Jurnal Infotel
Vol. 16, No. 4, November 2024, pp. 758–777. doi:10.20895/infotel.v16i3.1225

Research Article

Solar Radiation Prediction using Long Short-Term Memory with Handling of Missing Values and Outliers

Alfin Syarifuddin Syahab[1,*], MS Hendriyawan Achmad[2]

[1]Stasiun Klimatologi D.I Yogyakarta, Badan Meteorologi Klimatologi dan Geofisika, Yogyakarta 55285, Indonesia
[2]Master Program of Information Technology, Universitas Teknologi Yogyakarta, Yogyakarta, 55285 Indonesia

*Corresponding email: hendriyawanachmad@uty.ac.id

Received: August 22, 2024; Revised: October 28, 2024; Accepted: November 18, 2024.

Abstract: The pyranometer sensor measures Global Horizontal Irradiance (GHI), a key parameter for weather analysis and photovoltaic prediction. GHI data is essential for assessing solar power generation performance in distributed energy systems. However, GHI sensor data often contains missing values and outliers due to measurement errors. This research aims to develop a GHI prediction model that handles missing values and outliers to improve solar radiation prediction. Data preprocessing includes imputation of missing values using linear, polynomial, and Piecewise Cubic Hermite Interpolating Polynomials (PCHIP), and outlier elimination using Random Sample Consensus (RANSAC). Previous studies show that Long Short-Term Memory (LSTM) models outperform traditional machine learning in predictions. This study compares LSTM models with and without data preprocessing. The results show that PCHIP imputation achieved the best performance with a Mean Absolute Error (MAE) of 39.708 W/m^2, Root Mean Square Error (RMSE) of 76.224 W/m^2, Normalized Root Mean Square Error (NRMSE) of 0.433, and a Coefficient of Determination (R^2) of 0.903. After outlier elimination, the imputation yielded an MAE of 44.377 W/m^2, RMSE of 86.738 W/m^2, NRMSE of 0.500, and R^2 of 0.886, with RANSAC eliminating 100% of outliers. The LSTM model with preprocessing showed better results, with an MAE of 42.863 W/m^2, RMSE of 82.396 W/m^2, NRMSE of 0.396, and R^2 of 0.918. This study provides an effective GHI prediction model to support solar power plant operations by addressing missing values and outliers.

Keywords: data preprocessing, global horizontal irradiance, interpolation, long short-time memory, prediction, random sample consensus

1 Introduction

The pyranometer sensor, part of the Automatic Weather Station (AWS), measures solar radiation intensity as Global Horizontal Irradiance (GHI), which is used for research purposes and predicting weather conditions [1]. Solar radiation predictions are essential for providing information on weather, climate, and solar energy potential in an area, supporting solar panel installations for renewable energy [2]. In the solar power system, a solar radiation measurement system is installed to monitor solar radiation conditions to determine the performance of the generating system [3]. Weather variables, especially solar radiation, impact the uncertainty of a solar energy generation, posing challenges for integrating into the evolving electricity grid. Fluctuations in photovoltaic output can affect power stability and economic benefits. Thus, accurate solar radiation prediction is crucial for optimizing distributed energy networks [4].

Ground measurement sensor data has the weakness of low quality data [5]. Low data quality from the missing values filled in inaccurately results an invalid analysis [6]. Lost values in sensor data occur due to maintenance, data logger errors, communication failures, hardware damage, and power exhaustion [7,8]. Additionally, outliers, which are data points that significantly deviate from the usual pattern, can negatively affect the quality of sensor data [9]. Handling of outliers can be identified using statistical methods and then replacing or deleting outliers [10]. In response to sensor data issues, this research developed a solar radiation prediction model that includes data preprocessing techniques such as imputing missing values and removing outliers. These preprocessing steps enhance data quality by addressing missing values and eliminating anomalies [11]. For statistical methods in time series analysis, the estimation of missing values is required. This method enables the development and enhancement of prediction models [12]. Research related to the imputation of missing values can improve prediction performance with the best evaluation at Mean Absolute Error (MAE) 117.81 and Root Mean Square Error (RMSE) 201.58 at 60% of missing values in the dataset and improve the performance of the previous dataset with 80% missing values using the Imputed algorithm Gate Recurrent Unit (IGRU) [13]. In addition, outlier handling contribute to the fact that the model without outliers is the most appropriate model compared to the initial data with outliers producing higher coefficient of determination (R^2) [14]. Research shows that the best method for predicting optimal planting days based on weather forecasts used Histogram Gradient Boosting Regressor, achieving R^2 of 0.938 and MAE of 77.689. This model employed One-Class SVM for outlier removal and Random Forest for imputation. The data preprocessing significantly improved results compared to the initial model without imputation and outlier removal, which had an R^2 of 0.723 and MAE of 111.645 [15].

The prediction of solar radiation data achieving accurate results depends on choosing the right architecture to handle dynamic modeling data, determining the right algorithm, selecting input variables, and adjusting model hyperparameters [16]. Several studies show the advantages of Long Short-Time Memory (LSTM) in improving predictions using time series statistical techniques. Previous research regarding the GHI prediction model using Bidirectional LSTM has the best performance among Multilayer Perceptron (MLP), Random Forest, Extreme Gradient Boosting (XGboost), and linear regression [17]. In previous research, the LSTM prediction model for univariate GHI produced greater MAE, RMSE, and NRMSE than multivariate [18, 19]. Based on these findings, prior research concentrated on univariate models, which showed lower performance compared to multivariate models. Additionally, previous studies did not address data preprocessing techniques for handling missing values and outliers. Moreover, the data utilized was limited to a single measurement location in Andhra Pradesh, India, covering only a one-year period with five-minute intervals [18].

This research introduces a novel approach of LSTM-based model for predicting Global Horizontal Irradiance (GHI) using two different datasets. The first dataset (Model 1) does not address missing data or outliers, while the second dataset (Model 2) uses advanced preprocessing techniques. Model 2 applies imputation methods—linear interpolation, polynomial interpolation, and Piecewise Cubic Hermite Interpolating Polynomials (PCHIP)—chosen for their effectiveness in improving performance metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Normalized Root Mean Square Error (NRMSE), and R^2. It also removes outliers using the Random Sample Consensus (RANSAC). The study compares the LSTM model's performance between the preprocessed (Model 2) and non-preprocessed (Model 1) datasets. This new approach helps the Indonesian Meteorology, Climatology, and Geophysics Agency (BMKG) enhance their prediction models and supports solar power plants in improving energy management.

2 Research Method

This research focuses on a specific scope, with the following limitations: First, it is limited to GHI data from the pyranometer sensor at the AWS installed at BMKG's Climatology Station in Yogyakarta, Indonesia. Second, the study uses hourly GHI data from 2022–2023 as a single parameter. Third, data preprocessing involves cleaning through the imputation of missing values and outlier removal. The study compares solar radiation prediction models with and without data preprocessing to develop univariate prediction model using the LSTM algorithm. The research methodology phases are outlined in Figure 1.

