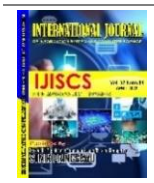

IJISCS (International Journal of Information System and Computer Science)
IJISCS, Volume 7, Number 3, December 2023
E ISSN: 2598-246X; P ISSN: 2598-0793, pp. 264-277
Accredited SINTA 4 Number 79/E/KPT/2023
https://jurnal.ftikomibn.ac.id/index.php/ijiscs

A HYBRID ARIMA-MLP ALGORITHM USING ARIMA AND MLP TO IMPROVE ESTIMATION MODEL PERFORMANCE IN SOLAR RADIATION SENSOR DATA

Alfin Syarifuddin Syahab[1], Arief Hermawan[2], Donny Avianto[3]
[1,2]Information Technology, Universitas Teknologi Yogyakarta
[3]Informatics, Universitas Teknologi Yogyakarta, Sleman
[1]Kabupaten Street No.Km. 5.5, Sendangadi, Mlati, Sleman
[1]*Stasiun Klimatologi D.I. Yogyakarta, Badan Meteorologi Klimatologi dan Geofisika*
[2,3]Siliwangi Street, Jombor, Mlati, Sleman

*Corresponding author
alfin.syahab@bmkg.go.id[1*]
ariefdb@uty.ac.id[2]
donny@uty.ac.idmail[3]

Article history:
Received November 23, 2023
Revised December 12, 2023
Accepted December 31, 2023

Keywords:
Solar Radiation;
Estimation;
ARIMA;
MLP;
ARIMA-MLP Hybrid;

Abstract
Various models have been developed to estimate solar radiation. Several additional models were created using improved machine learning. Currently, estimating solar radiation with the help of hybrid models is more efficient. In this research, the concepts of modeling for hybrid between the Autoregressive Integrated Moving Average (ARIMA) and the Multilayer Perceptron (MLP) are used to improve the performance of the ARIMA and MLP models in estimating solar radiation data from a pyranometer sensor. The test results of the estimation model based on the coefficient of determination (R^2) value and root mean square error (RMSE) show that the ARIMA model can provide a high coefficient of determination value in each data splitting scenario. The MLP estimation model shows a coefficient of determination value that is lower than the ARIMA model. On the other hand, MLP is able to improve the RMSE value in the ARIMA model in 70:30 and 90:10 splitting data. Furthermore, the ARIMA-MLP hybrid estimation model is able to improve the RMSE value of the ARIMA and MLP models even though the coefficient of determination value is not as good as the ARIMA model. This research shows that the ARIMA-MLP hybrid model is able to increase the accuracy value in RMSE compared to the ARIMA and MLP models and also increase the R^2 of previous research on the diffuse solar fraction prediction model. This study provides benefits to Badan Meterologi Klimatologi dan Geofisika (BMKG) by providing an accurate model to estimate solar radiation in drought predictions and provide appropriate of the public information in early warning of drought.

1.0 INTRODUCTION

Several measurement instruments using pyranometers have been developed to measure solar radiation parameters. These parameters are critical for atmospheric science analysis and renewable energy system design [1]. Automatic Weather Stations (AWS), are installed on land and have pyranometers that are used to record solar radiation levels [2]. Ground-based solar radiation measurements help solar energy projects and applications. These observations are

required to assess and enhance the accuracy of solar radiation data produced from satellite retrievals or numerical weather models, as well as to monitor the performance of solar panel installations. To obtain the requisite precision in solar resource data for solar power production projects and solar radiation projections, solar radiation measurement data must be made available [3]. This research discusses methods to support improving data quality by creating data estimation model to find data errors and gaps and make predictions.

