in panel data, enabling more context-aware and structure-sensitive imputation.**

Keywords—GAIN, gaussian random number generator, imputation missing value, palm oil, time series forecasting

I. INTRODUCTION

Palm oil plays a vital role in the economies of major producing countries, particularly Indonesia and Malaysia, where it serves as a key export commodity and a driver of the agribusiness sector [1]. Global demand for palm oil continues to grow due to its wide range of uses in food, cosmetics, and biofuels. However, this increasing demand is met with instability in production and supply, driven by economic pressures and climate variability [2]. These factors contribute to significant fluctuations in global palm oil prices, with production often affected by temperature and rainfall patterns and prices influenced by domestic consumption, exports, and import volumes [3].

Therefore, reliable forecasting of palm oil prices are essential to support trade policies and improve market stability. Achieving this, however, depends heavily on the quality and completeness of the available data. A persistent challenge in working with economic and environmental datasets is the presence of missing values. Incomplete data can introduce bias, reduce model accuracy, and lead to misleading conclusions—especially in forecasting and econometric analysis [4].

Previous techniques for imputing missing data include K-Nearest Neighbors (KNN) and Multiple Imputation by Chained Equations (MICE). A study by Maimunah et al. evaluated several imputation methods such as linear interpolation, Kalman Filter, MICE, and KNN on palm oil-related datasets using the Statistical-Temporal Evaluation Framework (STEF), and found that KNN consistently yielded the highest accuracy (95.00%) and the lowest mean absolute error (MAE)While these methods are widely used, they have notable limitations. KNN may fail to capture complex or nonlinear patterns in the data. MICE requires appropriate regression models for each variable and can suffer from convergence issues or overfitting in high-dimensional settings [5, 6]. In recent years, Generative Adversarial Imputation Networks (GAIN) have emerged as a promising deep learning-based approach for handling missing data [7]. Based on the principles of Generative Adversarial Networks (GANs), GAIN learns the underlying structure of the observed data and generates realistic imputations through adversarial training. GAIN also selected in this study based on previous study showing that it outperforms traditional imputation methods such as MICE, particularly in handling complex and high-dimensional data structures [8]. While GAIN has shown strong performance in various domains, certain aspects of the method still offer opportunities for improvement. In its original formulation, GAIN relies on a uniform random number generator to create the input noise matrix, which may not always reflect the underlying distribution of the data.

To address this limitation, this study proposes a modified version of GAIN incorporating a Gaussian random number generator instead of the standard uniform generator. By introducing a noise matrix sampled from a Gaussian distribution, the model is expected to capture real-world data's statistical characteristics better, leading to more accurate and stable imputations. This modification aims to improve the quality and reliability of the imputed values, especially in complex datasets that combine economic and climate-related variables across multiple countries and years. Dataset used in this study consisted of annual data from 1964 to 2024 across 16 major palm oil-producing countries. The dataset includes 20 economic and climate-related variables, with domestic price as the target variable.

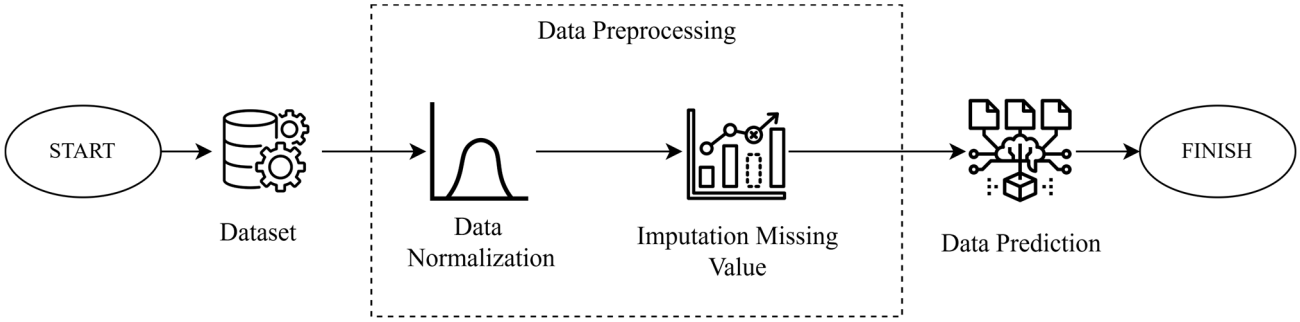

Fig. 1. General research methodology.

