JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 10 (ISSUE 1) JUNE 2026: 90-101
https://doi.org/10.21009/JSA.10108

OPTIMIZING UNIVARIATE TIME SERIES IMPUTATION USING RANDOM FOREST REGRESSION AND LSTM FOR ACCURATE FORECASTING

Maulana Baihaqi Ramadhan[*1], Emli Rahmi[2], Isran K. Hasan[3]

[1,3] Department of Statistics, Faculty of Mathematics and Natural Sciences, Gorontalo State University
[2] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Gorontalo State University
Jl. Prof. Dr. Ing. BJ. Habibie, Moutong, Bone Bolango, 96554, Gorontalo

Corresponding author's e-mail: [*] **baihaqimaulana82@gmail.com**

ABSTRACT

Indonesia possesses high solar radiation potential, making solar energy a strategic pillar for the national clean energy transition. However, its utilization is hindered by incomplete data due to instrument failure, which significantly reduces prediction accuracy. Starting from this problem, this study aims to evaluate the performance of the Machine Learning Based Univariate Time Series Imputation-Random Forest Regression (MLBUI-RFR) method by comparing it with the Mean Imputation method and evaluating it through Long Short-Term Memory (LSTM) forecasting. The methodology begins with data preprocessing using the MLBUI-RFR scheme to handle missing values, which are then used as input for the LSTM architecture to forecast solar radiation in Indonesia. The findings demonstrate that the use of the MLBUI-RFR method contributes significantly to improving data quality, where the LSTM model trained with MLBUI-RFR imputed data achieves higher accuracy compared to Mean Imputation. The evaluation results show a lower error rate, with an NRMSE of (15.68%) and a MAPE of (18.98%), whereas the Mean Imputation method yields an NRMSE of (16.06%) and a MAPE of (19.15%) proving that the proposed method is more effective in capturing non-linear patterns in the data. However, this study is based exclusively on data obtained from the Gorontalo Climatology Station in Gorontalo Province, Indonesia. The contribution of this study lies in evaluating the integration of MLBUI-RFR and LSTM for solar radiation forecasting, demonstrating how machine learning based univariate time series imputation can improve data quality and subsequently enhance forecasting performance on solar radiation data.

Article History:

Received: May 29, 2026
Revised: June 20, 2026
Accepted: June 26, 2026
Published: June 30, 2026
Available online.

Keywords:

Deep Learning;
Forecasting;
Imputation;
Machine Learning;
Solar Radiation.

How to cite this article:
Ramadhan, M. B., Rahmi, E., Hasan, I. K., "OPTIMIZING UNIVARIATE TIME SERIES IMPUTATION USING RANDOM FOREST REGRESSION AND LSTM FOR ACCURATE FORECASTING", Journal Statistika dan Aplikasinya, vol. 10, iss. 1, pp. 90 – 101, June 2026

Journal homepage: *https://journal.unj.ac.id/unj/index.php/statistika*
Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Indonesia receives high levels of solar radiation throughout the year [1]. These favorable geographical conditions make solar energy a highly strategic pillar in driving the national transition toward clean and sustainable energy in the future [2]. In line with national energy policy, the Indonesian government has set a target for new and renewable energy to reach 23% by 2025, with a further increase to 31% by 2050 [3]. Therefore, accelerating the development of solar infrastructure is a crucial step to support the sustainability of the national power grid and meet global environmental commitments [4]. Despite its immense potential, solar radiation data exhibits highly fluctuating patterns that are significantly influenced by local atmospheric and weather conditions [5]. This inherent variability introduces substantial uncertainty in solar power production, making it difficult to maintain grid stability. Therefore, a deeper understanding of radiation variation patterns through accurate forecasting is essential to support more effective solar energy planning and management in the future [6]-[7].

Forecasting is the process of predicting future conditions by utilizing historical data patterns to enable optimal planning [8]-[9]. For solar radiation, accurate forecasting is crucial yet challenging because the data is non-linear and influenced by both short-term and long-term weather variability [10]. In a time series structure, solar radiation data values are interdependent across time steps, a phenomenon known as temporal dependence [11]. To capture these complex relationships, advanced computational frameworks are essential [12]. Although various machine learning approaches have been applied due to their ability to map non-linear behavior [13], these approaches have inherent limitations in capturing long-term temporal dynamics [14]. To address this, deep learning offers specialized architectures such as Long Short-Term Memory (LSTM) networks [10]. As an extension of Recurrent Neural Networks (RNNs), LSTMs effectively process long-term temporal trends and seasonal variations without suffering from the vanishing gradient problem [15],[16],[17]. However, a key prerequisite for LSTMs is that the training data must be complete. The presence of missing values severely disrupts the patterns of time series sequences and reduces the model's ability to effectively learn temporal dependencies [18]. Hence, robust data imputation is a mandatory preprocessing phase to ensure sufficient data quality [19]. Traditional techniques like mean imputation are straightforward but flawed, as they smooth out data variance and disregard historical fluctuations [20]-[21]. Alternatively, Machine Learning-based Univariate Time Series Imputation (MLBUI) addresses this by restructuring univariate series into multivariate sequences to perform forward and backward regressions using Random Forest and Support Vector Regression [22]-[23].