Solar radiation estimation has significant impacts for many fields. Over time, various models have been developed to estimate solar radiation. Then, several additional models were created using improved machine learning. Currently, estimating solar radiation with the help of hybrid models is more efficient [4]. Data analysis becomes difficult if there are no observations. Missing values can cause problems such as lack of efficiency, difficulty handling and analyzing missing data, inaccurate estimates, and inefficient forecasting. Imputing missing values addresses the problem of handling complex patterns, which makes analysis easier by creating a complete data set. Although conventional imputation methods are easy to use, they introduce bias in the data. Under certain assumptions, modern and hybrid approaches are considered to have better performance [5]. The data collected is incomplete due to data gaps caused by delayed starts or early stops and measurement errors when measurements are taken. These errors occur mainly due to maintenance activities and battery failures and sometimes in the data logging process [6]. If these missing values cannot be filled in inaccurately, then existing analysis tools cannot be applied. If missing data is directly removed, a large amount of raw data will be lost thereby reducing the accuracy and reliability of the analysis results [7]. Solar radiation values play an important role in the recent hydrological drought. It is known that solar radiation is an important factor in evaporation. All meteorological readings known to influence drought must be known to determine its extent and take action. Any analysis or modeling requires complete and highly accurate data. If the data is invalid, it will provide wrong analysis, resulting in extreme climate mitigation errors in society, triggering the climate crisis [8]. This is an urgency to design a solar radiation sensor data estimation model to support the analysis of meteorological drought events to support appropriate climate mitigation.

In the energy sector, the ARIMA (Autoregressive Integrated Moving Average) model has been widely used because it is easy to use and versatile. The main advantages of this model are its accessibility and low computational complexity. At the same time, the possibility of incorporating the model into the theory and process structure is also an advantage. This, combined with the quality and reliability of the forecasts obtained, makes the ARIMA model one of the most popular methods for predicting time series values [9]. Besides ARIMA, this research uses the Multilayer Perceptron (MLP) method. MLP is a type of Artificial Neural Network (ANN) model that is widely applied in various fields. Since this particular type of neural network requires a desired output to be trained, it is called a supervised network and works as a simulator based on back propagation. this type of network attempts to create a model that accurately translates input to output using previous data [10]. ARIMA and ANN models are two statistical models and intelligent models that have been used in several applications to create hybrid models [11]. In this research, the basic concepts of modeling procedures for two ARIMA models and an MLP neural network are used to improve the performance of the ARIMA and MLP models in estimating solar radiation data from a pyranometer sensor installed on the automatic weather measuring instrument AWS (Automatic Weather Station) at Stasiun Klimatologi Daerah Istimewa Yogyakarta operated by BMKG.

In 2023 Woldegiyorgis et al. tested the performance of the ANN algorithm in estimating monthly solar radiation data with the statistical results of this metric, it was found that MAPE ranged from 1.554% to 7.343%, MSE ranged from 0.015 kWh/m^2/day to 0.127 kWh/m^2/day, and RMSE ranges from 0.124 kWh/m^2/day to 0.399 kWh /m^2/day. Research from Ho et al. in 2021 also estimated solar radiation data using ANN producing daily errors ranging from 0.06% to 2.04%, 0.08% to 5.88%, and 0.14% to 17.83 % respectively for the 30 day ANN model, 10 day ANN model, and 1 day ANN model. For a total of 30 days of predictions, the error percentages of the 30-day ANN model, 10-day ANN model, and 1-day ANN model are 0.25%, 1.67%, and 2.54%, respectively. This analysis assumes that all-day or 30-day data is not available, and the 30-day ANN model can predict the data with less than 3% error. For practical cases when only a small portion of data is lost, this ANN model is able to recover lost data with higher accuracy. Then

ARIMA is also used to carry out the ARIMA-MLP hybrid model to improve performance with low computational size [2]. Research from Qureshi et al. in 2022 tested the MLP model, on MLP with 20 hidden layers outperforming all other models for MLP 5 and 10 hidden layers as well as ARIMA in modeling and prediction purposes. The advantage of this research is that the ARIMA model gets confidence interval values of 95% and 90% [12]. In another study about MLP, hourly and daily diffuse solar fractions at Fez, Morocco, have been predicted using MLP models, which have been built and evaluated. The results show that the MLP model is a better predictor of diffuse solar fraction than the empirically tested models with 0.8896 of the coefficient of determination (R^2). This study has a purpose to get improvement on how to increase the coefficient of determination (R^2) between actual data and estimated data by using ARIMA-MLP hybrid from ARIMA and MLP with the coefficient of determination (R^2) and Root Mean Square Error (RMSE) evaluation in case of solar radiation data from sensor measurement. In addition, the analysis of evaluation of estimation model is equipped by the various proportion of data splitting for ratio of data training and data testing, consist of 70:30, 75:25, 80:20, 85:15, and 90:10 which it uses to analyze the best performance based on the difference in data segmentation.