The remainder of this paper was divided into following sections: Section II presents related research; Section III describes the dataset and proposed methodology; Section IV details the experimental results; and Section V concludes with findings and future research.

II. Related Work

Various studies have explored different strategies for imputing missing values. For instance, Alnowaiser [6] addressed missing values using KNN, effectively replacing invalid zeros in a medical dataset. This imputation step significantly improved data quality and contributed to the high performance of their proposed Tri-Ensemble model, which achieved 94.70% accuracy. Another research conducted by Maimunah et al. [9] used several statistical methods for imputating missing values, such as interpolation, KNN, and Kalman Filter. These methods were compared with the Statistical-Temporal Evaluation Framework (STEF), which evaluates based on distribution, independence, and the temporal relationship. The result showed that KNN outperformed all methods with 95.00% accuracy and lowest error values.

Several studies also used deep learning approaches to address the problem of missing data imputation. [10] proposed Labels Wasserstein Generative Adversarial Imputation Network (LWGAIN), an enhanced imputation method tailored for well-log data in lithology identification. It improves upon the original GAIN by incorporating Wasserstein distance into the loss function for more stable training and adding label information to the generator input. Experiments on this research demonstrate that LWGAIN

TABLE I. Dataset Description

Indicator	Description	Unit
Economic Indicator		
Area Harvested	Total harvested area of palm oil plantations	1000 ha
Beginning Stocks	Palm oil stocks available at the beginning of the year	1000 MT
Dom. Consumption	Total domestic consumption of palm oil	1000 MT
Ending Stocks	Remaining palm oil stocks at the end of the year	1000 MT
Exports	Total volume of exported palm oil	1000 MT
Feed Waste Cons.	Palm oil consumption for animal feed and domestic waste	1000 MT
Food Use Cons.	Palm oil consumption for domestic food purposes	1000 MT
Imports	Total volume of imported palm oil	1000 MT
Industrial Cons.	Palm oil consumption in the industrial sector	1000 MT
Production	Total volume of palm oil produced within one year	1000 MT
Total Distribution	Total distributed palm oil, including domestic consumption and exports	1000 MT
Total Supply	Total available palm oil in one year, including production, imports, and beginning stocks	1000 MT
Yield	Palm oil crop productivity measured by production per unit of harvested area	MT / ha
Total Demand	Total Supply minus Beginning Stocks of the following year	1000 MT
Price Domestic	Average domestic price calculated from export/import values and volumes of CPO and other palm oil	USD / MT
Price Global	Rata-rata Price Domestic pada tahun tertentu	USD / MT
Climate Indicators		
Mean Temperature	Annual average temperature of a country, calculated from daily mean temperatures	°C
Min. Temperature	Minimum annual temperature recorded in a country	°C
Max. Temperature	Maximum annual temperature recorded in a country	°C
Precipitation	Total annual rainfall in a country	mm

TABLE II. Dataset Statistics

Variable	Standard Deviation	Mean	Number of Missing Value	Missing Value Percentage (%)
Area Harvested	1657.02	572.36	376.00	38.52
Beginning Stocks	620.25	184.40	399.00	40.88
Dom. Consumption	2076.85	757.95	71.00	7.27
Ending Stocks	658.76	197.68	385.00	39.45
Exports	4037.50	1098.83	314.00	32.17
Feed Waste Cons.	44.49	11.84	796.00	81.56
Food Use Cons.	1322.17	507.20	91.00	9.32
Imports	1131.92	244.01	335.00	34.32
Industrial Cons.	1025.11	237.76	526.00	53.89
Production	5594.37	1626.05	95.00	9.73
Total Distribution	6301.67	2054.46	66.00	6.76
Total Supply	6301.67	2054.46	66.00	6.76
Yield	1.52	1.52	376.00	38.52
Total Demand	5448.19	1762.22	83.00	8.50
Price Domestic	347.51	655.34	0.00	0.00
Price Global	134.85	627.44	0.00	0.00
Mean Temperature	4.87	24.00	32.00	3.28
Min. Temperature	4.20	18.96	32.00	3.28
Max. Temperature	5.69	29.05	32.00	3.28
Area Harvested	724.21	1843.74	32.00	3.28