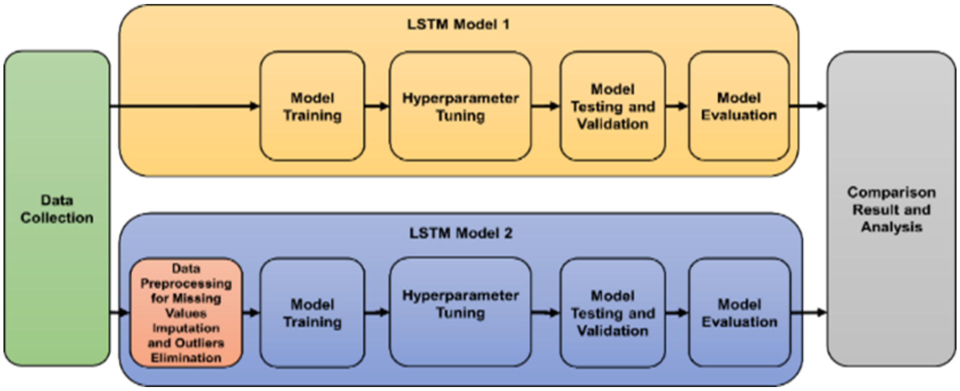
LSTM Model 1
Data Collection
Model Training
Hyperparameter Tuning
Model Testing and Validation
Model Evaluation
LSTM Model 2
Data Preprocessing for Missing Values Imputation and Outliers Elimination
Model Training
Hyperparameter Tuning
Model Testing and Validation
Model Evaluation
Comparison Result and Analysis

Figure 1: Research methodology diagram.

The methodology involves several detailed stages. First, GHI data from the pyranometer sensor at BMKG's Climatology Station in Yogyakarta was collected to create the dataset. Two prediction models were then developed: one without data preprocessing

(LSTM Model 1) and another with preprocessing (LSTM Model 2), which involved filling missing values using the best interpolation method (linear, polynomial, or PCHIP) and removing outliers with RANSAC. Both models were trained using GHI data, with hyperparameters such as the number of neurons, epochs, and window size optimized. The trained models were validated using test data to assess performance. Finally, the performance of both models was evaluated and compared using metrics like MAE, RMSE, NRMSE, and R^2.

2.1 Data Collection

The GHI data used in this study were obtained from AWS in the Yogyakarta Special Region, which is managed by the BMKG Climatology Station. GHI parameter data is one of the AWS measurement parameters of the Yogyakarta Climatology Station, which is operated in the observation equipment park. AWS data records weather parameters with data transmission intervals every ten minutes for the period January 2022 to December 2023. From these weather parameters, one solar radiation parameter is taken. The solar radiation parameter obtained from AWS is the average GHI per ten minutes in Watts$/m^2$ units. Microsoft Excel and Python3 are needed to support processing and analysis of the dataset.

2.2 Data Preprocessing

A collection of average solar radiation data at ten-minute intervals was obtained from this equipment for the period 2022 to 2023. By calculating an average GHI value for one-hour intervals, the raw data processing was completed. Hourly intervals were carried out to optimize the prediction model as in previous research from [19]. Then, the process flow for implementing the data preprocessing of filling in missing values and eliminating outliers is shown in Figure 2. The dataset simulation scenario is made into two models. The first model is to train the LSTM model by simulating the missing values with 500 random data in the data row range of 3000-6000 and simulating outlier values with 500 random data in the data row range of 6000-9000. The second model uses a dataset that has been improved with data cleaning. The second model is used to train a prediction model with the condition that the dataset has been corrected for outliers using RANSAC and filling in missing values using the best test results on MAE, RMSE, NRMSE, and R^2 values between linear, polynomial, and PCHIP interpolation methods. The dataset is then split into components for training and testing data. Referring to [20], the LSTM model which produced a good RMSE of 6.42 used the ratio 70% of train data and 30% of test data.

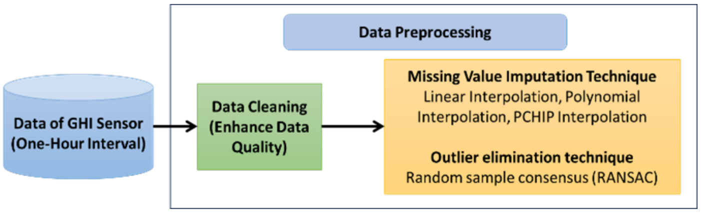
Data Preprocessing
Data of GHI Sensor (One-Hour Interval)
Data Cleaning (Enhance Data Quality)
Missing Value Imputation Technique
Linear Interpolation, Polynomial Interpolation, PCHIP Interpolation
Outlier elimination technique
Random sample consensus (RANSAC)

Figure 2: The flow of data cleaning stages in data preprocessing.

2.3 Interpolation

Linear interpolation is a method for performing linear calculations based on the straight-line distance at the values of two given points. This method estimates the values that lie between these points, and is the simplest way to replace missing values [21]. The linear interpolation function can be written as (1) [22].

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{(x_1 - x_0)(x_1 - x_0)}. \tag{1}$$

The equation (1) explains the x indicates the independent variable, where x_1 and x_0 are the known values of the independent variable and $f(x)$ is the variable dependent on the x of the independent variable. Then, polynomial interpolation can be used as an approximation function in numerical analysis problems because of its simple structure, so that polynomials can be used effectively. The function values at known points form a polynomial of degree less than or equal to n, this polynomial is called an interpolation polynomial. The equation of that method is in (2) [23].

$$P(x) = a_0 + a_1 x + a_2 x^2. \tag{2}$$

The equation (2) has a_0, a_1, a_2 which are coefficients calculated based on the data points which want to interpolate. This polynomial can be used to estimate the y value from x values between x_0, x_1, and x_2. In addition, the PCHIP is a third-order polynomial that has the characteristic of preserving shape by simply matching the first-order derivative of a data point with its neighbors, before and after) [24]. The mathematical equation of PCHIP interpolation is given as in (3) and (4).

$$P(x) = h_1 + (x - x_i)h_2 + (x - x_i)^2 h_3 + (x - x_i)^3 h_4, \tag{3}$$

$$h_1 = f_i, h_2 = \dot{f}_i, h_3 = \frac{3C_{i+\frac{1}{2}} - \dot{f}_{i+1} - 2\dot{f}_i}{\delta x_{i+\frac{1}{2}}}, h_4 = \frac{2C_{i+\frac{1}{2}} - \dot{f}_{i+1} - \dot{f}_i}{\delta x_{i+\frac{1}{2}}} \tag{4}$$

where $x_i \leq x \leq x_{i+1}$, the data point is denoted by f_i, and for $1 \leq I \leq n$, $\dot{f}_i$ is the slope of the x_i. The data point node value and its derivative values assigned to the data point node are used to calculate the PCHIP [25].

2.4 Random Sample Consensus

The RANSAC estimates mathematical model parameters from a set of data divided into inliers and outliers [26]. This method is used to obtain a model based on linear regression, carried out on input data which may include samples that have outliers. The basic assumption of this algorithm is that the measured $Y_{\text{measured}}(x)$ depends on a set of outlier independent variables added to it in (5).

$$Y_{\text{measured}}(x) = Y_{\text{outlier-free}}(x) + N \tag{5}$$

where $Y_{\text{outlier-free}}(x)$ is the expected measurement values that are free from noise and N is the internal noise that influences the measured outlier value. RANSAC assumes that outlier follow the assumption of having a constant distribution across all measurement values [27].

2.5 Long Short-Term Memory

LSTM developed to study long-term time dependencies of analysis. This is an effective model in dealing with sequential data problems containing long-term dependencies. Time-series prediction, sentiment analysis, and machine translation are a few uses for LSTM [18]. By employing memory cells with a range of gates, LSTM is intended to retain long-term temporal relationships. Figure 3 shows the single cell LSTM memory architecture and this picture is adapted from [28].