2.0 THEORETICAL FRAMEWORK

2.1. Completeness Check

For some applications, data completeness is a requirement. A comparison must be made between the observations actually received and the observations that are expected to be received [13]. Completeness check is carried out to determine the completeness of the data. Data completeness addresses the problem of missing sensor data values for a certain period of time, which may be determined based on application requirements. Based on the time interval between two sequentially recorded sensor data values, the count of a particular data value expected for a specified time period is symbolized by $\mathrm{count}(\mathrm{Exp}_v)$. If the sensor does not produce a value or a NULL data value within the specified time period, then the sensor is marked as lost. Because of these missing values, the actual number of recorded data values is denoted by $\mathrm{count}(\mathrm{Obt}_v)$ where the value is less than $\mathrm{count}(\mathrm{Exp}_v)$. Mathematically, data completeness C is calculated using equation 1.

$$\mathrm{C} = 1 - \frac{\mathrm{count}(\mathrm{Exp}_v) - \mathrm{count}(\mathrm{Obt}_v)}{\mathrm{count}(\mathrm{Exp}_v)} \qquad (1)$$

Sensor data completeness refers to the extent to which sensor data values are not lost within a certain period of time. In this approach, completeness is defined in such a way that the more missing values the less complete the data obtained. In checking data completeness, the program can classify each series into one of five quality levels, namely: complete quality (completeness = 100%), high quality (>90% completeness), medium quality (50–90% completeness), low quality (0–50% completeness) or undetermined quality (no estimate of completeness) [14].

2.2. Range Check

Aspects of climatological conditions in a region are things that support range checks. The specified range check refers to the climatological profile obtained through objective analysis and model forecasts. Thresholds can be determined using knowledge of climatology and varying observation errors [15]. Count counts of time-series data were examined to identify missing, negative, invalid, and outlier data points. Minimum and maximum range validation is carried out to ensure values are within the sensor measurement limits [6]. The measurement range check of limit value given in Table 1.

Table 1. Range Check Limit Values

No	Parameter	Symbol	Lower Limit	Upper Limit
1	Wind speed (m/s)	ws_avg	0	50
2	Air temperature (°C)	tt_air_avg	5	45
3	Realtive humidity (%)	rh_avg	5	100
4	Air pressure (mb)	pp_air	800	1050
5	Solar radiation (W/m^2)	sr_avg	0	3000

Range check helps to eliminate the outlier value. The range value is obtained from statistical calculations using long historical data series. The range values in the AWS Center system that can be applied generally to all AWS sites in Indonesia [16].

2.3. Stationery Check

A unit root test technique called the augmented Dickey-Fuller test (ADF) evaluates the stationarity of a time series. ADF test, a commonly used unit root test, can determine whether a time series is stationary by calculating the statistics of the parameters of a time series model and comparing them with the ADF distribution [17]. In addition, a function called autocorrelation function (ACF) shows the relationship between the observations at one time and the observations at previous times. The ACF also shows the autocorrelation coefficient, which is a measure of the relationship between observations at different times [18]. Then, the Partial Autocorrelation Function (PACF) is a partial correlation between observations at time t and observations at previous times. Also, the differential process is used to turn non-stationary data into stationary data (an ARIMA implementation requirement) [19].

2.4. ARIMA Model

The ARIMA model, often known as the Box-Jenkins approach, is one of the most effective classical time series models for short-term forecasting. This model [ARIMA (p, d, q)] is made up of three parts: auto regression (AR), which tells us how the series is dependent on its past lag and is denoted by a parameter p, moving average (MA), which tells us about the dependency of error terms on past lags and is denoted by q, and the integrated part, which is used when the series is not stationary and is denoted by d. This methodology consists of four procedures: model identification, parameter estimates, diagnostic checking, and forecasting. The series is tested for stationarity using certain tests, and the model is identified using the data correlogram. It then moves on to the estimation process, following which the estimated models are evaluated using diagnostic checking; if the candidate model meets the criteria, the model is used for forecasting [12]. The following equation 2 shows the ARIMA model's generic form.

$$Z_t = \mu + (\phi_1 + 1)Z_{t-1} - 1 + (\phi_2 - \phi_1)Z_{t-2} + \cdots$$
$$\cdots + (\phi_p - \phi_{p-1})Z_{t-1} - \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \quad (2)$$

Where Zt is data in period t, μ has constant value, then $\phi_2, \ldots, \phi_p$ are the values of *autoregressive* parameter, also $\theta_1, \ldots, \theta_q$ are the *moving average* parameter. Identify the model, plot time series data, and verify the mean and variance's stationarity [20].