achieved the lowest RMSE and Mean Absolute Error (MAE), outperforming KNN by an average RMSE reduction of 0.0220, Random Forest by 0.0072, Deep Convolutional Generative Adversarial Network (DCGAN) by 0.0052, and GAIN by 0.0046. Another research [11] proposed a novel imputation method combining dummy variable encoding with unsupervised learning. This dummy variable is concatenated with the original data and input into a generative-discriminative architecture based on a GAN-like unsupervised learning framework. The proposed method consistently achieves the lowest RMSE across various missing rates (20%–60%). For instance, at a 50% missing rate on the full dataset, the proposed method yields an RMSE of 0.0965, outperforming 0-imputation an RMSE of 0.3057, Random Forest an RMSE of 0.2001, Autoencoder an RMSE of 0.1810, and unsupervised learning without dummy variables an RMSE of 0.2029.

Based on past research, although the existing imputation methods for missing values can meet the requirements to some extent, certain aspects of the method still offer opportunities for improvement. While in original method GAIN used Uniform Random Number Generator to generate random matrix, this study proposed modified GAIN imputation method using a Gaussian Random Generator to generate the random matrix to stabilize the imputing of missing values. The rationale is that if the imputed values follow a distribution closer to the actual data, the resulting forecasting performance can be improved.

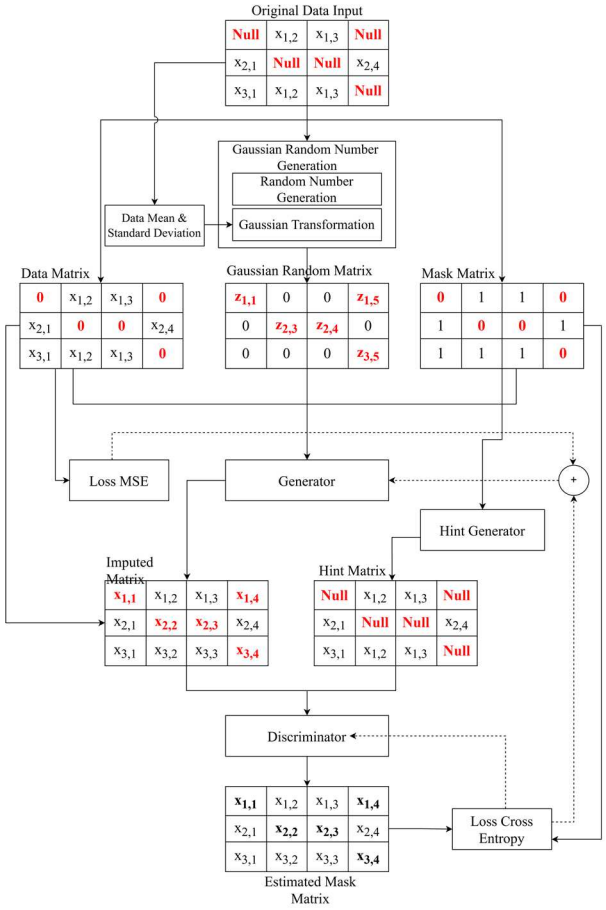

Fig. 2. Proposed Gaussian-GAIN method.

III. Methodology

In this chapter, the proposed method is illustrated by the flowchart shown in Fig. 1. This research was conducted across three primary phases: Data Preprocessing, which consists of Data Normalization and Imputation, followed by data forecasting.

A. Dataset

The dataset utilized in this study is structured as panel data, which combines both cross-sectional and time-series dimensions. The time-series component comprises annual data spanning from 1964 to 2024, while the cross-sectional dimension includes the world's major palm oil-producing countries: Indonesia, Malaysia, Thailand, Colombia, Nigeria, Guatemala, Papua New Guinea, Côte d'Ivoire, Honduras, Brazil, Cameroon, Ecuador, India, Ghana, the Democratic Republic of the Congo, and Peru. This dataset consists of 20 features from different categories: economic and climate indicators, with the response variable being domestic price. The details of each feature can be seen in Table I.

A descriptive statistical analysis was conducted for each variable to understand better the characteristics of the dataset used in this study, as presented in Table II. This analysis includes key metrics such as the number of available observations, the mean, the standard deviation, and the total number of missing values, which includes NaN and zero values. The missing value percentage then calculated by taking the ratio of the number of missing values to the total number of observations (976 observation) and expressing it as a percentage. The complete version of data can be accessed at https://github.com/noobPeople/palm-oil-forecasting.