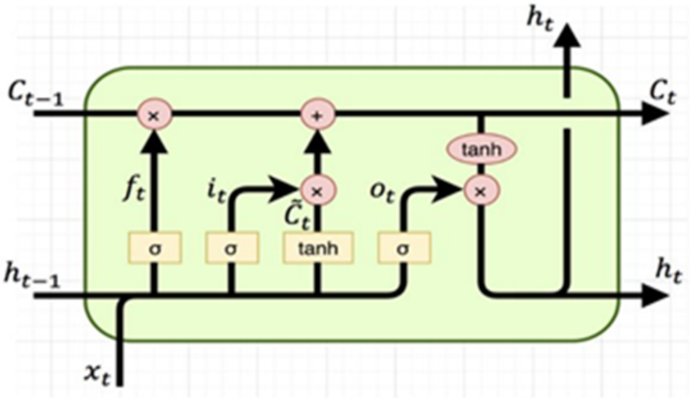

Figure 3: Single cell memory architecture in LSTM (Reproduced from [28]).

The equations related to the different gates of LSTM cells are discussed with gate descriptions [29]. For example, at time t, with the current input x_t and the previous hidden state $h(t-1)$, the current values of the forget gate f_t, input gate i_t, cell state c_t, output gate o_t, and hidden state h_t are calculated as follows (6)-(10).

$$\text{Forget gate}(f_t) = \sigma((w_f[h_{t-1}, x_t]) + b_f) \tag{6}$$

$$\text{Input gate}(f_t) = \sigma((w_i[h_{t-1}, x_t]) + b_i) \tag{7}$$

$$\text{Cell gate}(f_t) = \tanh((w_c[h_{t-1}, x_t]) + b_c) \tag{8}$$

$$\text{Output gate}(f_t) = \sigma((w_o[h_{t-1}, x_t]) + b_o) \tag{9}$$

$$\text{Hidden gate}(f_t) = o_t * tanh(c_t) \tag{10}$$

Weight matrices in this case are w_f, w_i, w_c, and w_o. The biases for each gate are b_f, b_i, b_c, and b_o. The $\tanh$ and $\text{sigmoid}(\sigma)$ are the activation function; symbol $*$ denotes multiplication of elements, while $+$ denotes addition.

2.6 Model Prediction LSTM

The study proposes a paradigm for univariate LSTM model-based GHI prediction. The time series data of the GHI measurement are ready for testing the LSTM model following data preprocessing and partitioning. Using a sliding window approach, the prediction makes estimates about future time stages by taking into account those of the past. The input and output vectors for the univariate LSTM model are shown in Figure 4.

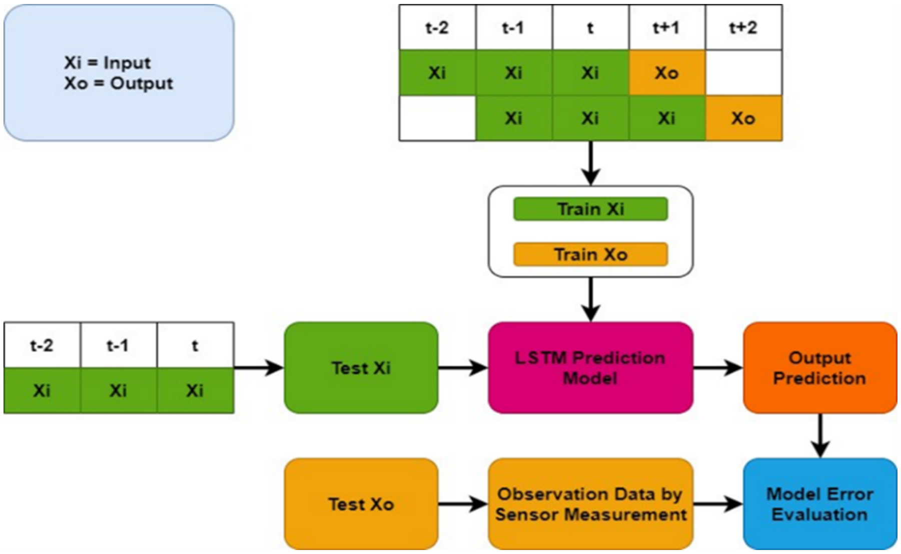

Figure 4: The prediction model using LSTM.

The input vector for the univariate prediction model only takes the GHI variable into account. Figure 4 explains that t indicates the current time step, where $t-1$ and $t-2$ are the lag values or window sizes, and $t+1$ and $t+2$ indicate the next step. In the LSTM model, hyperparameter settings are required, such as hidden layers, batch size, epochs, window size, and neurons. The initial settings shown in Table 1 describes the hyperparameters set for the univariate LSTM model. Hyperparameters in the LSTM determined based on the results of initial experiments as in Table 1. To determine other hyperparameters, the research tried variations of the activation function, optimizer, hidden layer, batch size, epoch, window size, and neuron to determine the best metric evaluation values.

Table 1: Initial setting of LSTM hyperparameters

Hyperparmeter	**Value**
Optimizer	Adam
Hidden Layer	1
Batch Size	5
Epoch	25
Neuron	50
Window Size	3
Activation Function	tanh

2.7 Model Evaluation

RMSE and MAE are two standard metrics used in model evaluation. For a sample of n observations with a value of y $(y_i, i = 1, 2, 3, \ldots, n)$ and $\hat{y}$ is the value of the model prediction or estimation result. The RMSE calculation is shown in (11).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{11}$$

Taking the root does not affect the relative ranking of the models, but produces a metric with the same units as y, which easily represents normally distributed errors. A common statistical tool for assessing model performance in studies on climate, air quality, and meteorology is the root mean square error (RMSE) [30]. The next evaluation model is MAE with the calculation formula in (12).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2 \tag{12}$$

MAE is a metric used to determine the absolute difference between the predicted and actual values. The range of MAE is $(0, +\infty)$; the smaller the MAE, the higher the accuracy of the predicted model. One of the advantages of MAE is that it is identical to the original data [31]. After that, this study uses NRMSE to compare data sets with different scales because it uses normalization. The NRMSE calculation is given in (13).

$$\text{NRMSE} = \frac{\text{RMSE}}{\bar{y}_i} \tag{13}$$

The symbol $\bar{y}_i$ is the average of observations. NRMSE does not differentiate between negative and positive errors and extreme errors will be penalized by NRMSE. The NRMSE value must be kept as small as possible for the estimation to be considered successful [19]. Evaluation also uses the R^2. The R^2 value is calculated in the range $(-\infty, 1)$ according to the reciprocal relationship between the actual value and the predicted value. The worst value range is $-\infty$ and the best value is $+1$. The equation (14) shows the calculation of the R^2 value.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{14}$$

The element $\hat{y}_i$ is the predicted value, y_i is the observation value, $\bar{y}_i$ is the average observation value, and n is the total data sample. The R^2 number indicates the percentage of variance in the dependent variable that can be predicted based on the independent variable [32].

3 Results

The one-hour interval was chosen for the GHI data prediction model because it is consistent with prior research data that yields a good evaluation of the prediction model utilizing solar radiation data with RMSE values of 71.25 and MAE 46.00 [33]. In testing data completeness, the amount of data recorded was 16.843, with the number of missing values being 677, which shows a percentage of data completeness of 96.14%. The dataset that will be used to the prediction model is 16.843 after elimination of missing values. Figure 5 shows the GHI dataset becomes data with an interval of one hour from raw data with an interval of ten minutes.