2.5. MLP Model

The multilayer perceptron (MLP) machine learning model is widely regarded as one of the most adaptable mathematical algorithms in terms of prospective applications and precision in time series prediction. The MLP model is useful for approximating any continuous, nonlinear, differentiable, and limited function. The MLP model is composed of an input layer, an output layer, and one or more hidden layers. Artificial neurons are utilized to transfer information from one layer to the next. Depending on the topic of interest, hidden layers collect information from the input layers and then transmit it in a nonlinear function to another region [12]. The input, hidden, and output layers combine to make MLP model. It uses neural network operations with specific adaptive weighting to process input data. The combination of the activation function, input weighting, and bias is the MLP ANN output [21]. It is put forward as follows in equation 3.

$$\mathrm{Y} = f\left(\sum_{n=1}^{n} w_i x_i + b\right) \quad (3)$$

Neuron output is denoted by Y, input by x, weighting by w, and bias by b. The activation function is defined as equation 2. The sigmoid, hyperbolic tangential, and rectified linear unit functions are used in this study. The equations 4,5 and 6 express these functions.

$$F_s = f(x) = \frac{1}{1+e^x} \quad (4)$$
$$F_t = tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (5)$$
$$F_r = f(x) = \max(0, x) \quad (6)$$

F_s stands for sigmoid function. F_t represents for hyperbolic tangential function. F_r means for rectified linear unit function. The steps involved in setting up an MLP ANN model include data selection, model training, and validation [21].

2.6. ARIMA-MLP Hybrid Model

The autoregressive integrated moving average solar radiation forecasting model (ARIMA) with multilayer perceptron (MLP) is discussed in this section. The linear ARIMA model is often used to predict time-series data. Nonlinear residuals are produced by the linear model. A nonlinear model, multiple layer perception networks (MLP), is used to analyze these residuals. The residual values of the nonlinear model will have a linear structure [22]. First, the ARIMA model is applied to the linear solar radiation data. The nonlinear residuals from ARIMA model are trained by the multilayer perceptron. Solar radiation data have been continuously analyzed using the ARIMA model, and respective nonlinear residual errors are successfully resolved with the help of the MLP model. Figure 1 is the framework of ARIMA-MLP Hybrid model.

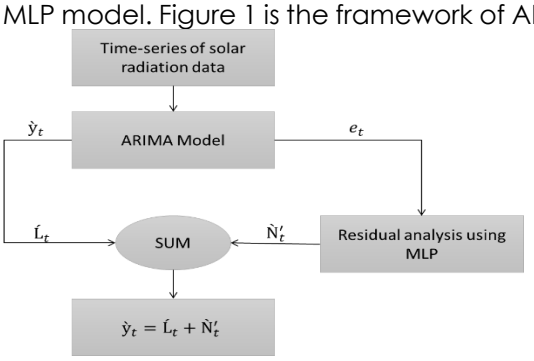

Figure 1. The Framework of ARIMA-MLP Hybrid Model

The ARIMA is employed in the first permutation of the ARIMA-MLP model to simulate the linear component, in accordance with the series models' process. If we assume that e_t represents the ARIMA model's residual at time t, then the formula is displayed in equation 7.

$$e_t = \hat{y}_t - \acute{L}_t \qquad (7)$$

Where $\acute{L}_t$ is the ARIMA model's output at time t. The nonlinear patterns were retained in the residual of the ARIMA model. Thus, nonlinear relationships can be found in the second stage by modeling residuals using MLPs. The MLP model for the residuals with n input nodes is as follows in equation 8.

$$e_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) + \varepsilon_t \quad \Rightarrow \grave{N}'_t = \tilde{e}_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) \qquad (8)$$

Where $\grave{N}'_t$ is the predicted value at time t from the MLP model on the residual data, ε_t is the random error, and f is a nonlinear function determined by the MLP.

