ISSN 2355-3286

Solar Radiation Intensity Imputation in Pyranometer of Automatic Weather Station Based on Long Short Term Memory

Richat Pahlepi[1,3], Santoso Soekirno[1], Haryas Subyantara Wicaksana[2,4]
[1]Department of Physics, Faculty of Mathematics and Science, Universitas Indonesia, Depok, Indonesia
[2]Faculty of Industrial Technology, Institut Teknologi Bandung, Bandung, Indonesia
[3]Climatology Station of West Java, BMKG, Bogor, Indonesia
[4]Center of Instrumentation, Calibration and Engineering, BMKG, Jakarta, Indonesia
[1]richat.pahlepi@ui.ac.id, [2]santoso.s@sci.ui.ac.id, [3]haryas.wicaksana@bmkg.go.id

Accepted on 17 October 2023
Approved on 06 November 2023

***Abstract*— Automatic Weather Station (AWS) experienced problems in the form of component damage and communication system failure, resulting in incomplete parameter data. Component damage also occurs in pyranometers. Decreased pyranometer performance results in deviations, uncertainty in measuring solar radiation intensity, and data gaps. Data imputation is one solution to minimize measurement deviations and the occurrence of missing AWS pyranometer data. This research aims to design and analyze the accuracy performance of the multisite AWS pyranometer solar radiation intensity data imputation model when a data gap occurs. This research attempts to utilize the spatio-temporal relationship of multisite AWS solar radiation intensity in the imputation model. Long-Short Term Memory (LSTM) algorithm is used as an estimator in the multisite AWS pyranometer network. Data imputation modeling stage includes data collection, data pre-processing, creating missing data scenarios, LSTM design and model testing. Overall, LSTM-based imputation model has ability of filling gap data on AWS Cikancung pyranometer with maximum missing sequence of 12 hours. Imputation model has MAPE 1.76% - 5.26% for missing duration 30 minutes-12 hours. It still it meet World Meteorological Organization (WMO) requirement for solar radiation intensity measurement with mean absolute percentage error (MAPE) < 8%.**

***Index Terms*— imputation; pyranometer; Long Short Term Memory**

I. INTRODUCTION

Solar radiation is electromagnetic energy emitted by the sun as a result of nuclear fusion in the sun's core. Solar radiation is one of the elements of weather and climate. Solar radiation parameter measurements are mainly used in seasonal forecast modeling [1]. Apart from that, measuring this parameter is also useful for the energy, architecture, astronomy and environmental sectors. Almost all companies and manufacturers related to solar power generation use solar radiation measurement data to plan generating system mechanisms [2].

Shortwave solar radiation with a spectrum range from 290 to 3000 nm. One of automatic digital instruments for measuring shortwave solar radiation is pyranometer. Pyranometers use photodiode or thermophile type sensors [3]. This instrument is installed on the Automatic Weather Station (AWS) and Automatic Solar Radiation Station (ASRS) [4]. Photodiode type pyranometers use photodiodes to detect the intensity of sunlight. Photodiodes convert changes in light intensity values into changes in electric current values. Thermophile type pyranometer uses a thermophile detector. Thermophile element is able to convert changes in temperature values into changes in electrical voltage values. The thermophile consists of a black layer and a reference layer [5].

The decline in pyranometer performance is characterized by degradation changes in sensor sensitivity values. Degradation of pyranometer sensitivity is caused by drift of internal components, leveling shifts, changes in spectral response, presence of scratches, as well as installation environmental conditions [6]. Decreased pyranometer performance results in deviations, uncertainty in measuring solar radiation intensity, and data gaps. This reduces the accuracy of synoptic surface weather analyzes [7]. Imputation is able to minimize the occurrence of missing data on pyranometer.

Several previous studies have been carried out regarding the imputation of meteorological data including the intensity of solar radiation. Turrado et al. designed a model for estimating the intensity of solar radiation per 10 minutes using the Multivariate

Imputation by Chained Equation (MICE) algorithm based on the spatial links between 9 ground station pyranometers in Galicia, Spain [8]. This algorithm has an RMSE performance of 13.37% and is lower than multilinear regression and interpolation methods based on Inverse Distance Weighting (IDW).