Empirical literature confirms the individual strengths of these methods. According to [10] and [24] demonstrated LSTM's superior capability in handling non-linear, seasonal meteorological data, while [23] verified that MLBUI outperforms conventional interpolation and filtering frameworks. However, previous research has separated data imputation from deep learning-based forecasting modeling, treating the two as distinct fields. Although various imputation techniques exist, no study has explicitly explored how MLBUI-RFR imputation affects LSTM-based solar radiation forecasting, particularly when using local datasets such as Indonesian solar radiation data. This separation creates a significant gap, as unaddressed missing values can severely disrupt forecasting systems. To bridge this critical gap, this study develops an innovative integrated framework that combines advanced data cleaning with deeply learning-based time series forecasting to deliver more reliable results focused on the local context. The analysis utilizes daily solar radiation data collected from the Gorontalo Climatology Station, Gorontalo Province, Indonesia, covering the period from 2021 to 2025. This dataset was selected because solar radiation in tropical regions such as Gorontalo is strongly influenced by local atmospheric variability, including cloud cover, rainfall patterns, and seasonal weather fluctuations, making it a suitable case study for evaluating imputation and forecasting methods. The comparative performance is evaluated using Mean Absolute Percentage Error (MAPE) and Normalized Root Mean Square Error (NRMSE) metrics.

2. METHODS

Material and Data

The data used in this study are secondary data in the form of daily solar radiation intensity for the period 2021–2025, obtained from the Gorontalo Climatology Station, comprising a total of 1826 daily observations. The dataset contained 107 naturally occurring missing observations resulting from data collection and recording issues at the Gorontalo Climatology Station. No artificial missing-data mechanism (e.g., MCAR, MAR, or MNAR simulation) was introduced in this study. The imputation methods were applied directly to these naturally occurring missing values prior to the forecasting stage.

Research Method

Machine Learning Based Univariate Time Series Imputation-Random Forest Regression

Machine Learning based Univariate Time Series Imputation (MLBUI), introduced by [23], is an approach designed to handle missing values in univariate time series data by transforming the segments before and after a gap into multivariate structures through forward and backward conversion. The missing values are then estimated using machine learning through both forward and backward predictions, which are subsequently aggregated via an ensemble mechanism to enhance stability and accuracy. Within this framework, MLBUI is highly compatible with Random Forest Regression (RFR), an ensemble learning method that aggregates multiple decision trees to produce robust predictions [25]. The integration of RFR within MLBUI ensures highly accurate imputation performance by preserving the underlying temporal dynamics and non-linear characteristics of the time series data [23]. The stages of the MLBUI method consist of 5 phases, namely:

1. Data Segmentation (Dividing Data)

The initial phase involves isolating each data gap of size T into two independent segments. The first segment, D_b, contains all data preceding the gap, while the second segment, D_a, contains the data following it.

Data after gap:

$$Da = X[N : t + T] \tag{1}$$

Data before gap:

$$Db = X[1 : t - 1] \tag{2}$$

Where X is the incomplete series, t is the starting index of the gap, and T is the gap size.

2. Data Transformation

In this phase, each univariate segment is transformed into a multivariate structure with $(T + 1)$ dimensions. This process uses T preceding values to estimate the next value (Forward Converting for D_b) and T succeeding values to estimate the previous value (Backward Converting for D_a).

3. Model Training

The transformed datasets are used to train the Random Forest Regression (RFR) models. RFR functions by aggregating the predictions of q decision trees built on bootstrap samples.

Forward Prediction (for D_b):

$$\widehat{f_b}(X) = \frac{1}{q}\sum_{t=1}^{q}.\hat{h}\left(X, S_{\theta_b^t}\right) \tag{3}$$

Backward Prediction (for D_a):

$$\widehat{f_a}(X) = \frac{1}{q}\sum_{t=1}^{q}.\hat{h}\left(X, S_{\theta_a^t}\right) \quad (4)$$

Where $\hat{h}$ represents the individual tree prediction, S is the transformed dataset, and q is the total number of trees.

4. Multi-step Forecasting

The trained model will estimate missing data by predicting it one step at a time. This model uses a recursive method, meaning that the prediction at t time step serves as the basis for predicting the next $t+1$ time step.

5. Data Completion

In the final step, the two different prediction sequences from the forward and backward models are combined. The final imputed value for each point in the gap is determined by calculating the average of these two results, thereby producing a stable and balanced estimate.