2.7. Model Evaluation

The coefficient of determination is defined as the proportion of the variance in the dependent variable that can be predicted by the independent variables. In addition, if there are outliers to be recognized, MSE can be employed. In fact, MSE is excellent for assigning higher weights to such points. MSE and RMSE have a monotonically connected relationship (via the square root). A regression model ordering based on MSE will be equivalent to a model ordering based on RMSE [23]. This study uses R^2 and RMSE to evaluate the performance model. These formulas R^2 and RMSE are displayed in equation 9 and 10.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\hat{Y} - Y_i)^2} \quad (9)$$

The observation value to i is Y, while the prediction value to i is X. The $\hat{Y}$ is the mean of all the observation data gathered for the dataset. The technique predicts the Y for the matching X in the solar radiation dataset. Moreover, the number of periods employed in the computations is m.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \qquad (10)$$

By dividing the total squared prediction errors by the number of prediction time data, the Root Mean Square Error (RMSE) can be computed. The accuracy of the model increases with decreasing RMSE value. the RMSE calculation was performed to identify the optimal solar radiation estimation model.

3.0 METHOD FRAMEWORK

This research was carried out in several stages. The stages consist of data collecting, data preprocessing, algorithms testing, evaluation, and model of estimation. The diagram of processing stages can be seen in Figure 2.

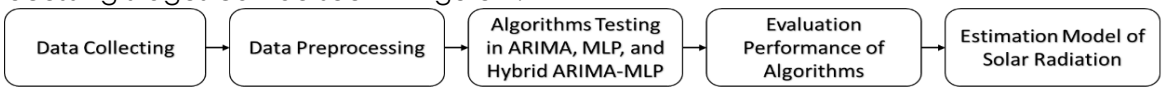

Figure 2. The Solar Radiation Estimation Model Processing Stages

Firstly, the step begins with data collecting, it needs solar radiation data from pyranometer sensor. The sample using site named Automatic Weather Station (AWS) in Stasiun Klimatologi Daerah Istimewa Yogyakarta (Staklim) Mlati, Sleman, Special Region of Yogyakarta. Then, the raw datasets that collected successfully consist of one month in ten minutes' interval include several weather parameters; air temperature, air pressure, wind speed, relative humidity, and global horizontal irradiance. Secondly, the raw dataset should be handled by preprocessing data. Before checking the valid data, the process needs calculate the completeness percentage of data in each weather parameters. The result of data preprocessing is dataset with validation process using the tolerance range measurement data. Algorithm testing scenarios include variations in data segmentation. By using data splitting between training data and testing data with a ratio of 70:30, 75:25, 80:20, 85:15, and 90:10.

In the subsequence stage, dataset is analyzed to stationary check. Dataset which they are stationary, will be optimized by using ARIMA and MLP testing. The stationary data also can be check by ACF and PACF. Then, dataset is calculated to estimate the global horizontal irradiance using ARIMA, MLP, and hybrid of ARIMA-MLP. The final stage is an estimation model of solar radiation data is trained using performance evaluation such as; R^2 and RMSE.

4.0 RESULANTS

3.1. Data Collection

Dataset was collected from AWS Staklim Mlati in the Sleman, Yogyakarta Special Region with latitude coordinates -7.73118 and longitude 110.3537, then altitude of 182 meters. The location of this AWS can be seen in Figure 3.

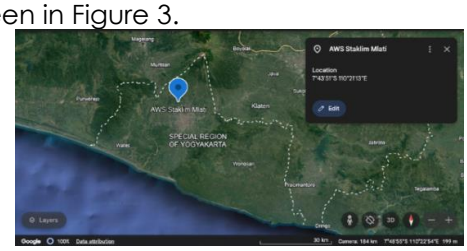

Figure 3. Location of AWS Staklim Mlati

Raw dataset was obtained from the AWS database for the period January 2022 with a data interval of ten minutes contains 4465 rows of raw data . From the AWS data, sensor data for temperature, relative humidity, wind speed, air pressure and global horizontal irradiance are obtained. To maintain the accuracy of sensor data, this AWS equipment is routinely field calibrated by Badan Meteorologi Klimatologi dan Geofisika (BMKG).

The raw data was downloaded into excel format, it consists of table. The attributes have date time, air temperature, wind speed, air pressure, relative humidity, and solar radiation in average each ten minutes. Table 2 shows AWS Staklim Mlati dataset.