Karaman et al. used Extreme Learning Machine (ELM) to estimate daily solar radiation with inputs including sun duration, air temperature and wind speed in Turkey [9]. This algorithm shows better performance than ANN with an RMSE of 0.0297 kJ/m^2. ELM also has a shorter computing time than ANN. Yamoah et al. applied the multilinear regression (MLR) method to impute hourly wind speed data. Regression predictor variables include temperature, humidity, air pressure and rainfall [10]. The multilinear regression method produces an average error of 2.38 m/s. This method has better performance than ARIMA and Kalman filter. The limitations of the methods are risks related to incomplete availability of multivariable inputs.

Previous studies have not involved spatial and temporal links simultaneously. This research attempts to utilize the spatio-temporal relationship of solar radiation intensity in the imputation model. Multisite AWS has advantages on imputes pyranometer data spatially. Long short term memory algorithm is chosen based on its capability of extracting long term temporal features on time series data [11].

II. DATA

Physical AWS pyranometer used for modeling is located in West Java Province. These sensors are spread across 3 AWS locations in Sumedang Regency and Tasikmalaya Regency. AWS locations include AWS Cipasung, AWS Cikancung and AWS Cimalaka. Figure 1 shows AWS pyranometer location map in the research area.

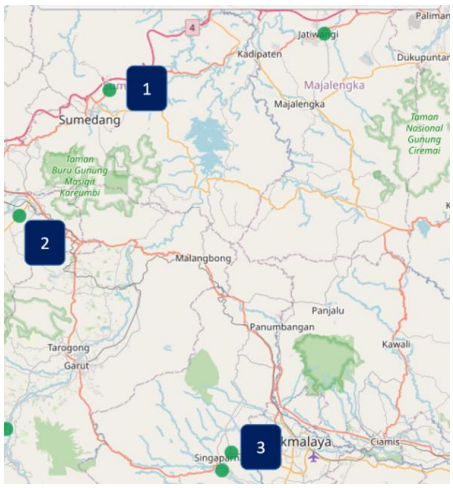

Fig. 1. Sites of AWS pyranometer

Figure 1 shows that all sites are covered adjacently area. Physical AWS pyranometer used for modeling is located in West Java Province. Table 1 describes site coordinates in more detail.

TABLE I. SITES COORDINATES

No	Site	Latitude	Longitude	Elevation
1	AWS Cikancung	-6,9988	107,8168	680 m
2	AWS Cimalaka	-6,8154	107,9475	545 m
3	AWS Cipasung	-7,3416	108,1280	417 m

Table 1 states study area based on label number of Figure 1 respectively. Variations in distance and elevation is possibly produce further analysis related to the results of spatial imputation model for the three pyranometer data [12]. AWS pyranometer data is generated and sent every 10 minutes. This data is collected from the BMKG Database Center via the BMKG AWS Center website. Three AWS pyranometers data was taken in the period 2021 – 2022.

III. METHODS

Imputation modeling stage includes literature study, data collection, data pre-processing, creating missing data scenarios, Long Short Term Memory (LSTM) design and model testing. Figure 2 shows imputation modeling flowchart.