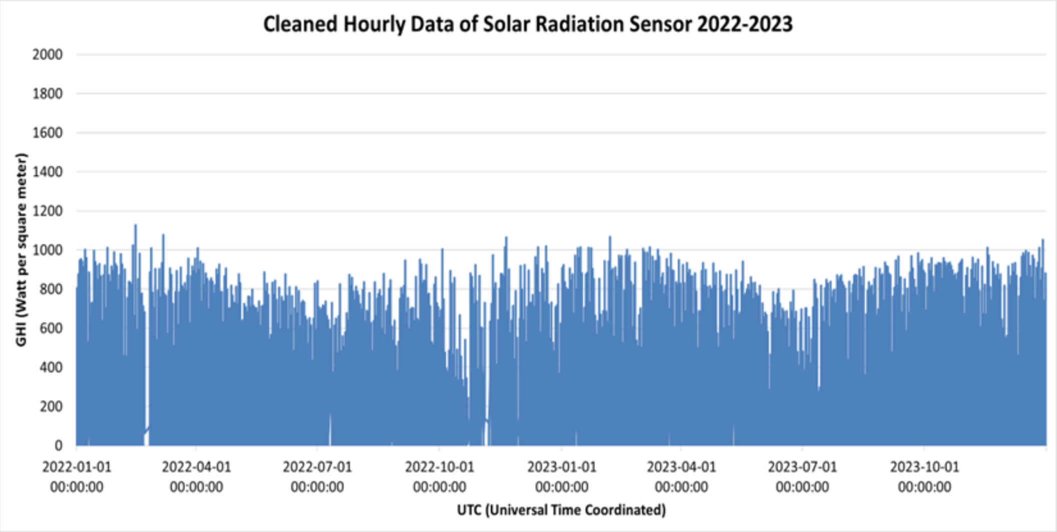

Figure 5: One-hour interval GHI solar radiation dataset.

Dataset at one hour intervals on sensor measurements is carried out using an average technique of measuring every ten minutes of data. In this case, if there are missing values then there should be six data in one hour, then the average number of values carried out is reduced. For example, if there are only three data recorded, then the divider becomes three. In cases where there are more than six rows of missing data, it will be a missing value which is then eliminated. At the data preprocessing stage, the collected dataset has 100% completeness and no outliers are detected based on the sensor measurement value range specifications of 0-2000 W/m^2. In testing, a dataset scenario with missing values and outliers is created by collecting random samples from the dataset and replacing them with missing values and outliers in order to construct a model for filling in the gaps and removing outliers. The dataset simulation scenario is explained in Table 2.

Table 2: Dataset simulation scenarios

Dataset	Missing Values	Outliers	Number of Data
Model 1	500	500	16343
Model 2	0	0	16843

The scenario in the GHI dataset is determined with 500 missing data values randomly in the data row range of 3000-6000 and simulated outlier values with 500 random data in the data row range of 6000-9000. The percentage of missing data and outliers is 5.937% of the dataset. The results of the GHI dataset scenario model 1 with missing sample values and outliers are shown in Figure 6.

Data samples that have missing values are in the data period range 3000 – 6000. The interpolation approach is used to fill in missing values, and the resulting MAE, RMSE, NRMSE, and R^2 evaluation values are tested toward the original values. Then, the dataset that has outliers is in the data period range 6000 – 9000. Outlier values are then eliminated using the RANSAC method, then validated with the percentage of outlier values that can

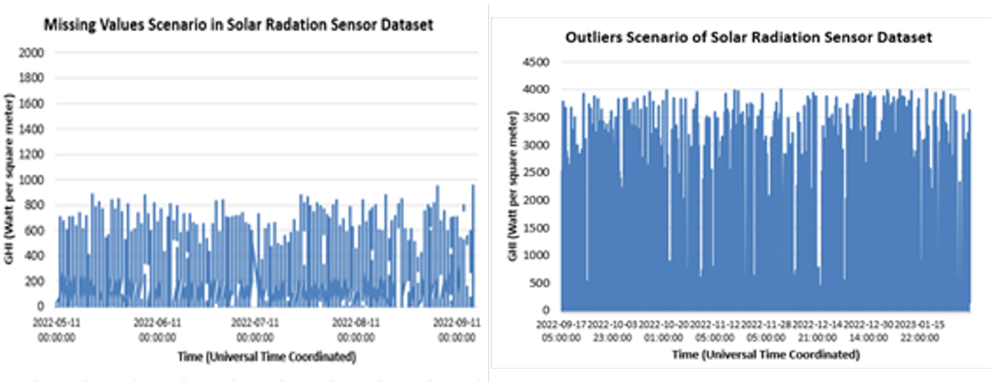

Figure 6: The GHI dataset with missing values and outliers scenario.

be eliminated using data completeness testing Following the removal of outliers, the outlier values become missing values, which are subsequently filled in using interpolation. When validating with MAE, RMSE, NRMSE, and R^2, the initial value of the missing item is utilized. The dataset is then used to train and test the prediction model using the LSTM algorithm, with the best evaluation value for filling in the missing data determined from the linear interpolation, polynomial interpolation, and PCHIP interpolation methods.

3.1 Interpolation Testing

Scenario model 1 is utilized in the interpolation tests to fill in the GHI dataset's missing values. Using RANSAC, Scenario model 2 is utilized to fill in any missing data from the outcomes of outlier reduction. The results of testing on each GHI dataset scenario will be compared and further analyzed using prediction model evaluation metrics for interpolation results in scenario model 1 (MAE 1, RMSE 1, NRMSE 1, and R^2 1) then scenario model 2 (MAE 2, RMSE 2, NRMSE 2, and R^2 2). Figure 7 illustrates MAE, RMSE, NRMSE, and R^2 evaluation of interpolation for two models.

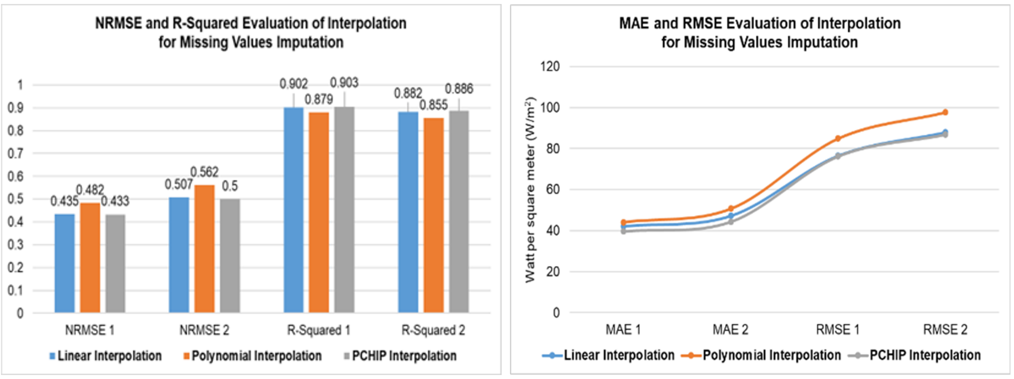

Figure 7: Evaluation metrics of interpolation for model 1 and model 2.

Testing to fill in the missing values in model 1 using interpolation methods produced an evaluation matric based on linear interpolation with MAE 42,227 W/m^2, RMSE 76,583 W/m^2, NRMSE 0.435, and R^2 0.902. Then, polynomial interpolation produces MAE 44.135 W/m^2, RMSE 84.891 W/m^2, NRMSE 0.482, and R^2 0.879. Furthermore, PCHIP interpolation obtained MAE 39.708 W/m^2, RMSE 76.224 W/m^2, NRMSE 0.433, and R^2 0.903. From the results of the interpolation test, PCHIP shows the best performance in missing values filling in the GHI dataset. In the model 2, The dataset which had outliers was 100% successful in being eliminated using the RANSAC to become missing values, the dataset was interpolated to fill in those missing values. Linear interpolation produces MAE 47.302 W/m^2, RMSE 87.917 W/m^2, NRMSE 0.507, and R^2 0.882. Furthermore, polynomial interpolation shows MAE 50.683 W/m^2, RMSE 97.592 W/m^2, NRMSE 0.562, and R^2 0.855. Then, PCHIP interpolation obtained MAE 44,377 W/m^2, RMSE 86,738 W/m^2, NRMSE 0.500, and R^2 0.886. From the test results of the interpolation tests, the PCHIP was chosen to fill in the missing values from the outlier elimination results in the dataset because it has the best performance.