Table 2. Dataset of AWS Mlati

Date Time	ws_avg	tt_air_avg	rh_avg	pp_air	sr_avg
1/1/2022 0:00	0.017	25.44	86	990.2285	255.9
1/1/2022 0:10	1.003	25.88	83.2	990.3819	252.6
1/1/2022 0:20	0.423	26.21	81.8	990.3892	235.3
1/1/2022 0:30	0.158	26.48	81.1	990.5874	388
1/1/2022 0:40	0.796	26.87	78.08	990.6114	403

The air global horizontal irradiance sensor is a KIPP & Zonen mode CMP3. Sensor data is recorded by using Campbell Scientific CR6 logger. There are some parameters used, ws_avg is

wind speed in m/s unit, tt_air_avg is average of air temperature in celcius unit, rh_avg is relative humidity in percentage unit, pp_air is air pressure in milibar.

3.2. Completeness Check

The first result obtained from data preprocessing is data completeness, this helps in knowing the initial condition of the data, whether there is null data or not, which will affect the data testing process. Checking completeness data can be helped using the graph in Figure 4 below.

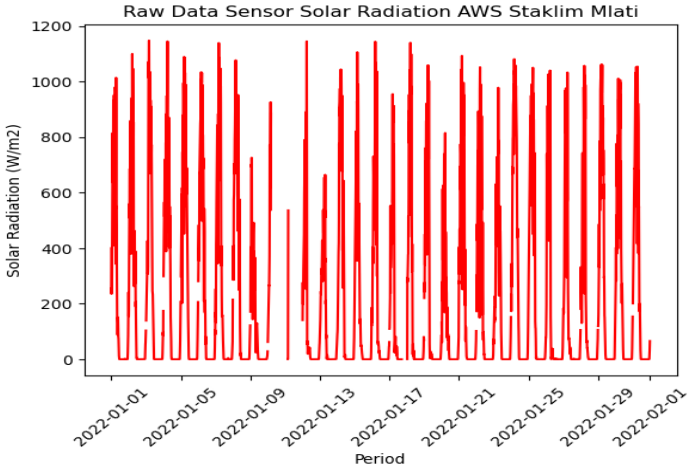

Figure 4. The Raw Data of Solar Radiation Sensor AWS Staklim Mlati

Raw data visualization plots help find missing pieces of data over a certain period of time. In detail, percentage calculations can be carried out by comparing the recorded data with the ideal data that should be obtained. Figure 4 shows the results of checking data completeness in percent for each parameter.

Table 4. Completeness Check

Parameter	Completeness (%)
ws_avg	92.743561
tt_air_avg	92.743561
rh_avg	92.743561
pp_air	92.743561
sr_avg	92.743561

At this stage, all parameters have the same completeness percentage value above 90%. This figure shows that there is not significant of missing data. This is according to previous research [14], if it is classified according to data quality, it is included in the high quality label.

3.2. Range Check

Range check is a way to deal with outlier values in sensor measurements when there is a measurement error. In this research, the way to eliminate outlier values uses the range check method to eliminate values that are outside the specified range. Table 4 shows the percentage of valid data after eliminating outlier data.

Table 5. Range Check

Parameter	Data Valid (%)
ws_avg	92.743561
tt_air_avg	92.743561
rh_avg	90.951848
pp_air	92.743561
sr_avg	92.698768

Valid data shows that the data is within the data range with a measurement range based on historical data for each parameter. The results for the relative humidity and solar radiation parameters contained outlier data which were eliminated from the dataset. Figure 5 is an image of the dataset graph after the data elimination process.