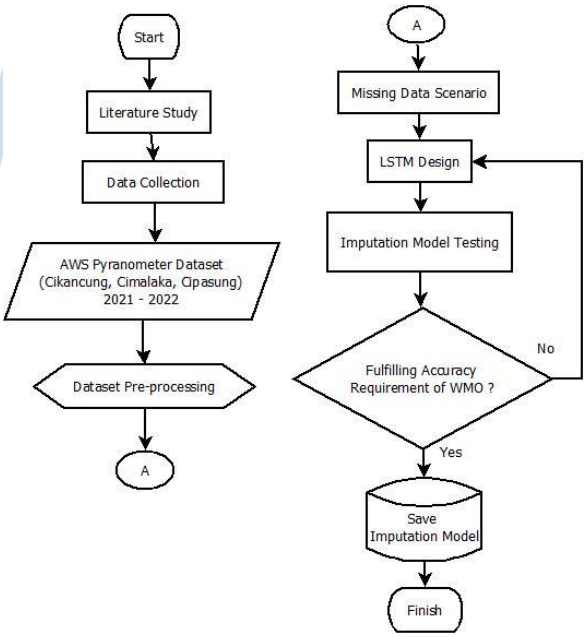

Fig. 2. Imputation Modeling Flowchart

Based on Figure 2, a literature study was carried out on literature reviews of previous research. Data collection is carried out by downloading AWS data on

Badan Meteotologi Klimatologi dan Geofisika (BMKG) Database Center page. Before the imputation model is designed, the dataset needs to be pre-processed to select well-qualified data. Missing data simulations were implemented on normal datasets. The imputation model is designed using LSTM algorithm, then tested on the missing data scenario that has been created. If the test imputation model does not meet the World Meteorological Organization (WMO) accuracy criteria, then the LSTM imputation model is redesigned until the accuracy criteria are achieved.

A. *Dataset Pre-processing*

Pre-processing functions to prepare modeling data. Data pre-processing was carried out to remove outliers in the AWS pyranometer raw data. Outliers are deviations from normal data values according to certain criteria.

Outliers are identified through data quality control (QC) based on range check and step check. Range check is checking data based on the historical range per parameter. If the data is not within the normal range, then it becomes suspect based on the range check. A normal range of solar radiation intensity data is 0-1500 W/m^2 [12].

Step check is a check based on the temporal relationship of current data to previous data. If the difference between the current data and previous data exceeds a certain threshold, then the data becomes suspect based on the step check [13]. Step check threshold value for solar radiation intensity data is that there is no change in the value exceeding 800 W/m^2 for a period of 5 minutes. Outliers are removed via a listwise deletion mechanism. Listwise deletion is the process of deleting all outlier data [14]. This mechanism is carried out at the three AWS pyranometer locations.

B. *Missing Data Scenario*

Missing data scenario is simulated to AWS Cikancung pyranometer data. AWS pyranometer dataset is only taken from 07.00 - 19.00 Local Time according to sunshine hours in Indonesia [15]. Pyranometer dataset on 2021-2022 is divided into two parts. The 2021 dataset is the imputation model data and the 2022 dataset is the test data. Missing data scenario is only carried out on the 2022 AWS pyranometer dataset. There are 5 missing data scenarios:

i. For every 3 days, pyranometer data are missing for 30 minutes respectively;
ii. For every 6 days, pyranometer data are missing for 60 minutes respectively;
iii. For every 18 days, pyranometer data are missing for 3 hours respectively;
iv. For every 36 days, pyranometer data are missing for 6 hours respectively;
v. For every 72 days, pyranometer data are missing for 12 hours respectively.

All of these missing data are then saved as actual data for testing stage. These scenarios produces balance number of missing data. It purposes to perform imputation model on vary missing data possibilities in one day.

C. *Imputation Model Based on LSTM*

Imputation techniques fill in empty data based on available data using certain methods [16]. LSTM is one method of imputation technique. LSTM has a memory cell in each neuron. LSTM model consists of connections between cells in three gates: input gate, forget gate and output gate. Forget gate has a role to regulate flow of information, so that LSTM has management of cell memory [17]. Table II shows LSTM algorithm.