B. Normalization

Min-max normalization was applied to the dataset to ensure all variables contribute proportionally to the analysis and mitigate the effects of differing units and scales among features. This method transforms each variable to a 0 to 1 scale [12] by rescaling values based on the minimum and maximum observed within each feature, which can be described as Eq. (1).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X is the original value, X_min is the minimum value of a variable, while X_max is the maximum value of a variable.

C. Imputation Missing Value

This study used the GAIN method to impute missing data. GAIN [8] is a variant of GAN designed explicitly for imputing missing data. The architecture consists of two components: a generator G and a discriminator D. The generator receives three inputs: the incomplete data $\tilde{X}$, a binary mask M indicating the location of missing values, and a noise matrix Z generated using a uniform random distribution. The generator outputs a completed version of the data $\bar{X}$, which is then combined with the observed data to produce the imputed dataset $\hat{X}$, as shown in Eq. (2) and Eq. (3).

$$\bar{X} = G(\tilde{X}, M, (1 - M) \odot Z) \quad (2)$$

$$\hat{X} = M \odot \tilde{X} + \bar{X} \odot (1 - M) \quad (3)$$

To guide the discriminator, a hint vector H is generated by partially revealing the mask M, controlled by a predefined hint

TABLE III. GAIN IMPUTATION RESULTS

Random Tools	Range Value	Mean : Std	Mean Difference	Std Difference	SWD	ED	RMSE
Mersenne Twister	0.00 : 0.99	-	**9.18**	**55.41**	**209.49**	249087.23	4145.61
	0.00 : 0.50	-	28.55	208.22	234.14	**204554.30**	**3404.45**
	0.00 : 0.10	-	93.61	408.02	399.25	219706.00	3657.09
	0.00 : 0.01	-	98.57	409.37	395.34	218021.11	3630.35
Mersenne Twister + Box-Muller	0.00 : 1.00	0.15 : 0.30	**42.63**	**251.11**	**264.17**	237594.90	3956.94
		0.15 : 0.20	42.99	276.71	265.93	217297.17	3618.73
		0.15 : 0.10	44.95	284.74	270.41	**198073.17**	**3297.69**
	0.00 : 0.50	0.15 : 0.30	**16.45**	**170.30**	**263.37**	258311.68	4302.65
		0.15 : 0.20	38.99	234.42	271.16	232558.85	3872.49
		0.15 : 0.10	36.78	273.80	282.49	**211784.67**	**3527.07**
	0.00 : 0.10	0.15 : 0.30	129.50	341.00	395.46	241653.35	4024.81
		0.15 : 0.20	68.16	79.42	265.96	197103.60	3281.02
		0.15 : 0.10	**25.33**	**76.75**	**225.51**	**176020.27**	**2931.61**

rate h, which is shown in Eq. (4). The discriminator then attempts to predict the probability that each component of $\hat{X}$ is observed or imputed, as defined in Eq. (5). Unlike standard GANs, which classify data as real or fake, the GAIN discriminator operates at the feature level.

$$H = M \odot h \tag{4}$$

$$\hat{D} = D(\hat{X}, H) \tag{5}$$

GAIN employs different loss functions for each component: Mean Squared Error (MSE) for the generator and Cross-Entropy (CE) for the discriminator. The generator is trained to minimize a combined loss that encourages accurate imputation, while the discriminator is trained to maximize the ability to distinguish between actual and imputed values. The overall objective function is presented in Eq. (6).

$$V(D,G) = M \odot log\,\hat{D} + (1-M) \odot log(1-\hat{D}) \tag{6}$$

To enhance the quality and accuracy of imputed values, this study proposes an extension of the original GAIN framework by modifying the random number generation algorithm used to create the random matrix Z. Specifically, instead of using the default uniform distribution generated by the Mersenne Twister algorithm, this research adopts a Gaussian (normal) distribution generated via the Box-Muller Transform. Gaussian distribution is symmetric around the mean, increasing the likelihood that random values cluster near the average. This approach is expected to improve the quality of imputation and, consequently, the accuracy of downstream forecasting tasks. The Box-Muller Transform generates normal distributed random variables Z_1 and Z_2 from two independent uniform random variables U_1 and U_2 [13], as shown in Eq. (7) and Eq. (8).

$$Z_1 = \sqrt{-2\,ln\,U_1} \cdot cos(2\pi U_2) \tag{7}$$

$$Z_2 = \sqrt{-2\,ln\,U_1} \cdot cos(2\pi U_2) \tag{8}$$

The proposed workflow of the imputation missing value is illustrated in Fig. 2.