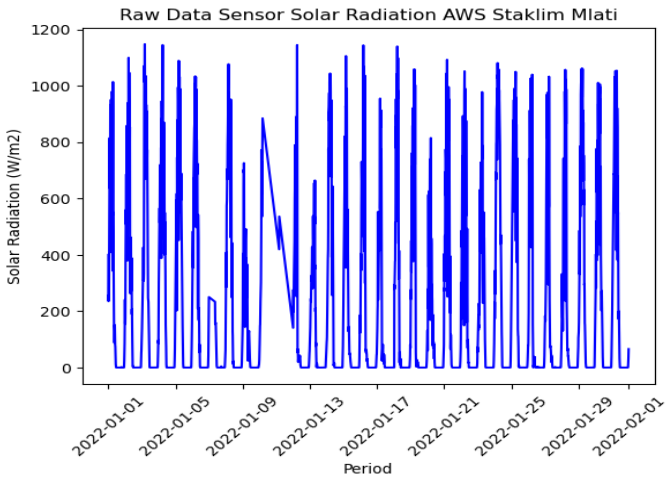

Figure 5. The Result of Range Check in Dataset

After a range check is carried out, empty data and data outside the applicable range are eliminated. In the visualization plot, it can be seen the complete dataset and within the applicable measurement range, now the dataset consist of 4059 rows of data. The dataset after eliminated by range check will be continued to the next steps in ARIMA and MLP, also in Hybrid ARIMA-MLP.

3.3. Data Splitting

In order to segment the solar radiation data for the ARIMA, MLP, and ARIMA-MLP hybrid models, the training and testing parts are divided. 70%, 75%, 80%, 85%, and 90% make up the percentage distribution of training data for each scenario. A specific proportion from the start of the data to a predetermined limit is taken to divide the data. On the other hand, 30%, 25%, 20%, 15%, and 10% of testing data sharing is allocated to each scenario. A specific percentage from the end of the data to a predetermined boundary is taken to divide the data.

3.4. Stationery Check

Data on solar radiation is plotted every ten minutes. A graphic of the original solar radiation data from January 2022 is displayed using PACF and ACF in Figure 6. Because the correlation value degradation in this image moves stationery down from one lag to the next, it appears to be a sequence of stationary data in variance.

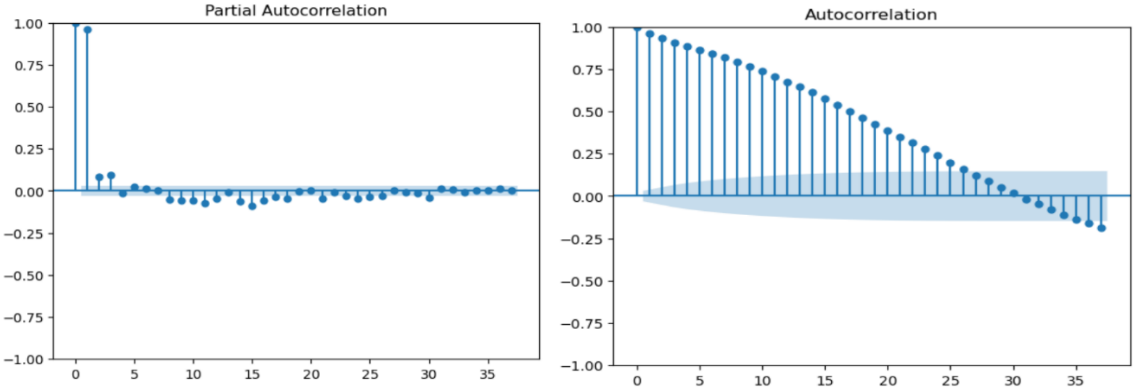

Figure 6. PACF and ACF Graphic of Original Solar Radiation Sensor Data

The data from the sun radiation sensor is differenced once in the ensuing steps. We then test the original data's stationary characteristic and the data's differencing data. ACF and PACF graphics of differencing data are displayed in Figure 7.