Mean Imputation

Mean Imputation (MI) is one of the simplest and most frequently used techniques to handle missing values. The fundamental principle of this method is to replace missing data points in a dataset with the average value of all available observations. Mean Imputation can be mathematically expressed using Equation below [26]:

$$\widehat{x_t} = \sum_{i:x_t \in C_k}.\frac{x_t}{n_k} \quad (5)$$

Where $\widehat{x_t}$ is the imputed value at time period t, x_t is the available data value at time period t, Σ denotes the summation operation, C_k represents all data points within a single variable, and n_k i denotes the total count of available.

Long Short Term Memory

Long Short-Term Memory (LSTM) is an advanced variant of the Recurrent Neural Network (RNN) designed to process both short- and long-term dependencies using an internal memory cell mechanism [27]. Introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997, LSTM resolves the vanishing gradient problem in standard RNNs, where gradients become too small to effectively update network weights during training [28]-[29]. LSTMs are highly effective for time series analysis due to their ability to selectively retain and discard information [30].

The architecture of LSTM outputs two main states: the hidden state (passed to the next time step and subsequent layers) and the cell state (acting as the internal memory for long-term information). The internal operations are driven by three gates and a candidate cell state [31].

1. Forget Gate

The forget gate determines which information from the previous cell state should be discarded or retained using a sigmoid activation function, yielding values between 0 and 1 [29], [32]:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (6)$$

Where f_t is the forget gate, σ is the sigmoid activation function, x_t is the input, h_{t-1} is the hidden state from time $t-1$, W_f is the weight, and b_f is the bias.

2. Input Gate

The input gate determines which new information to store in the cell state. It consists of a sigmoid gate and a hyperbolic tangent (tanh) candidate state:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (7)$$

$$\widetilde{C_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (8)$$

Where i_t is the input gate, $\widetilde{C_t}$ is the candidate cell state, W and b represent the respective weights and biases.

3. Cell State Update

The cell state is updated by combining the outputs of the forget and input gates [33] :

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \quad (9)$$

Where C_t is the new cell state, and C_{t-1} is the previous cell state.

4. Output Gate

The output gate decides what information should be output from the cell state and updates the hidden state [32]:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (10)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (11)$$

Where o_t is the output gate, W_o and b_o are its weights and biases, and h_t is the updated hidden state.

3. RESULTS

Descriptive analysis is conducted as an initial step prior to the solar radiation training data processing phase. This phase is designed to offer a comprehensive insight into the data characteristics and patterns, ensuring that the applied analytical methods are well-suited to the nature of the observed data. This analysis utilizes daily solar radiation data from Gorontalo Province covering the period from January 1, 2021, to December 31, 2025. A summary of the descriptive statistics is presented in Table 1.

Table 1. Descriptive Statistics

N	Mean	Std	Min	Max	Missing Value
1,826	194.80	56.74	51	534	107

Based on the descriptive analysis of the daily solar radiation data from January 1, 2021, to December 31, 2025, the dataset exhibits significant variability, making it highly suitable for time-series forecasting. Across the 1,826 total observations, there are 107 missing values identified. The data indicates a mean radiation of 194.80 W/m^2 with a standard deviation of 56.74, along with a wide range from a minimum of 51 W/m^2 to a maximum of 534 W/m^2. The high fluctuation and broad span of these values are crucial for modeling, as they allow the forecasting model to comprehensively learn and capture complex change patterns over time.

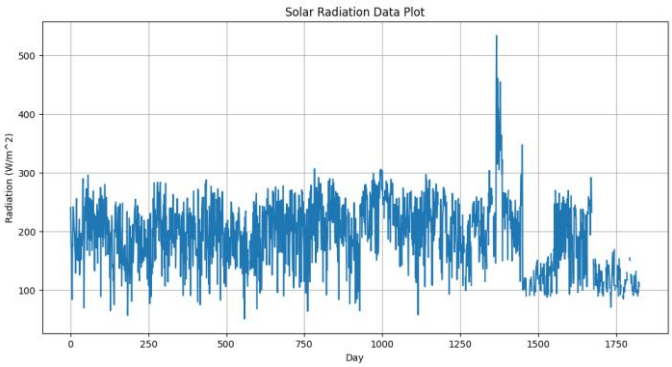

Figure 1. Solar Radiation Data Plot

As depicted in Figure 1, the solar radiation time series displays a high degree of daily variability and volatility, with most values fluctuating dynamically between 100 and 300 W/m^2. An extreme spike exceeding 500 W/m^2 is observed in days 1,368, which likely indicates a weather anomaly or unusually clear sky conditions during this specific timeframe. To address the missing values, the data handling process in this study utilizes two different approaches, the MLBUI-RFR method and Mean Imputation as a comparative baseline. The implementation of these methods ensures the quality of the time series dataset before proceeding to the forecasting stage. A comparison of the imputation results between the MLBUI-RFR and Mean Imputation methods is presented in Figure 2, which illustrate how effectively each technique preserves the underlying temporal structure and distribution of the solar radiation data.