TABLE II. LSTM ALGORITHM

Input :	Input (x_t), previous cell state (c_{t-1}), and previous hidden state (s_{t-1})
Output :	Current cell state (c_t) and current hidden state (s_t)
1	Initialization x,c_{t-1},s_{t-1} -> ReLU(x) = max(0,1)
2	n = epoch
3	For i = 0,1,….n, do:
4	$f_t = \sigma(W_f.[s_{t-1}, x_t] + b_f)$
5	$i_t = \sigma(W_f.[s_{t-1}, x_t] + b_i)$
6	$c_t = \tanh(W_c.[s_{t-1}, x_t] + b_c)$
7	$c_t = f_t * c_{t-1} + i_t * c_t$
8	$o_t = \sigma(W_o.[s_{t-1}, x_t] + b_o)$
9	$s_t = o_t * \tanh(c_t)$
10	Calculate RMSE of c_t and c_{actual}
11	End of iteration

Design of LSTM-based imputation model aims to determine the structural configuration of LSTM algorithm. This design consists of preparation stages, data transformation, data segmentation, hyperparameter design, training and model validation. Model design uses the 2021 dataset.

Preparation stage for LSTM design is determining input and output of the model. Imputation of AWS pyranometer data is carried out spatially. LSTM input is data from two pyranometers around target pyranometer: AWS Cipasung and AWS Cimalaka. Meanwhile, LSTM output is data from one target pyranometer: AWS Cikancung.

Input and output data are then transformed. Data transformation functions to simplify computational learning. Input and output are transformed into variables on a smaller scale range. Transformation equation is stated as follows [18]:

$$Z_{norm} = \frac{z_i - z}{stdev(z)} \quad (1)$$

Each input and output value (Z_i) is reduced to its average value (Z), then divided by its standard deviation, to obtain a new variable as a result of the transformation. Next, the input and output data of the 2021 dataset are segmented into 80% training data and 20% validation data. Testing data are taken from the 2022 dataset based on missing data scenarios.

LSTM hyperparameter design is adopted from previous literature. In 2022, Parasyris et.al. developed an LSTM model for estimating temperature, humidity and wind speed. This time, the model is applied to impute solar radiation intensity of pyranometer [11]. Table III shows hyperparameter design of LSTM for imputation model

TABLE III. LSTM HYPERPARAMETER DESIGN

Hyperparameter	Value
Hidden Layer	2
Neuron Number	[40,40]
Activation Function	Rectified Linear Unit
Batch Size	32
Epoch	100
Number of inputs	2 (pyranometers of AWS Cimalaka and AWS Cipasung)
Number of output	1 (pyranometer of AWS Cikancung)

Value of hidden layer and neuron number are determined by best trial process during training. Validation aims to review model training performance on a smaller number of data distributions. This process ensures that the model experience no underfitting or overfitting [19]. Validation results are then compared against modeling performance criteria based on the WMO requirements for solar radiation intensity measurement, namely a maximum of 8% error [20].

D. Model Testing

AWS pyranometer imputation model needs to be tested to determine the level of LSTM accuracy performance. Model is tested against the 2022 dataset. The imputed data is compared to actual AWS pyranometer data. Accuracy performance of the imputation model is measured using 3 parameters, namely correlation coefficient (R), root mean square error (RMSE) and mean average percentage error (MAPE). These equations are stated respectively as follows [21]:

$$R = \frac{\sum_{i=1}^{m}(x_i - \bar{x})\sum_{i=1}^{m}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - x_i)^2} \quad (3)$$

$$MAPE = \frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{x_i - y_i}{y_i}\right| \quad (4)$$

y is denoted for actual data, while x is denoted as imputed data. correlation coefficient states the relationship between the imputed data variables and the actual data on a scale of 0-1. RMSE states the error value in units of solar radiation intensity. MAPE expresses the error value in percentage form.

IV. RESULT AND ANALYSIS

Initial total raw dataset for AWS Cikancung, AWS Cipasung and AWS Cimalaka are respectively 49,817; 49,697; and 49,820 for 2021, then 51,278; 51,318; and 50,130 for 2022. AWS Cimalaka contains 10 outliers in 2021 and 5 outliers in 2022, so it remains 49,810 in 2021 and 50,125 in 2022. AWS Cipasung contains only 2 outliers in 2022, so it remains 51,316 data in 2022. AWS Cimalaka contains no outliers in 2021 and 2022. Next, a listwise deletion is performed on the outlier. This process removes all raw data in the same sampling time when there are outliers.