D. Data Forecasting

This study employs three predictive modeling approaches: LSTM, Random Forest, and Linear Regression, each selected for their respective strengths in handling time series data. LSTM is a Recurrent Neural Network (RNN) specifically designed to capture temporal dependencies in sequential data [14][15]. Its architecture includes memory cells and gating mechanisms that enable the model to retain long-term contextual information while mitigating issues such as vanishing gradients. The LSTM is particularly suitable for complex time series tasks where historical trends significantly influence future outcomes [16].

Random Forest is a robust ensemble learning method based on decision trees. It constructs multiple trees by randomly sampling subsets of the training data and aggregates their forecasting to enhance accuracy and reduce overfitting [17]. Its non-parametric nature allows it to model nonlinear relationships effectively, making it a powerful tool for structured time series forecasting [18].

Lastly, Linear Regression is utilized as a benchmark statistical method. It models the relationship between input variables and a continuous target variable under the assumption of linearity [19]. Despite its simplicity, Linear Regression provides interpretability and serves as a valuable baseline for evaluating the performance of more complex models.

TABLE IV. GAIN AND FORECASTING RESULTS COMPARISON

Random Tools	Range Value	Mean : Std	Mean Difference	Std Difference	SWD	ED	LSTM RMSE	RF RMSE	LR RMSE
Mersenne Twister	0.00 : 0.99	-	1037.48	1508.09	1854.80	669690.21	1443.22	1621.53	1633.68
	0.00 : 0.50	-	813.92	938.51	1369.28	527274.84	968.52	1110.14	**1137.90**
	0.00 : 0.10	-	628.20	620.81	1360.90	429378.29	984.34	1033.78	1320.09
	0.00 : 0.01	-	**533.93**	**462.35**	**1228.28**	**373200.13**	**953.60**	**954.83**	1298.72
Mersenne Twister + Box-Muller	0.00 : 1.00	0.15 : 0.30	721.95	899.69	1644.58	510458.10	1588.82	1582.67	1744.91
		0.15 : 0.20	**540.25**	**377.35**	**1129.67**	**346503.78**	**818.45**	**932.12**	**858.18**
		0.15 : 0.10	684.70	633.11	1264.92	437764.08	1018.02	1111.29	1066.78
	0.00 : 0.50	0.15 : 0.30	779.43	904.88	1678.68	516415.21	1159.66	1315.53	1314.08
		0.15 : 0.20	774.12	838.82	1389.74	499676.98	973.35	1073.92	**952.01**
		0.15 : 0.10	**605.46**	**598.52**	**1384.00**	**420900.43**	**936.84**	**1018.35**	1123.55
	0.00 : 0.10	0.15 : 0.30	1304.25	2188.03	2667.59	819164.17	1231.47	1193.98	1691.48
		0.15 : 0.20	1044.54	1505.25	1986.16	669474.74	1110.97	1057.34	1862.13
		0.15 : 0.10	**778.04**	**852.71**	**1456.19**	**503207.83**	**980.19**	**1026.94**	**1208.25**

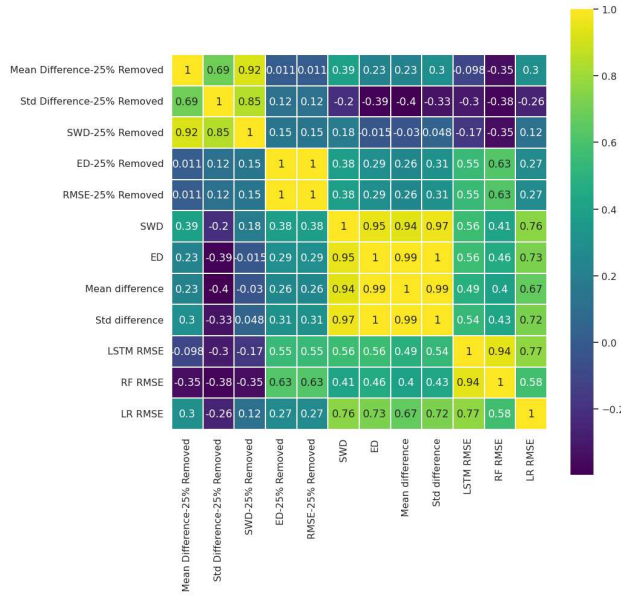

Fig. 3. Pearson correllation on forecasting and GAIN results.