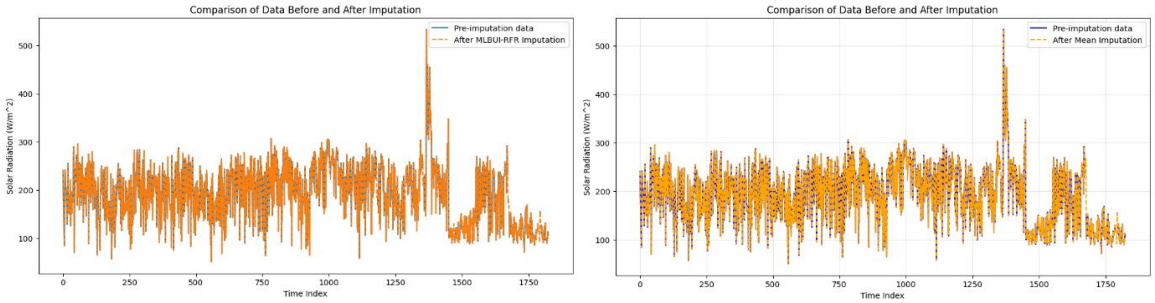

Figure 2. Comparison of the MLBUI and MI Imputation Methods

The figures 2 illustrate the comparison of the solar radiation data distribution before and after imputation using both the MLBUI-RFR and Mean Imputation methods. Overall, both graphs demonstrate that the imputed data, represented by the orange dashed line, effectively fills the missing values without distorting the primary structure or long-term trends of the original baseline data. The MLBUI-RFR approach produces imputed values that are more in line with the local fluctuations of the surrounding data, thanks to the predictive nature of the algorithm. Conversely, the Mean Imputation technique tends to pull the missing values toward the global mean, which results in a lower variance. Ultimately, these plots confirm that both techniques successfully preserve the continuity of the time-series data, which is essential for maintaining model stability and performance during the subsequent forecasting phase.

Following this data preparation stage, the LSTM model was implemented using the Python programming language with the normalized dataset as its input. At this phase, the LSTM model was constructed using 96 LSTM units, a batch size of 32, and 365 timesteps. These parameters were selected through a trial and error approach by testing various combinations of LSTM units, batch sizes, and timesteps. Each configuration was evaluated based on the model's performance on the validation data, utilizing loss values and error metrics such as MAPE. The test results indicated that the combination of 96 LSTM units, a batch size of 32, and 365 timestep provided the best performance, characterized by lower error rates and greater stability throughout the training process.

Furthermore, the 365 timestep parameter was explicitly chosen to represent the annual seasonal patterns within the solar radiation data. The model was then trained using the Adam optimizer in accordance with the research methodology.

Following the development of LSTM model, the next phase involved evaluating the model's performance in learning the underlying patterns within the solar radiation data. The training process was conducted over 50 epoch, allowing the model to iteratively adjust its weights and parameters to minimize the error. Based on the training outcomes, when using the data imputed with the MLBUI-RFR method, the loss steadily decreased from 0.0255 in the first epoch to 0.0090 by the 50 epoch. A similar consistent decline was observed for the dataset imputed using Mean Imputation, where the loss fell from 0.0257 at epoch 1 to 0.0090 at epoch 50. This steady reduction demonstrates an enhanced proficiency of the model in capturing data patterns over time. Additionally, the validation loss maintained a relatively stable trend without any significant increases, indicating that the model did not suffer from notable overfitting. Therefore, the LSTM model proves capable of capturing the nonlinear relationships within the solar radiation data. The solar radiation forecasting results are visually organized within table 2.

Table 2. Comparison of MLBUI and MI Forecast Results

Date	Actual Data		Forecasting Result Data		Difference	
	MLBUI	MI	MLBUI	MI	MLBUI	MI
3/14/2025	157	157	124.270554	122.924034	32.729446	34.075966
3/15/2025	138	123	130.804489	129.46611	7.195511	-6.466110
3/16/2025	132	118	134.946686	131.169571	-2.946686	-13.169571
3/17/2025	125	125	136.478195	130.395691	-11.478195	-5.395691
3/18/2025	134	134	135.779129	130.304688	-1.779129	3.695312
⋮	⋮	⋮	⋮	⋮	⋮	⋮
12/27/2025	98	98	110.935379	110.10894	-12.935379	-12.108940
12/28/2025	96	96	110.203911	109.317535	-14.203911	-13.317535
12/29/2025	114	114	109.711662	108.77951	4.288338	5.22049
12/30/2025	112	112	112.182816	111.198097	-0.182816	0.801903
12/31/2025	107	107	114.626488	113.645737	-7.626488	-6.645737

Based on the table 2, the model utilizing the MLBUI-imputed data demonstrates superior performance compared to the one using the Mean Imputation (MI) method. This advantage is evident from the residuals in the MLBUI column, which are generally smaller and closer to zero at several observation points such as the samples on March 18 and December 30, 2025 than those in the MI column. This indicates that the MLBUI effectively preserves the original data's characteristics and variability patterns. As a result, it provides a more reliable foundation for the forecasting model, enabling more accurate forecasts that are highly responsive to fluctuations in solar radiation.