Data are only selected at 07.00 - 19.00 Local Time or 00.00 - 12.00 am UTC. Hence, total raw selected dataset are 26,280 data for every sites, whether in 2021 or 2022. There are no outlier inside the selected dataset. Simulation is performed on Spyder application. Spyder utilizes Python 3.9 programming language. Library used in compiling LSTM algorithm is keras-tensorflow.

Missing data scenarios are carried out in AWS Cikancung 2022 dataset, then they are saved as actual data for testing. LSTM is trained utilizing 80% of 2021 dataset. Training stage duration spends 305 s and yields 374.40 W/m^2 of RMSE. Table IV shows validation result of LSTM algorithm.

TABLE IV. VALIDATION RESULT OF LSTM

Total Missing Data	Validation		
	R	RMSE (W/m^2)	MAPE (%)
5256	0.763	194.04	5.13

LSTM is validated using 20% of 2021 dataset. According to Table IV, validation RMSE has insignificant difference compared to training result. It means LSTM model experience no underfitting or overfitting. Imputed data in validation process has

moderate correlation against actual data with R > 0.7. MAPE value is less than 8%, so it indicates that LSTM model fulfill WMO requierement for solar radiation intensity measurement. LSTM is then tested using 2022 dataset based on arranged scenarios. There are 5 variations of missing data scenarios in different duration. Table V shows model testing results.

TABLE V. TEST RESULT OF LSTM

Total Missing Data	Missing Data Scenario	Testing		
		R	RMSE (W/m^2)	MAPE (%)
1095	30 minutes	0.868	243.90	5.26
1098	60 minutes	0.894	235.85	4.14
1098	3 hours	0.915	222.05	3.32
1116	6 hours	0.910	208.48	2.93
1152	12 hours	0.911	176.18	1.76

Based on Table V, all imputed data ara strongly correlated to actual data, because R>0.85 in overall scenarios. RMSE testing values are still in range of training and validation results. Meanwhile, MAPE values are <8%, so it meet WMO requirement for solar radiation intensity measurement. In accordance with missing data duration, longer imputed values produces higher accuracy. It is proven by an increase of R values, and a decrease of RMSE and MAPE values. It shows that LSTM has capability on long term data imputation, because it has forget gate as memory holder in each neuron [22].

Figure 3 shows imputed data versus actual data with 60 minutes of missing data. Missing data occurs on January 1, 2022 at 04.30 - 05.30 UTC. Solar radiation intensity may reach zenith point on this period. Figure 3 determines that imputed data has difference compared to actual data, but it follows the fluctuation adjacently.

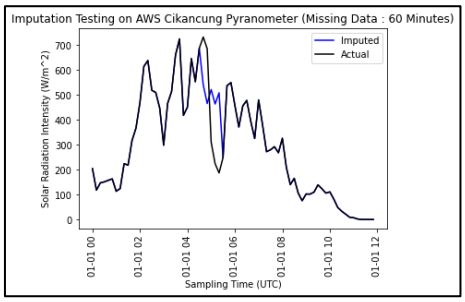

Fig. 3. Imputation testing plot with 60 minutes of missing data

Figure 4 shows imputed data versus actual data with 12 hours of missing data. Missing data occurs on January 1, 2022 at 00.00 - 12.00 UTC. It happens on one day of sunshine duration.. Figure 4 determines that imputed data is able to coincide actual data patterns adjacently. However, it has weakness on mid day of solar time, because actual data is tend to higher than imputed data.