3.2 RANSAC Testing

In the case of outliers in the GHI dataset, the outliers are handled in two stages. First, the outliers' removal stage uses the RANSAC by randomly detecting outliers in the dataset, then the detected outliers are deleted from the dataset. The Figure 8 shows the result of outlier removal by RANSAC. In this research, five experiments were carried out with 500 samples of outlier values that could be eliminated, including 328 with a percentage of 65.6%, 467 with a percentage of 93.4%, 270 with a percentage of 54%, 476 with a percentage of 95.2%, and 500 with a percentage of 100%.

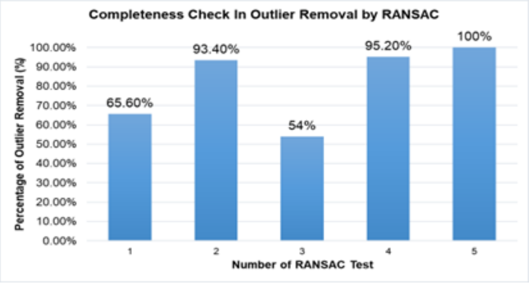

Figure 8: Completeness check in outlier removal by RANSAC.

3.3 LSTM Testing

Testing of the GHI prediction model was carried out using the LSTM with scenarios model 1 and model 2. The first model uses the dataset which has a random sample of missing values in the data row range 3000 – 6000 and outliers in the data row range 6000 – 9000. The second model shows the preprocessed dataset that was cleaned up by using RANSAC to eliminate outliers and PCHIP interpolation to fill in missing values. The dataset is divided into training data and testing data. To run the algorithm and conduct LSTM testing, hyperparameters need to first be configured. This research focuses on the GHI prediction model to get the best results by carrying out experiments on the combination of the best hyperparameter values in epoch, batch size, and hidden layer. In the initial setup, hidden layer 1, epoch 25, batch size 5, window size 3, and neuron 50. The results that have the best performance for the prediction model in two models are displayed in Figure 9.

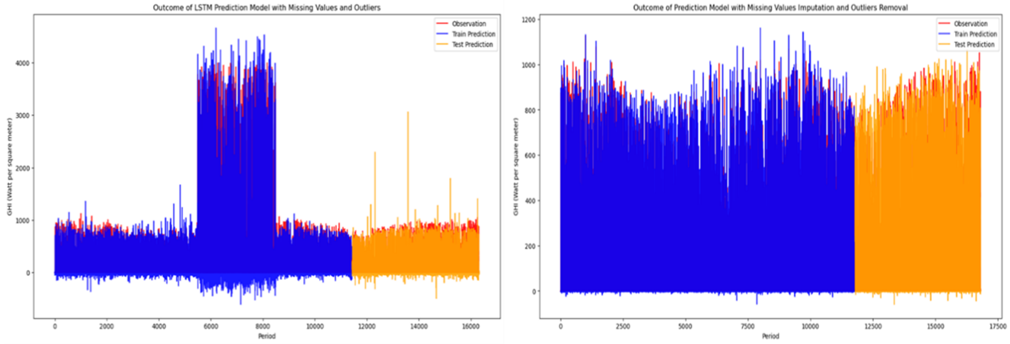

Figure 9: The result of LSTM prediction for model 1 and model 2.

Based on Figure 9, the blue line shows the predicted value from training data and the orange line shows the prediction result on test data. Also, the red line on the graph shows the observed value of the pyranometer sensor measurement for GHI. The results of testing the prediction model in model 1 using a simulated dataset that has random missing values which are then eliminated and the dataset has outliers that exceed the sensor measurement range at a threshold of 0 W/m^2 to 2000 W/m^2. Furthermore, the results of testing the prediction model in model 2 using a dataset that has been interpolated on missing values and eliminated outliers using RANSAC. Figure 10 shows the results of the prediction model on train data by the various window sizes.

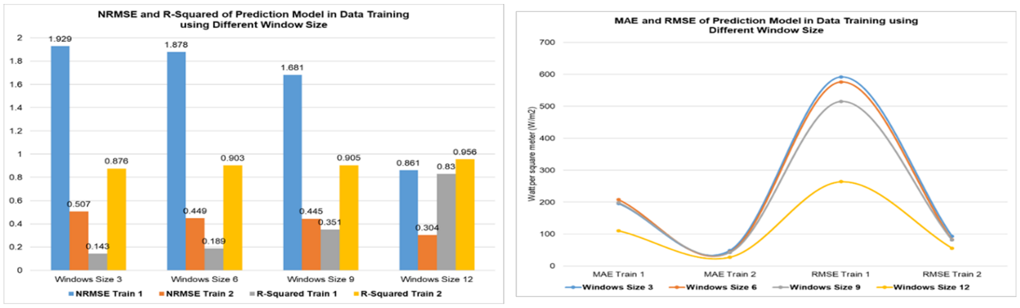

Figure 10: The result of the LSTM prediction model on train data using different of window sizes.

Metric evaluation using MAE, RMSE, NRMSE and R^2 testing of the performance of the LSTM prediction model was carried out using train data based on changes in window size. The training data for model 1 has outlier values and the data is reduced because there are missing values. Meanwhile, model 2 has been improved with data preprocessing. The best performance is found in window size 12. MAE, RMSE, NRMSE values in dataset model 2 have decreased compared to model 1. Then, R^2 in dataset model 2 has increased compared to model 1. This shows that the performance of model 2 with data preprocessing techniques can improve the performance of predictions. The performance in the model 2 test data have decreased compared to model 1. Figure 11 displays the results of the LSTM prediction model's evaluation on test data by various window sizes.

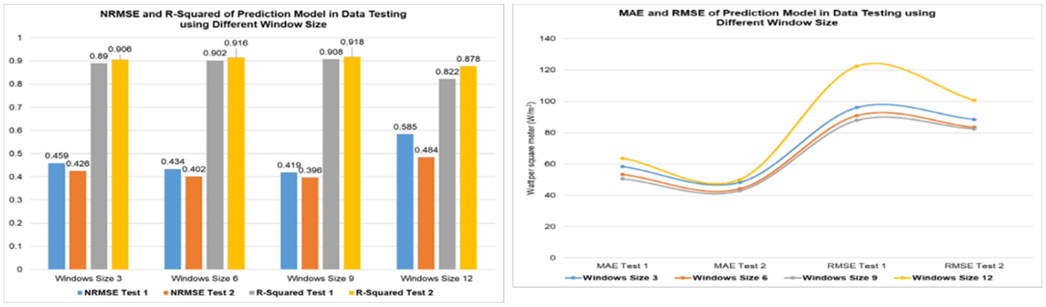

Figure 11: The result of LSTM prediction model on test data using different window sizes.

Metric evaluation using MAE, RMSE, NRMSE, and R^2 tests on the performance of the LSTM prediction model were also applied to test data with a data portion of 30% of the dataset. The model 1 test data has outlier values and eliminates missing values. Meanwhile, model 2 has been improved by data preprocessing. In the test data, model 2 values for MAE, RMSE, and NRMSE have decreased. Then, R^2 increases. This shows that the performance of model 2 is better than model 1 on test data. The best performance window size is found at a value of 9. After conducting hyperparameter analysis with previous related research, this research summarizes the process of determining the best hyperparameters for changes in batch size, epoch, neurons, hidden layers, and the window size. Table 3 is an evaluation of the performance of the LSTM prediction model on the GHI model 1 dataset without using data preprocessing.