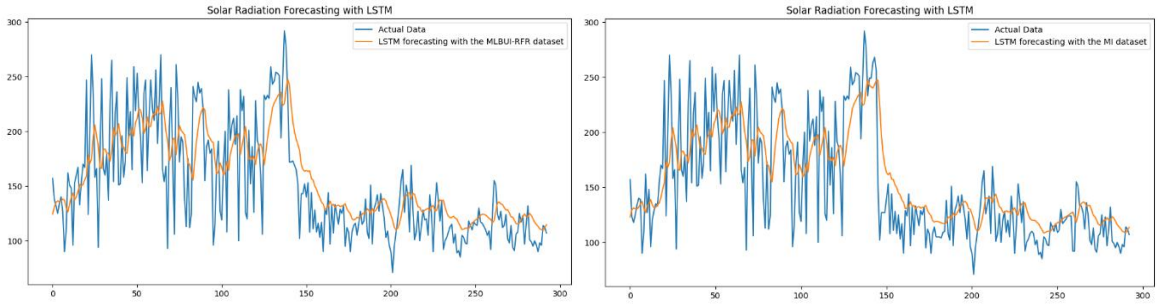

Figure 3. Comparison of the MLBUI and MI Forecasting

Figure 3 shows a visualization of the solar radiation forecasting results using the LSTM model built on the MLBUI-RFR and Mean Imputation datasets, respectively. Visually, the model utilizing

the MLBUI-RFR dataset (shown on the left) demonstrates superior performance, as the predicted line (orange) tracks the sharp fluctuations of the actual data (blue) with greater responsiveness and precision, particularly at the radiation peaks prior to index 150. Conversely, the predictions on the Mean Imputation dataset (shown on the right) appear smoother and exhibit a lag in responding to abrupt changes, leading to larger residual errors. The advantage of the MLBUI-RFR method indicates that machine learning based imputation delivers a richer representation of feature patterns, which directly enhances the generalization ability of the LSTM model in forecasting volatile atmospheric phenomena.

Based on these findings, the trained model was used to forecast solar radiation for the next 10 days. After undergoing training and evaluation, the model was applied to forecast values for future, unobserved periods. The forecasting process was performed sequentially using the last data point from the dataset as the initial input, allowing the model to forecast solar radiation values for each subsequent day until reaching a 10 day forecast horizon. The results of this 10 day solar radiation forecast are detailed thoroughly within Table 3.

Table 3. Comparison of 10 Day Forecast Results Using the MLBUI and MI Datasets

Date	MLBUI-RFR Forecasting	MI Forecasting
01/01/2026	115.669189	114.736435
01/02/2026	117.454788	116.411575
01/03/2026	119.329620	118.156235
01/04/2026	121.119896	119.831909
01/05/2026	122.794945	121.414696
01/06/2026	124.366982	122.915138
01/07/2026	125.858513	124.351913
01/08/2026	127.291832	125.743332
01/09/2026	128.685852	127.104752
01/10/2026	130.055374	128.448013

Table 3 presents the results of solar radiation forecasts for the next 10 days, from January 1, 2026, to January 10, 2026, comparing forecasts from the LSTM model when applied to data imputed using the MLBUI-RFR and Mean Imputation (MI) methods. As shown in the table, both approaches predict a steady upward trend in daily solar radiation, increasing from approximately 114.74 W/m^2 and 115.66 W/m^2 at the start of the period to 128.45 W/m^2 and 130.05 W/m^2 on the tenth day. Furthermore, the values estimated by the MLBUI-RFR approach are consistently higher than those obtained from the MI method throughout the forecast period. This difference reflects the model's ability to capture the dynamic variations of the original dataset more effectively, indicating that machine learning-based imputation techniques provide a more responsive and accurate foundation for predicting future atmospheric conditions. To evaluate the forecasting performance of the LSTM models trained on different imputed datasets, Mean Absolute Percentage Error (MAPE) and Normalized Root Mean Square Error (NRMSE) were calculated using the testing data. These metrics were employed to assess forecasting accuracy and were not intended to directly evaluate the imputation performance.