IV. Result and Discussion

This section presented the experimental findings and analyzed the effectiveness of employing the Gaussian Random Generator in the GAIN method. This experiment was evaluated with several metrics, such as RMSE, Sliced Wasserstein Distance (SWD), and Euclidean Distance (ED), to compare the difference and distribution between original and predicted data. This study also aimed to evaluate and compare the imputation performance across various settings of range values for uniform random generation and standard deviation parameters for Gaussian random transformation.

A. GAIN Imputation Result

To evaluate the performance of the proposed imputation method, an additional 25% of the observed (non-missing) data was intentionally removed at random from the original dataset. These artificially removed values served as ground truth for assessing the accuracy of the imputation results. The model was trained in over 1,000 iterations with a fixed hint rate of 0.9. Evaluation metrics such as RMSE were employed to measure the quality of imputations, while SWD and ED were also employed to measure the distribution of imputation results. The experimental results provide insight into the proposed GAIN's ability to recover missing values accurately compared to the regular GAIN imputation method, as seen in Table III.

Based on the results presented in Table III, the original GAIN implementation using the Mersenne Twister uniform random generator demonstrates relatively lower values in terms of statistical distance and deviation metrics. Specifically, it achieves a SWD of 209.49, a mean difference of 9.18, and a standard deviation difference of 55.41—indicating the distribution of imputed values closely follows the original data distribution. The uniform distribution produces well-aligned synthetic samples regarding overall statistical shape.

However, the proposed GAIN variant significantly outperforms the original version when evaluating imputation accuracy using ED and RMSE. Among all configurations tested, the Gaussian generator with a mean of 0.15 and standard deviation of 0.10 achieves the lowest RMSE of 2931.61 and ED of 176020.27. These values indicate a more precise reconstruction of the missing values than the original GAIN despite having slightly less ideal distributional alignment.

B. Influence of Imputation Models on Forecasting Results

The imputation was applied directly to the real-world dataset containing naturally missing values without removing any originally complete entries. The results presented in Table IV showed how different imputation settings influence the performance of subsequent forecasting using LSTM, RF, and LR models. Among all tested configurations, the best-performing setup was achieved using the proposed GAIN variant with a Gaussian random generator with a mean of 0.15 and std of 0.20, which resulted in the lowest overall RMSE across all models: LSTM of 818.45, RF of 932.12, and LR of 858.18. This configuration also achieved the lowest ED of 346503.78 and SWD of 1129.67, indicating that the imputations were both distributionally sound and numerically close to the expected values. The strong performance of the configuration with mean 0.15 and std 0.20 can be attributed to its alignment with the statistical characteristics of the scaled original data, where 0.15 approximates the average value. The standard deviation of 0.20 likely provides a balanced level of variation—neither too narrow nor too dispersed—allowing the generator to explore meaningful imputations without introducing excessive randomness. Nevertheless, since the values are generated randomly, some level of uncertainty remains, and the outcomes cannot be precisely determined for every run.

By contrast, the baseline imputation using Mersenne Twister with uniform distribution showed higher RMSE values across the three forecasting models, highlighting the advantage of the proposed Gaussian-based approach in preserving temporal and statistical patterns relevant for forecasting. These findings suggest that the quality of the imputed values directly and significantly impacts the downstream forecasting performance and that careful selection of the random generation method in GAIN can substantially improve model performance. Thus, incorporating Gaussian-distributed noise during the imputation process appears to capture the underlying structure of the data better and supports more reliable forecasting outcomes.

C. Correlation Analysis

This study conducts a correlation analysis using the Pearson correlation coefficient to gain a deeper understanding of the factors influencing forecasting performance. In this study, Pearson correlation was selected for the correlation analysis because it measures the strength and direction of linear relationships between continuous numerical variables, which aligns with the objective of evaluating linear dependencies between the evaluation of imputed data and forecasting performance. Fig. 3 provides insight into the relationship between imputation quality metrics and errors across three models: LSTM, RF, and LR. Notably, LSTM RMSE exhibits a positive correlation with Standard Deviation Difference of 0.54, Mean Difference of 0.49, and SWD of 0.56, suggesting that LSTM performance is susceptible to the distributional accuracy of the imputed data. A remarkably high correlation of 0.94 is observed between LSTM RMSE and RF RMSE, indicating a shared sensitivity to the quality of imputations across both models.