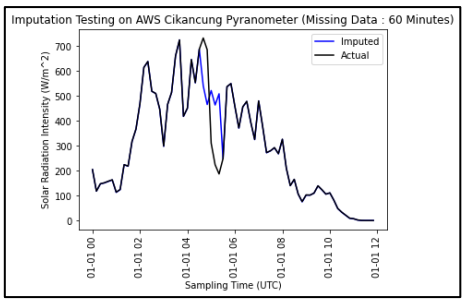

Fig. 4. Imputation testing plot with 60 minutes of missing data

Table VI describes statistical components of imputed data versus actual data. Imputed data give no significant changes to actual statistical components value based on its mean, median, maximum, minimum and standar deviation. Imputation affects only standar deviation. Greater missing data sequences yield a little decrease on deviation. It becomes a limitation of any imputation model.

TABLE VI. TEST RESULT OF LSTM

Component	Actual Data W/m^2	Imputed Data (W/m^2)				
		30 min	60 min	3 hours	6 hours	12 hours
Mean	284	285	285	284	284	284
Median	214	217	217	216	216	217
Max	1141	1141	1141	1141	1141	1141
Min	0	0	0	0	0	0
Stdev	255	254	254	254	253	253

Based on Figure 5, distribution of pyranometer data is a normal distribution type with a certain skewness. Results of imputation in all missing data scenarios have trivial impact on data variability, so it unchange distribution shape.

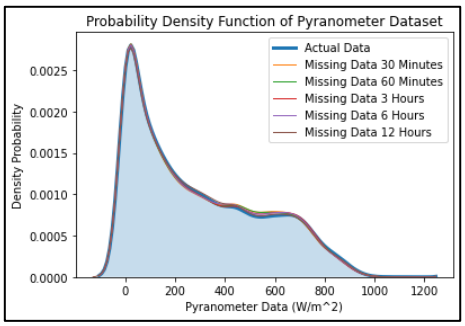

Fig. 5. Imputation testing plot with 60 minutes of missing data

Nonetheless, LSTM-based imputation model needs improvement in order to minimize bias. Difference of imputed versus actual data is caused by input compability, hyperparameter setting and internal bias of the algorithm. Solar radiation intensity measurement is also affected by cloud cover. Spatial relationship among three sites of pyranometer may be interrupted by different total cloud cover value at same time. Cloud has irregular shape and area, meanwhile pyranometer site has fixed position and elevation. Hyperparameter setting of LSTM also contributes bias of imputation model [23]. Optimal number of hidden layers and neuron should be tuned in future works.

V. CONCLUSION

Overall, LSTM-based imputation model has ability of filling gap data on AWS Cikancung pyranometer with maximum missing sequence of 12 hours. It yields strong correlation to actual data with R>0.85. Imputation model has MAPE 1.76% - 5.26% for missing duration 30 minutes-12 hours. It still it meet WMO requirement for solar radiation intensity measurement with MAPE<8%. Difference of imputed versus actual data is caused by input compability, hyperparameter setting and internal bias of the algorithm.

REFERENCES

[1] L. Wald, Fundamentals of Solar Radiation. CRC Press, 2021. https://doi.org/10.1201/9781003155454.

[2] J. Haigh, Solar influences on Climate. Grantham Institute for Climate Change Briefing paper No 5. Imperial College London. 2011

[3] J. Fraden, Handbook of Modern Sensors. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-19303-8

[4] Badan Meteorologi Klimatologi dan Geofisika, Peraturan Kepala BMKG Nomor 7 Tahun 2014 tentang Standar Teknis dan Operasional Pemeliharaan Peralatan Pengamatan Meteorologi, Klimatologi dan Geofisika. 2014.

[5] M. Al-Rasheedi, Gueymard, A. Ismail, , & T. Hussain, Comparison of two sensor technologies for solar irradiance measurement in a desert environment. Solar Energy, 161, 194–206, 2018. https://doi.org/10.1016/j.solener.2017.12.058

[6] J. D. Wood, T. J. Griffis, & J. M. Baker, Detecting drift bias and exposure errors in solar and photosynthetically active radiation data. Agricultural and Forest Meteorology, 206 33–44, 2015. https://doi.org/10.1016/j.agrformet.2015.02.015