Table 4. Evaluation Metrics

MAPE (%)		Interpretation	NRMSE (%)		Interpretation
MLBUI	MI		MLBUI	MI	
18.98%	19.15%	Good	15.68%	16.06%	Good

Based on Table 4, the MAPE value was 18.98% for the model using the MLBUI-RFR method and 19.15% for the model using the Mean Imputation method both utilized LSTM networks for forecasting. These figures fall into the good category as they remain within the 10% to 20% range. Additionally, the NRMSE values were recorded at 15.68% for the MLBUI approach and 16.06% for the Mean Imputation approach. the MLBUI approach demonstrates slightly superior performance by

yielding lower MAPE and NRMSE values compared to the Mean Imputation approach. This indicates that the LSTM model integrated with MLBUI-RFR imputation achieves a higher level of accuracy, making it suitable for forecasting highly fluctuating solar radiation values.

4. DISCUSSIONS

The results indicate that the choice of imputation method influences the forecasting performance of the LSTM model. Based on Table 4, the LSTM model trained using MLBUI-RFR-imputed data achieved lower forecasting errors (MAPE = 18.98% and NRMSE = 15.68%) than the model trained using Mean Imputation data (MAPE = 19.15% and NRMSE = 16.06%). Although the performance difference is relatively small, the consistent reduction in both error metrics suggests that the quality of missing-value treatment contributes to the effectiveness of subsequent forecasting models.

The improved performance of MLBUI-RFR can be explained by the characteristics of Random Forest Regression (RFR), which serves as the predictive engine within the MLBUI framework. Unlike conventional statistical imputation techniques, RFR estimates missing values by learning patterns from neighboring observations through an ensemble of decision trees. This mechanism enables the model to capture complex non-linear relationships and local temporal dependencies that are commonly present in solar radiation data. Consequently, the imputed values are more representative of the underlying data-generating process and preserve important fluctuations surrounding the missing observations. As a result, the reconstructed time series retains more information relevant to forecasting, allowing the LSTM model to learn temporal patterns more effectively.

In contrast, Mean Imputation replaces all missing observations with a single average value derived from the available data. While this method is computationally simple, it ignores temporal dynamics and the contextual information surrounding each missing observation. Consequently, the variability of the series is reduced, and extreme values tend to be suppressed. This phenomenon is commonly referred to as a smoothing effect, whereby the imputed data become more homogeneous than the original observations. For highly variable meteorological data such as solar radiation, the loss of local variability may weaken the ability of LSTM to capture abrupt changes and nonlinear fluctuations, ultimately leading to less accurate forecasts.

The findings of this study are consistent with those reported by [23], who introduced the Machine Learning Based Univariate Time Series Imputation (MLBUI) framework and demonstrated its ability to preserve the characteristics of incomplete time-series data more effectively than conventional imputation approaches. Phan showed that machine learning-based imputation methods can exploit information embedded in neighboring observations to generate more realistic estimates of missing values. Similar evidence was observed in the present study, where the MLBUI-RFR approach produced lower forecasting errors than Mean Imputation. This consistency suggests that maintaining the intrinsic structure of time-series data during the imputation stage contributes positively to downstream predictive performance.

The results also emphasize the importance of considering data preprocessing as an integral component of forecasting systems. Previous studies have extensively examined LSTM for solar radiation forecasting and MLBUI for missing-value treatment as separate research topics. However, limited attention has been given to evaluating how imputation quality affects the predictive performance of deep learning models, particularly for solar radiation data from Indonesia. By assessing MLBUI-RFR and Mean Imputation within the same LSTM forecasting framework using daily solar radiation observations from the Gorontalo Climatology Station, this study provides empirical evidence that preserving temporal characteristics during the imputation stage can contribute to improved forecasting accuracy. Therefore, the findings highlight the interdependence between data quality and forecasting performance in deep learning-based time-series applications.

5. CONCLUSION

This study was conducted to address two main objectives related to the imputation and forecasting of solar radiation in Gorontalo Province. The first objective was to evaluate data preprocessing

approaches for handling missing values by comparing machine learning-based imputation methods with conventional mean imputation methods. The second objective focuses on the development and evaluation of a LSTM deep learning model to forecast solar radiation patterns in tropical regions.

Regarding the first objective, the analysis results demonstrate that machine learning-based imputation approaches are far superior to conventional mean-based methods in preserving data authenticity and reducing bias. This preprocessing step preserves the diversity and extreme values in the time-series data without losing important details. Consequently, this step successfully prevents the propagation of errors into subsequent forecasting stages.

In line with the second objective, the research results demonstrate that LSTM model can capture the complexity and temporal characteristics of solar radiation with high precision. This model not only minimizes prediction errors and learns from past data fluctuations but also effectively responds to seasonal variations. Furthermore, by addressing the vanishing gradient problem, this architecture has proven consistent in generating radiation estimates that closely match actual data.