Similarly, RF RMSE shows positive correlations with a Mean Difference of 0.43, Standard Deviation Difference of

0.4, and SWD of 0.46. These results suggest that deep learning and tree-based models respond similarly to distributional discrepancies in the imputed dataset. In contrast, although LR RMSE generally shows weaker correlations overall, it still demonstrates moderate to strong correlations with SWD of 0.76, ED of 0.73, and Standard Deviation Difference of 0.72. The LR performance implies that while Linear Regression is a simpler model, its performance is still influenced by how well the imputation process preserves the statistical properties of the original data.

V. Conclusion

This study presents a comprehensive evaluation of missing data imputation using GAIN and a proposed enhancement that integrates Gaussian-distributed noise generated through the Box-Muller transform. While the original GAIN, utilizing uniform random noise via the Mersenne Twister, produced imputed values with distributional characteristics closely aligned to the original data, the modified GAIN demonstrated superior performance in terms of imputation accuracy reflected in lower RMSE and Euclidean Distance values. This improvement directly improved forecasting outcomes when using LSTM, Random Forest, and Linear Regression models, particularly in scenarios involving real-world missing values. Furthermore, correlation analysis using Pearson's coefficient revealed that accurate recovery of statistical distribution—particularly mean and standard deviation—is critical for ensuring high accuracy, regardless of the predictive model employed. Despite these improvements, a notable limitation of GAIN remains its inability to capture cross-sectional dependencies within panel data structures. Therefore, future work could focus on developing an extended version of GAIN capable of modeling interdependencies across cross-sectional units, enabling more context-aware imputation in panel data settings. Additionally, future research may consider explicitly addressing seasonality and stationarity characteristics in the data, particularly to improve the effectiveness of time-series forecasting models such as LSTM, which are sensitive to such temporal patterns.

Acknowledgment

This work was supported by Institut Teknologi Sepuluh Nopember (ITS) under Penelitian Keilmuan, Penelitian Dana Departemen Program, Penelitian Flagship, Riset Kolaborasi Indonesia, and under the Project Scheme of the Publication Writing and Intellectual Property Rights (IPR) Incentive (Penulisan Publikasi dan Hak Kekayaan Intelektual/PPHKI) Program 2025

References

[1] N. Sylvia, W. Rinaldi, A. Muslim, H. Husin, and Yunardi, "Challenges and possibilities of implementing sustainable palm oil industry in Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 969, no. 1, 2022, doi: 10.1088/1755-1315/969/1/012011.

[2] R. R. M. Paterson, "Oil palm survival under climate change in Kalimantan and alternative SE Asian palm oil countries with future basal stem rot assessments," *For. Pathol.*, vol. 50, no. 4, p. e12604, Aug. 2020, doi: 10.1111/EFP.12604.

[3] Y. Shigetomi, Y. Ishimura, and Y. Yamamoto, "Trends in global dependency on the Indonesian palm oil and resultant environmental impacts," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-77458-4.

[4] H. Ahn, K. Sun, and K. P. Kim, "Comparison of missing data imputation methods in time series forecasting," *Comput. Mater. Contin.*, vol. 70, no. 1, pp. 767–779, 2021, doi: 10.32604/cmc.2022.019369.

[5] M. D. Samad, S. Abrar, and N. Diawara, "Missing value estimation using clustering and deep learning within multiple imputation framework," *Knowledge-Based Syst.*, vol. 249, p. 108968, 2022, doi: 10.1016/J.KNOSYS.2022.108968.

[6] K. Alnowaiser, "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model," *IEEE Access*, vol. 12, no. February, pp. 16783–16793, 2024, doi: 10.1109/ACCESS.2024.3359760.

[7] A. T. Haryono, R. Sarno, R. N. E. Anggraini, and K. R. Sungkono, "Imputation Missing Stock Prices Using Generative Adversarial Networks and Attention Mechanism," *Proc. - 2024 2nd Int. Conf. Technol. Innov. Its Appl. ICTIIA 2024*, 2024, doi: 10.1109/ICTIIA61827.2024.10761233.