In determining the best hyperparameters, this research uses the technique of averaging the MAE and RMSE values in model 1 and model 2. The average MAE and RMSE values in two models with the lowest values are used to determine the best. In Table 2 is a dataset that has data reduction due to missing values and outliers. The results of LSTM testing on these dataset conditions produced the best performance evaluation at window size 12 with MAE 110,475, RMSE 264,191, NRMSE 0.861 and R^2 0.830 on the training data and at window size 9 with MAE 50.539, RMSE 87.882, NRMSE 0.419 and R^2 0.908 on the testing data. Table 4 is an evaluation of the performance of the LSTM prediction model on the GHI model 2 dataset using data preprocessing.

The results of LSTM testing on dataset using data preprocessing on model 2 have better performance than model 1 without data preprocessing with MAE 27,581, RMSE 55,838, NRMSE 0.304 and R^2 0.956 on train data and MAE 42.863, RMSE 82.396, NRMSE 0.396, and R^2 0.918 on test data. Based on the average MAE and RMSE on model 1 and model 2, the LSTM test results of the GHI prediction model get the best hyperparameter at batch size 15, epoch 75, neurons 50, hidden layer 3, and window size 12.

Table 3: Performance Evaluation of LSTM Model 1

		Model 1 (Dataset with missing values and outlier)							
Hyper-	Value		Data Training				Data Testing		
parameter		MAE	RMSE	NRMSE	R^2	MAE	RMSE	NRMSE	R^2
Batch Size	5	206.814	602.088	1.962	0.114	74.210	108.167	0.516	0.861
	10	207.803	600.646	1.957	0.118	77.442	110.750	0.529	0.854
	15	**203.187**	**596.789**	**1.944**	**0.130**	**69.052**	**100.412**	**0.479**	**0.880**
	20	213.158	605.776	1.974	0.103	86.875	120.772	0.577	0.826
Epoch	25	203.187	596.789	1.944	0.130	69.052	100.412	0.479	0.880
	50	196.447	597.843	1.948	0.127	64.644	98.334	0.469	0.885
	75	**196.895**	**597.289**	**1.946**	**0.128**	**63.511**	**95.309**	**0.455**	**0.892**
	100	198.525	602.293	1.962	0.114	72.041	104.740	0.500	0.869
Neuron	**50**	**196.895**	**597.289**	**1.946**	**0.128**	**63.511**	**95.309**	**0.455**	**0.892**
	75	198.457	598.163	1.949	0.126	66.745	100.176	0.478	0.881
	100	196.610	598.623	1.950	0.124	65.426	98.735	0.471	0.884
Hidden Layer	1	196.895	597.289	1.946	0.128	63.511	95.309	0.455	0.892
	2	198.711	593.440	1.933	0.139	62.935	97.257	0.464	0.887
	3	**195.761**	**592.081**	**1.929**	**0.143**	**58.319**	**96.048**	**0.459**	**0.890**
	4	202.267	593.388	1.933	0.140	63.254	103.491	0.494	0.872
Window Size	3	195.761	592.081	1.929	0.143	58.319	96.048	0.459	0.890
	6	207.703	576.175	1.878	0.189	53.302	90.935	0.434	0.902
	9	198.555	515.577	1.681	0.351	50.539	87.882	0.419	0.908
	12	**110.475**	**264.191**	**0.861**	**0.830**	**63.709**	**122.479**	**0.585**	**0.822**

4 Discussion

This research uses linear, polynomial, and PCHIP interpolation to fill in missing values in the GHI dataset and the RANSAC method to help retrieve outlier data in the GHI dataset, then eliminate and also fill in missing values using linear interpolation, polynomial interpolation, and PCHIP interpolation. This interpolation method was chosen to fill in missing values because it has better capabilities than the simple method, namely the mean. Based on previous research showing that on missing data 5% of the dataset, the interpolation method has an RMSE evaluation value of 16.818, which is better than the mean method with an RMSE value of 167.756 [34]. The test results are in accordance with the theory contained in the research of [35] which states that the simplest method is to define a piecewise linear function between each number of points. Although linear interpolation does not result in smooth curves, linear methods are quick and simple to use. In order to get around this, PCHIP is used to manage the overshoot problem—that is, the occurrence of the interpolation curve beyond the actual value of the data interpolated between two given data points—and preserve the monotonicity of the points on the interpolation curve. Furthermore, the technique yields interpolation curves that are smoother. The dataset is then utilized to train an LSTM prediction model after it has been filled in and outliers removed. In the RANSAC test, it shows the completeness of the outlier data that was successfully removed in a sample of 500 outliers. The results with the completeness of the outlier data that was successfully removed 100% were selected to be filled with missing values using interpolation. The interpolation technique is used to fill in missing values in the GHI dataset after the elimination process.

Table 4: Performance Evaluation of LSTM Model 2

		Model 2 (Dataset with missing values and outlier)							
Hyper-	**Value**	**Data Training**				**Data Testing**			
parameter		**MAE**	**RMSE**	**NRMSE**	R^2	**MAE**	**RMSE**	**NRMSE**	R^2
Batch Size	5	52.677	95.843	0.521	0.869	52.781	91.293	0.440	0.900
	10	54.231	97.149	0.528	0.866	54.251	92.538	0.446	0.897
	15	**50.747**	**94.550**	**0.514**	**0.873**	**50.625**	**90.483**	**0.436**	**0.901**
	20	55.295	98.578	0.536	0.862	55.463	94.395	0.455	0.893
Epoch	25	50.747	94.550	0.514	0.873	50.625	90.483	0.436	0.901
	50	49.661	93.911	0.510	0.875	49.107	89.612	0.432	0.903
	75	**49.326**	**93.283**	**0.507**	**0.876**	**48.539**	**88.852**	**0.428**	**0.905**
	100	50.619	94.020	0.511	0.874	49.901	89.455	0.431	0.904
Neuron	**50**	**49.326**	**93.283**	**0.507**	**0.876**	**48.539**	**88.852**	**0.428**	**0.905**
	75	51.645	94.734	0.515	0.872	50.941	90.014	0.434	0.902
	100	49.935	94.385	0.513	0.873	49.479	89.935	0.433	0.903
Hidden	1	49.326	93.283	0.507	0.876	48.539	88.852	0.428	0.905
Layer	2	49.138	93.790	0.510	0.875	47.841	88.432	0.426	0.906
	3	**49.224**	**93.352**	**0.507**	**0.876**	**48.194**	**88.504**	**0.426**	**0.906**
	4	48.112	93.428	0.508	0.876	47.401	88.603	0.427	0.905
Window	3	49.224	93.352	0.507	0.876	48.194	88.504	0.426	0.906
Size	6	44.059	82.626	0.449	0.903	44.100	83.426	0.402	0.916
	9	42.801	81.744	0.445	0.905	42.863	82.396	0.396	0.918
	12	**27.581**	**55.833**	**0.304**	**0.956**	**50.082**	**100.763**	**0.484**	**0.878**

By adjusting the hyperparameters for the optimizer, activation function, number of neurons, epochs, batch size, window size, and hidden layers, the LSTM prediction model was evaluated. Train and test data are used to evaluate the LSTM following the completion of the hyperparameter phase. In this study, it was determined that the train data had a portion of 70% and the test data had a portion of 30% the of GHI dataset. After conducting data training with 70% in the initial period, data testing is carried out on 30% at the end of the period sequentially. After testing on 30% of the data in the final period, the test results were validated using the evaluation MAE, RMSE, NRMSE, and R^2. Adam optimizer is used in this study to extract neural network parameters. It is an optimization approach based on stochastic gradients. Furthermore, it has demonstrated an efficacy in resolving sparse gradient domains and a deep machine learning issues. Consequently, when working with LSTM networks—which typically have several times more parameters than simple neural networks—this optimizer is a good option. Adam optimizer is more efficient at this task than standard stochastic gradient descent (SGD), which requires a lot of computing power to optimize such networks [36]. Then, the tanh activation function was chosen because it has the best LSTM prediction model performance between sigmoid and ReLU. Research for epoch 100 resulted in an accuracy of 80.1% for tanh, 78.81% for ReLU, and 75.76% for sigmoid [37].