[7] Manfred Georg Kratzenberg, Hans Georg Beyer, S. Colle, and A. Albertazzi, "Uncertainty Calculations in Pyranometer Measurements and Application," Solar Energy, Jan. 2006, https://doi.org/10.1115/ISEC2006-99168

[8] C. Turrado, M. López, F. Lasheras, B. Gómez, J. Rollé, and F. Juez, "Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions," Sensors, vol. 14, no. 11, pp. 20382–20399, Oct. 2014, https://doi.org/10.3390/s141120382

[9] Ö. A. Karaman, T. Tanyıldızı Ağır, and İ. Arsel, "Estimation of solar radiation using modern methods," Alexandria Engineering Journal, vol. 60, no. 2, pp. 2447–2455, Apr. 2021, https://doi.org/10.1016/j.aej.2020.12.048

[10] E. Afrifa-Yamoah, U. A. Mueller, S. M. Taylor, and A. J. Fisher, "Missing data imputation of high-resolution temporal climate time series data," Meteorological Applications, vol. 27, no. 1, Jan. 2020, https://doi.org/10.1002/met.1873

[11] A. Parasyris, G. Alexandrakis, G. V. Kozyrakis, K. Spanoudaki, and N. A. Kampanis, "Predicting Meteorological Variables on Local Level with SARIMA, LSTM and Hybrid Techniques," Atmosphere, vol. 13, no. 6, p. 878, May 2022, https://doi.org/10.3390/atmos13060878

[12] C. A. Fiebrich, C. R. Morgan, A. G. McCombs, P. K. Hall, and R. A. McPherson, "Quality Assurance Procedures for Mesoscale Meteorological Data," Journal of Atmospheric and Oceanic Technology, vol. 27, no. 10, pp. 1565–1582, Oct. 2010, https://doi.org/10.1175/2010JTECHA1433.1

[13] F. Vejen, C. Jacobsson, U. Fredriksson, M. Moe, L. Andresen, E. Hellsten, P. Rissanen, Þ. Pálsdóttir, & Þ. Arason, Quality Control of Meteorological Observations Automatic Methods Used in the Nordic Countries. KLIMA Report. 2002.

[14] T. B. Pepinsky, "A Note on Listwise Deletion versus Multiple Imputation," Political Analysis, vol. 26, no. 4, pp. 480–488, Aug. 2018, https://doi.org/10.1017/pan.2018.18

[15] S. Hamdi, "Mengenal Lama Penyinaran Matahari Sebagai Salah Satu Parameter Klimatologi," Berita Dirgantara, vol. 15, no. 1, 2014, https://jurnal.lapan.go.id/index.php/berita_dirgantara/article/view/2068.

[16] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," Journal of Big Data, vol. 8, no. 1, Oct. 2021, https://doi.org/10.1186/s40537-021-00516-9

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: https://doi.org/10.1162/neco.1997.9.8.1735.

[18] F. L. Gewers et al., "Principal Component Analysis," ACM Computing Surveys, vol. 54, no. 4, pp. 1–34, May 2022, https://doi.org/10.1145/3447755

[19] M. Kubat, An Introduction to Machine Learning. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-63913-0

[20] "Guide to Instruments and Methods of Observation (WMO-No. 8) | World Meteorological Organization," community.wmo.int. https://community.wmo.int/en/activity-areas/imop/wmo-no_8

[21] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, vol. 7, no. 5, p. e623, Jul. 2021, https://doi.org/10.7717/PEERJ-CS.623

[22] T. T. K. Tran, S. M. Bateni, S. J. Ki, and H. Vosoughifar, "A Review of Neural Networks for Air Temperature Forecasting," Water, vol. 13, no. 9, p. 1294, May 2021, doi: https://doi.org/10.3390/w13091294.

[23] J. Kalliola, J. Kapočiūtė-Dzikienė, and R. Damaševičius, "Neural network hyperparameter optimization for prediction of real estate prices in Helsinki," PeerJ Computer Science, vol. 7, p. e444, Apr. 2021, https://doi.org/10.7717/peerj-cs.444