[8] J. Yoon, J. Jordon, and M. Van Der Schaar, "GAIN: Missing Data Imputation using Generative Adversarial Nets," *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 13, pp. 9042–9051, Jun. 2018, Accessed: May 12, 2025. [Online]. Available: https://arxiv.org/pdf/1806.02920

[9] Maimunah, J. L. Buliali, and A. Saikhu, "A Statistical-Temporal Framework for Evaluating Missing Value Imputation on Daily Waste Data," *ICADEIS 2025 - 2025 Int. Conf. Adv. Data Sci. E-learning Inf. Syst. Integr. Data Sci. Inf. Syst. Proceeding*, 2025, doi: 10.1109/ICADEIS65852.2025.10933416.

[10] H. Qian, Y. Geng, H. Wang, X. Wu, and M. Li, "LWGAIN: An Improved Missing Data Imputation Method Based on Generative Adversarial Imputation Network," *2024 5th Int. Conf. Comput. Vision, Image Deep Learn. CVIDL 2024*, pp. 835–839, 2024, doi: 10.1109/CVIDL62147.2024.10603600.

[11] P. Lian, Q. Zhao, and Y. Cui, "An Imputation Method Based on Dummy Variable and Unsupervised Learning for Electricity Consumption Data with Missing Values," *2021 IEEE 2nd China Int. Youth Conf. Electr. Eng. CIYCEE 2021*, 2021, doi: 10.1109/CIYCEE53554.2021.9676809.

[12] I. N. G. A. M. Wardhiana, M. Z. Ghufron, R. Sarno, and A. T. Haryono, "Price Forecasting of West Java Rice using Multivariate Decomposition SARIMAX-Gated Recurrent Unit Combination," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 1, pp. 769–778, 2025, doi: 10.22266/IJIES2025.0229.54.

[13] G. E. P. Box and M. E. Muller, "An Inverse Method for The Generation of Random Normal Deviates on Large-Scale Computers," *Math. Tables Other Aids to Comput.*, vol. 12, no. 63, p. 167, 1958, doi: 10.2307/2002017.

[14] B. S. Kwon, R. J. Park, and K. Bin Song, "Short-Term Load Forecasting Based on Deep Neural Networks Using LSTM Layer," *J. Electr. Eng. Technol.*, vol. 15, no. 4, pp. 1501–1509, 2020, doi: 10.1007/s42835-020-00424-7.

[15] G. G. S. Putra, A. W. C. D'Layla, D. Wahono, R. Sarno, and A. T. Haryono, "American Sign Language to Text Translation Using Transformer and Seq2Seq with LSTM," *Proc. - 2024 2nd Int. Conf. Technol. Innov. Its Appl. ICTIIA 2024*, 2024, doi: 10.1109/ICTIIA61827.2024.10761837.

[16] L. Muflikhah *et al.*, "Single nucleotide polymorphism based on hypertension potential risk prediction using LSTM with Adam optimizer," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 2, pp. 1126–1139, 2024, doi: 10.11591/ijeecs.v33.i2.pp1126-1139.

[17] W. Farsal, M. Ramdani, and S. Anter, "An Effective Random Forest Approach for Mining Graded Multi-Label Data," *Int. J. Intell. Eng. Syst.*, vol. 16, no. 4, pp. 548–557, Aug. 2023, doi: 10.22266/IJIES2023.0831.44.

[18] I. N. G. A. M. Wardhiana, M. Z. Ghufron, A. T. Haryono, R. Sarno, S. I. Sabilla, F. Taufany, Sholiq, and K. R. Sungkono "Comparative Study of Statistical, Machine Learning, and Deep Learning for Rice Retail Price Forecasting in West Java," *Proc. - 2024 2nd Int. Conf. Technol. Innov. Its Appl. ICTIIA 2024*, 2024, doi: 10.1109/ICTIIA61827.2024.10761587.

[19] R. A. Putri, D. Sunaryono, R. Sarno, S. S. Lee, H. Rahman, G. E. S. Setiawan, T. C. Amri, A. T. Haryono, and F. Taufany, "Lead (Pb) Contaminant Measurement on Distilled Water Using Hybrid Sensors and Linear Regression," *2024 Beyond Technol. Summit Informatics Int. Conf. BTS-I2C 2024*, pp. 274–279, 2024, doi: 10.1109/BTS-I2C63534.2024.10942213.