In testing the LSTM model, the batch size testing is carried out from the smallest value to the largest to improve the performance of the LSTM. Batch size 5 to 15 experienced a gradual increase in performance, then from 15 to 20 experienced a decrease. This is the same as a previous research which shows that increasing the batch size does not necessarily improve the performance of the LSTM [38]. Next, the best epoch value is 75 from 25 and 50, then at 100, performance decreases. As in earlier studies on solar radiation LSTM prediction models, the test findings on the influence of epoch values show that continuously increasing the epoch does not necessarily improve the performance of the prediction model [19]. Furthermore, there was no discernible difference in the performance of LSTM when the number of neurons was increased or decreased. It is same as the research in [39]. In this study, the performance of LSTM did not increase with the continuous addition of hidden layers. Previous studies have shown that the adjustment ability improves with the number of hidden layers, however, as layers rise, the structure becomes more complex and challenging to train, and an overfitting scenario may occur, which would degrade the generalization ability [40]. And the last, test on the window size show that increasing the window size value can improve the performance of the prediction model, but a large window size can cause an increased computational load due to the addition of the calculated values [19]. On the other hand, based on research in [41], when the window size is expanded from one to fifteen days, the RMSE of the LSTM model significantly drops. The RMSE then begins to rise until the window size reaches sixty days. The RMSE decreases with increasing window size when the window size is greater than sixty days.

The best influence on the window size in the prediction model's performance was determined by analyzing the outcomes of LSTM model testing. With training data, window size 12 yields the greatest results; with testing data, window size 9 yields the best results for both models 1 and 2. For training data, the second rank performance is window size 9, and for testing data, it is window size 6 for two models. On the training data, the third rank performance is window size 6, and on the testing data, window size 3 for two models. On the training data, the last rank is window size 3, and on the testing data, it is window size 12 for two models. The results indicate that the model's performance significantly declines if the window is too large. Previous research related to solar radiation prediction also shows the performance of the best univariate model with NRMSE 0.213652 using the largest window size. This result shows that solar radiation depends on past observations. In this study, the best NRMSE result was 0.304 and could not exceed previous research [19]. In the MAE and RMSE results for train data at window size 12, it was found that the performance was better than the solar radiation prediction performance in previous research in [33].

5 Conclusion

Testing for the imputation of missing values at the data preprocessing was carried out using the interpolation method to produce MAE, RMSE, NRMSE, and R^2 evaluations, according to the findings of the research that was conducted in a prediction model using the data preprocessing stage on the GHI dataset using the LSTM. The best is PCHIP interpolation from linear and polynomial interpolation. In the test of filling in missing values from data that successfully eliminated outliers using RANSAC, the best test results were also obtained in PCHIP interpolation. The results of PCHIP interpolation on the GHI dataset were selected for use in training and testing prediction models using the LSTM algorithm. Outlier elimination testing using the RANSAC method resulted in five trials to obtain the percentage of data completeness. In the fifth experiment, all outliers were successfully eliminated. The elimination results in the fifth experiment were chosen to be used as a dataset that will be processed to fill in missing values and as a dataset for the GHI prediction model. The test results of the GHI solar radiation prediction model using the LSTM for datasets that underwent data preprocessing with the best evaluation of MAE, RMSE, NRMSE, and R^2 for train data and test data had better performance than the prediction model for datasets that did not undergo data preprocessing. The best evaluation was obtained from the LSTM prediction model using hyperparameters with a batch size 15, epoch 75, neuron 50, hidden layer 3, and window size 12. Future research can carry out deeper analysis regarding hyperparameters in the prediction model using LSTM for prediction data results that have negative values for the GHI solar radiation parameter. In addition, research using a multivariate model is carried out selectively on parameters that contribute significantly to solar radiation as input in order to anticipate excessive memory load in the model computing process.

Acknowledgments

We express our gratitude to the Ministry of Higher Education, Research and Technology of the Republic of Indonesia for its assistance in the form of a Master's Thesis Research Grant with contract number 107/E5/PG.02.00.PL/2024.

References

[1] W. Sawadogo, J. Bliefernicht, B. Fersch, S. Salack, S. Guug, B. Diallo, K. O. Ogunjobi, G. Nakoulma, M. Tanu, S. Meilinger, *et al.*, "Hourly global horizontal irradiance over west africa: A case study of one-year satellite-and reanalysis-derived estimates vs. in situ measurements," *Renewable Energy*, vol. 216, p. 119066, 2023.

[2] S. Rahman, S. Rahman, and A. B. Haque, "Prediction of solar radiation using artificial neural network," in *Journal of physics: conference series*, vol. 1767, p. 012041, IOP Publishing, 2021.

[3] P. Megantoro, M. A. Syahbani, S. D. Perkasa, A. R. Muzadi, Y. Afif, A. Mukhlisin, and P. Vigneshwaran, "Analysis of instrumentation system for photovoltaic pyranometer used to measure solar irradiation level," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 6, pp. 3239–3248, 2022.

[4] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead pv generation forecasting," *Applied Sciences*, vol. 9, no. 1, p. 204, 2019.

[5] A. Forstinger, S. Wilbert, A. R. Jensen, B. Kraas, C. F. Peruchena, C. A. Gueymard, D. Ronzio, D. Yang, E. Collino, J. P. Martinez, *et al.*, "Expert quality control of solar radiation ground data sets," in *SWC 2021: ISES Solar World Congress*, pp. 1037–1048, International Solar Energy Society, 2021.

[6] W. M. Organization, ed., *Guide to Meteorological Instruments and Methods of Observation Volume V*. BOCA RATON, FL: World Meteorological Organization, 2018.

[7] M. Bayray, Y. Gebreyohannes, H. Gebrehiwot, S. Teklemichael, A. Mustefa, A. Haileslassie, P. Gebray, A. Kebedom, and F. Filli, “Temporal and spatial solar resource variation by analysis of measured irradiance in geba catchment, north ethiopia,” *Sustainable Energy Technologies and Assessments*, vol. 44, p. 101110, 2021.

[8] R. N. Faizin, M. Riasetiawan, and A. Ashari, “A review of missing sensor data imputation methods,” in *2019 5th International Conference on Science and Technology (ICST)*, vol. 1, pp. 1–6, IEEE, 2019.

[9] M. Shcherbakov, A. Brebels, N. Shcherbakova, V. Kamaev, O. M. Gerget, and D. Devyatykh, “Outlier detection and classification in sensor data streams for proactive decision support systems,” in *Journal of Physics: Conference Series*, vol. 803, p. 012143, IOP Publishing, 2017.

[10] H. Hissou, S. Benkirane, A. Guezzaz, M. Azrour, and A. Beni-Hssane, “A novel machine learning approach for solar radiation estimation,” *Sustainability*, vol. 15, no. 13, p. 10609, 2023.