Based on these results, this study provides in-depth insights into how the integration of machine learning imputation methods and deep learning forecasting models responds to weather dynamics in tropical regions. These findings have significant practical implications for the region, which can serve as a strategic foundation for government policymakers and the energy sector to guide the deployment of large-scale solar infrastructure, enhance smart grid stability, and support the renewable energy transition.

6. ACKNOWLEDGMENTS

Deepest gratitude is extended by the author to the Statistics Study Program within the Faculty of Mathematics and Natural Sciences. Universitas Negeri Gorontalo (UNG), for providing academic resources and support during the research process. Furthermore, the authors extend their deepest appreciation to the National Amil Zakat Agency (BAZNAS) for the educational scholarship provided through the Beasiswa Cendekia BAZNAS (BCB), which greatly supported and facilitated the completion of this study.

7. REFERENCES

[1] F. Afif, “Tinjauan Potensi dan Kebijakan Energi Surya di Indonesia,” *J. Engine Energi, Manufaktur, dan Mater.*, vol. 6, no. 1, pp. 43–52, 2022.

[2] L. Satria, “Pemerintah targetkan 1 GW energi surya pada 2025, perkuat transisi energi nasional,” 2025. [Online]. Available: https://esgnow.republika.co.id/berita/t1ya9p423/pemerintah-targetkan-1-gw-energi-surya-pada-2025-perkuat-transisi-energi-nasional

[3] A. Y. Salile, S. Nisworo, and S. Sumardi, “Analisis Fluktuasi Radiasi Matahari dan Implikasinya Terhadap Penempatan PLTS,” *J. Profesi Ins. Indones.*, vol. 2, no. 6, pp. 354–358, 2024.

[4] F. Luthfianingsih, S. N. Asyari, T. Y. Sidabutar, M. A. Fitri, and U. Kamal, “Strategi dan Implementasi Net Zero Emission melalui PLTS Terapung Cirata Menuju Target Indonesia 2060,” *J. Intelek Insa. Cendikia*, vol. 2, no. 6, pp. 11699–11708, 2025.

[5] A. Asrori and E. Yudiyanto, “Kajian Karakteristik Temperatur Permukaan Panel terhadap Performansi Instalasi Panel Surya Tipe Mono dan Polikristal,” *FLYWHEEL J. Tek. Mesin Untirta*, vol. 1, no. 1, pp. 68–73, 2019.

[6] L.-E. Ordoñez-Palacios, V.-A. Bucheli-Guerrero Ph, H.-A. Ordoñez-Eraso Ph, D.-A. León-Vargas M Sc, and *et al.*, “Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático,” *Rev. Fac. Ing.*, vol. 29, no. 54, 2020.

[7] J. Ramirez-Vergara, L. B. Bosman, E. Wollega, and W. D. Leon-Salas, “Review of forecasting methods to support photovoltaic predictive maintenance,” *Clean. Eng. Technol.*, vol. 8, 2022, doi: 10.1016/j.clet.2022.100460.

[8] F. S. Jamil and F. Faisol, “Peramalan Hasil Penjualan Sandal Menggunakan Metode Kalman Filter,” *Zeta-Math J.*, vol. 2, no. 2, pp. 37–40, 2016.