[11] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data,” *Frontiers in energy research*, vol. 9, p. 652801, 2021.

[12] C. C. Turrado, M. d. C. M. López, F. S. Lasheras, B. A. R. Gómez, J. L. C. Rollé, and F. J. de Cos Juez, “Missing data imputation of solar radiation data under different atmospheric conditions,” *Sensors*, vol. 14, no. 11, pp. 20382–20399, 2014.

[13] S. Shan, X. Xie, T. Fan, Y. Xiao, Z. Ding, K. Zhang, and H. Wei, “A deep-learning based solar irradiance forecast using missing data,” *IET Renewable Power Generation*, vol. 16, no. 7, pp. 1462–1473, 2022.

[14] I. Daut, Y. Irwan, I. Safwati, M. Irwanto, N. Gomesh, and M. Fitra, “Finding the outliers on solar radiation in northern malaysia, perlis,” *Asian Transactions on Engineering (ATE ISSN: 2221-4267)*, vol. 2, pp. 35–40, 2012.

[15] A. D. Călin, A. M. Coroiu, and H. B. Mureşan, “Analysis of preprocessing techniques for missing data in the prediction of sunflower yield in response to the effects of climate change,” *Applied Sciences*, vol. 13, no. 13, p. 7415, 2023.

[16] W. T. Handoko and A. Handayani, “Forecasting solar irradiation on solar tubes using the lstm method and exponential smoothing,” *J. Ilm. Tek. Elektro Komput. dan Inform*, vol. 9, no. 3, pp. 649–660, 2023.

[17] C. N. Obiora, A. N. Hasan, and A. Ali, “Predicting solar irradiance at several time horizons using machine learning algorithms,” *Sustainability*, vol. 15, no. 11, p. 8927, 2023.

[18] A. K. Mandal, R. Sen, S. Goswami, and B. Chakraborty, “Comparative study of univariate and multivariate long short-term memory for very short-term forecasting of global horizontal irradiance,” *Symmetry*, vol. 13, no. 8, p. 1544, 2021.

[19] M. C. Sorkun, Ö. D. Incel, and C. Paoli, "Time series forecasting on multivariate solar radiation data using deep learning (lstm)," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 1, pp. 211–223, 2020.

[20] M. Munoz-Organero, "Deep physiological model for blood glucose prediction in t1dm patients," *Sensors*, vol. 20, no. 14, p. 3896, 2020.

[21] E. Cho, T.-W. Chang, and G. Hwang, "Data preprocessing combination to improve the performance of quality classification in the manufacturing process," *Electronics*, vol. 11, no. 3, p. 477, 2022.

[22] N. M. Noor, M. M. Al Bakri Abdullah, A. S. Yahaya, and N. A. Ramli, "Comparison of linear interpolation method and mean method to replace the missing values in environmental data set," in *Materials science forum*, vol. 803, pp. 278–281, Trans Tech Publ, 2015.

[23] N. Fatimah, "Aplikasi interpolasi newton menggunakan borland delphi 5.0," *Jurnal Ilmiah Teknologi dan Rekayasa*, vol. 20, no. 1, 2015.

[24] R. Kumar, S. Bhattacharya, and G. Murmu, "Exploring optimality of piecewise polynomial interpolation functions for lung field modeling in 2d chest x-ray images," *Frontiers in Physics*, vol. 9, p. 770752, 2021.

[25] A. Jaffar, N. M. Thamrin, M. S. A. M. Ali, M. F. Misnan, A. I. M. Yassin, and N. M. Zan, "Spatial interpolation method comparison for physico-chemical parameters of river water in klang river using matlab," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 4, pp. 2368–2377, 2022.

[26] S. Jia, Z. Zheng, G. Zhang, J. Fan, X. Li, X. Zhang, and M. Li, "An improved ransac algorithm for simultaneous localization and mapping," in *Journal of Physics: Conference Series*, vol. 1069, p. 012170, IOP Publishing, 2018.

[27] O. Kaspi, A. Yosipof, and H. Senderowitz, "Random sample consensus (ransac) algorithm for material-informatics: application to photovoltaic solar cells," *Journal of cheminformatics*, vol. 9, pp. 1–15, 2017.

[28] P. Dey, E. Hossain, M. I. Hossain, M. A. Chowdhury, M. S. Alam, M. S. Hossain, and K. Andersson, "Comparative analysis of recurrent neural networks in stock price prediction for different frequency domains," *Algorithms*, vol. 14, no. 8, p. 251, 2021.

[29] S. Malakar, S. Goswami, B. Ganguli, A. Chakrabarti, S. S. Roy, K. Boopathi, and A. Rangaraj, "Designing a long short-term network for short-term forecasting of global horizontal irradiance," *SN Applied Sciences*, vol. 3, pp. 1–15, 2021.

[30] T. O. Hodson, "Root mean square error (rmse) or mean absolute error (mae): When to use them or not," *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.

[31] A. Jierula, S. Wang, T.-M. Oh, and P. Wang, "Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data," *Applied Sciences*, vol. 11, no. 5, p. 2314, 2021.

[32] D. G. da Silva, M. T. B. Geller, M. S. dos Santos Moura, and A. A. de Moura Meneses, "Performance evaluation of lstm neural networks for consumption prediction," *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 2, p. 100030, 2022.

[33] S. Liebermann, J.-S. Um, Y. Hwang, and S. Schlüter, "Performance evaluation of neural network-based short-term solar irradiation forecasts," *Energies*, vol. 14, no. 11, p. 3030, 2021.

[34] H. Liu, Y. Wang, and W. Chen, "Three-step imputation of missing values in condition monitoring datasets," *IET Generation, Transmission & Distribution*, vol. 14, no. 16, pp. 3288–3300, 2020.

[35] J. He, L. Yuan, H. Lei, K. Wang, Y. Weng, and H. Gao, "A novel piecewise cubic hermite interpolating polynomial-enhanced convolutional gated recurrent method under multiple sensor feature fusion for tool wear prediction," *Sensors*, vol. 24, no. 4, p. 1129, 2024.

[36] R. Solgi, H. A. Loaiciga, and M. Kram, "Long short-term memory neural network (lstm-nn) for aquifer level time series forecasting using in-situ piezometric observations," *Journal of Hydrology*, vol. 601, p. 126800, 2021.

[37] K. Vijayaprabakaran and K. Sathiyamurthy, "Towards activation function search for long short-term model network: A differential evolution based approach," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 2637–2650, 2022.

[38] M. Khosravi, B. M. Duti, M. M. S. Yazdan, S. Ghoochani, N. Nazemi, and H. Shabanian, "Multivariate multi-step long short-term memory neural network for simultaneous stream-water variable prediction," *Eng*, vol. 4, no. 3, pp. 1933–1950, 2023.

[39] A. W. Saputra, A. P. Wibawa, U. Pujianto, A. P. Utama, and A. Nafalski, "Lstm-based multivariate time-series analysis: A case of journal visitors forecasting," *ILKOM Jurnal Ilmiah*, vol. 14, no. 1, pp. 57–62, 2022.

[40] C. Liu, A. Zhang, J. Xue, C. Lei, and X. Zeng, "Lstm-pearson gas concentration prediction model feature selection and its applications," *Energies*, vol. 16, no. 5, p. 2318, 2023.

[41] H. Fan, M. Jiang, L. Xu, H. Zhu, J. Cheng, and J. Jiang, "Comparison of long short term memory networks and the hydrological model in runoff simulation," *Water*, vol. 12, no. 1, p. 175, 2020.