[9] B. Li, C. Yao, F. Zheng, L. Wang, J. Dai, and Q. Xiang, "RETRACTED ARTICLE: Intelligent Decision Support System for Business Forecasting Using Artificial Intelligence," *Arab. J. Sci. Eng.*, vol. 48, no. 3, 2023.
[10] D. P. Utomo, "Peramalan Penyinaran Matahari Per Jam Satu Hari Ke Depan Menggunakan Model Long Short-Term Memory (LSTM)," vol. 13, no. 3, pp. 876–886, 2025.
[11] A. Arwansyah, S. Suryani, H. S. Y. H. SY, A. Ahyuna, U. Usman, and S. Alam, "Model Prediksi Deret Waktu Menggunakan Deep Convolutional LSTM," *SISITI: Seminar Ilmiah Sistem Informasi dan Teknologi Informasi*, pp. 21–25, 2024.
[12] A. N. A. Mazmee, N. Zaini, and M. F. A. Latip, "Enhancing Solar Energy Forecasting Accuracy Using LSTM Networks for Global Horizontal Irradiance," *14th International Conference on System Engineering and Technology (ICSET)*, 2024, pp. 174–179.
[13] B. Mehmood *et al.*, "Development Of A Hybrid Artificial Intelligence Framework For Accurate Forecasting Of Solar Power Generation Using Machine Learning Algorithms And Time-Series Analysis," *Spectr. Eng. Sci.*, vol. 3, no. 5, pp. 613–636, 2025.
[14] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, vol. 148, pp. 461–468, 2018.
[15] R. Julian and M. R. Pribadi, "Peramalan Harga Saham Pertambangan Pada Bursa Efek Indonesia (BEI) Menggunakan Long Short Term Memory (LSTM)," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 3, pp. 1570–1580, 2021, doi: 10.35957/jatisi.v8i3.1159.
[16] M. Lim and T. Handhayani, "Penerapan lstm dan gru untuk prediksi harga cabai merah di kota jawa timur," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 2, 2025.
[17] T. Trinh, A. Dai, T. Luong, and Q. Le, "Learning longer-term dependencies in rnns with auxiliary losses," *International Conference on Machine Learning*, pp. 4965–4974, 2018.
[18] L. S. Hasibuan and Y. Novialdi, "Prediksi Harga Minyak Goreng Curah dan Kemasan Menggunakan Algoritme Long Short-Term Memory (LSTM)," *J. Ilmu Komput. dan Agri-Informatika*, vol. 9, no. 2, pp. 149–157, 2022.
[19] G. M. Yagli, D. Yang, O. Gandhi, and D. Srinivasan, "Can we justify producing univariate machine-learning forecasts with satellite-derived solar irradiance?," *Appl. Energy*, vol. 259, 2020, doi: 10.1016/j.apenergy.2019.114122.
[20] D. N. F. Murjito, "Perbandingan Metode Imputasi: Metode Mean Dan Metode K Nearest Neighbor (KNN) Untuk Mengatasi Data Hilang Pada Data Survei," 2021.
[21] D. Ramadhani, A. M. Soleh, and E. Erfiani, "Characteristics of Machine Learning-based Univariate Time Series Imputation Method," *JUITA J. Inform.*, vol. 12, no. 2, pp. 279–288, 2024.
[22] C. A. Keller and M. J. Evans, "Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10," *Geosci. Model Dev.*, vol. 12, no. 3, pp. 1209–1225, 2019, doi: 10.5194/gmd-12-1209-2019.
[23] T. T. H. Phan, "Machine Learning for Univariate Time Series Imputation," *2020 Int. Conf. Multimed. Anal. Pattern Recognition,* October 2020, doi: 10.1109/MAPR49794.2020.9237768.
[24] Y. Arsanti, L. O. A. Minsaris, and W. A. Arifin, "Perbandingan Model Prediksi Suhu Permukaan Laut Menggunakan Smoothing dan Long Short-Term Memory," *J. Algoritm.*, vol. 22, no. 1, pp. 1026–1038, 2025, doi: 10.33364/algoritma/v.22-1.2113.
[25] D. Manurung, B. Zealtiel, and A. H. Lubis, "Prediksi Produksi Tanaman Padi di Indonesia dengan Menggunakan Algoritma Random Forest Regressor," *J. Comput. Informatics Res.*, vol. 4, no. 3, pp. 337–345, 2025.
[26] A. Anggreni, U. Ardhita, N. Sulistianingsih, and A. Rahman, "Penanganan Missing Data Hujan Kabupaten Sumbawa," *Bakti Sekawan J. Pengabdi. Masy.*, vol. 5, no. 1, pp. 39–44, 2025.
[27] I. I. Zulfa, D. Candra, R. Novitasari, F. Setiawan, A. Fanani, and M. Hafiyusholeh, "Prediction of sea surface current velocity and direction using lstm," *Indones. J. Electron. Instrum. Syst*, vol. 11, no. 1, pp. 93–102, 2021.
[28] R. Rowan, L. Muflikhah, and I. Cholissodin, "Peramalan kasus positif covid-19 di jawa timur

menggunakan metode hybrid arima-lstm," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 9, pp. 4146–4153, 2022.

[29] E. F. Syahram, M. M. Effendy, and N. Setyawan, "Sun position forecasting menggunakan metode rnn--lstm sebagai referensi pengendalian daya solar cell," *SinarFe7*, vol. 3, no. 1, 2020.

[30] M. Hussein and Y. Azhar, "Prediksi Harga Minyak Dunia Dengan Metode Deep Learning," *Fountain Informatics J.*, vol. 6, no. 1, pp. 29–34, 2021.

[31] J. S. Prasetyo, "Stock Price Prediction Using Machine Learning With Long Short Therm Memory Method (LSTM)," *Kilat*, vol. 12, no. 1, pp. 64–78, 2023.

[32] L. Wiranda and M. Sadikin, "Penerapan long short term memory pada data time series untuk memprediksi penjualan produk PT. Metiska Farma," *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 8, no. 3, pp. 184–196, 2019.

[33] R. Akbar, R. Santoso, and B. Warsito, "Prediksi Tingkat Temperatur Kota Semarang Menggunakan Metode Long Short-Term Memory (Lstm)," *J. Gaussian*, vol. 11, no. 4, pp. 572–579, 2023.