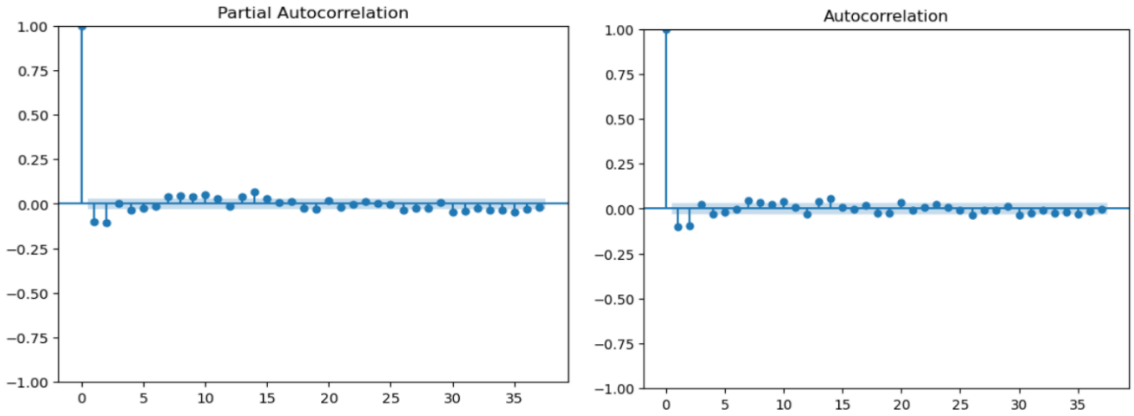

Figure 7. PACF and ACF Graphic of First Differenced Data

The fifth lag correlation is still strong, as the ACF figure demonstrates, thus the moving average order is between 1 and 5. Although the third lag partial correlation is substantial, as seen by the PACF image, the auto regression order falls between 1-3. Significant correlation value decline is depicted in the ACF figure. Table 6 displays both data's stationary test results.

Table 6. Stationary Data Checking

Test	ADF	
	p-value	**characteristic**
Original Data	$3.5902038991479224\times10^{-22}$	Weakly Stationery
Differenced Data	$1.8614173540331579\times10^{-22}$	Stationery

Data that was previously first order differenced is now more stationary. The original data has a lower p-value than the differenced data. The original data is less stationery than the differencing data.

3.5. Estimation Model using ARIMA

Then, by determining the AIC values, these ARIMA-based order estimates are verified. The AIC values for every ARIMA model run with p values = range (1,11), d values = range (1,2), and q values = range (1,11) are listed in Table 7.

Table 7. AIC Value for ARIMA Model

ARIMA Model	AIC
ARIMA(1, 1, 1)	47717.546
ARIMA(1, 1, 2)	47706.809
ARIMA(1, 1, 5)	47680.725
ARIMA(1, 1, 6)	47668.029
ARIMA(1, 1, 7)	47661.617
ARIMA(4, 1, 7)	47658.903
ARIMA(5, 1, 10)	47657.481
ARIMA(7, 1, 2)	47653.676
ARIMA(9, 1, 3)	47630.607

ARIMA (9,1,3) has the minimum AIC value, according to Table 7. The model selected for the sun radiation sensor data estimator is ARIMA (9,1,3).

3.6. Estimation Model using MLP

MLP ANN input is selected by Principal Component Analysis (PCA) method. Figure 7 is variance ratio percentage graphic as PCA result. This Figure 8 states that only four input variables are significant to be injected into the model.

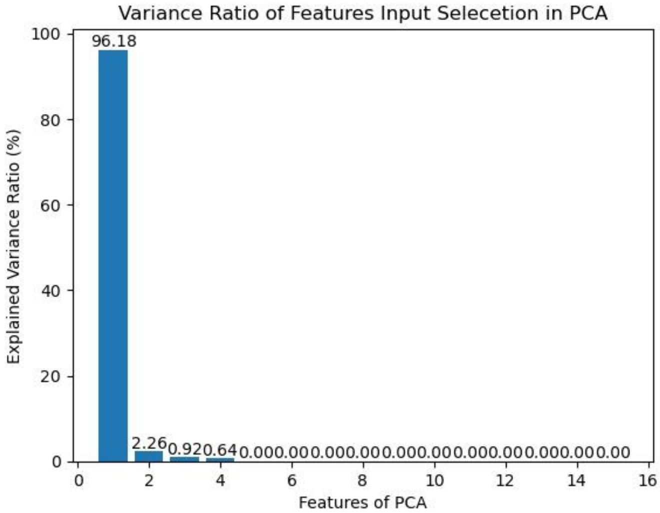

Figure 8. Proportion of PCA Result at Variance Ratio

Significant ratio values have been established for a number of parameters. Based on this graph, the inputs that will be examined using MLP are sun radiation and relative humidity. Four important inputs for the MLP ANN model are displayed in Table 8.

Table 8. Significant Input Parameters

Significant input	Eigen Value	Variance Ratio (%)
Solar radiation (t-2)	0.581229758	96.18
Solar radiation (t-1)	0.575159073	2.26
Solar radiation (t-3)	0.575134398	0.92
Relative humidity (t-1)	0.0137194877	0.64

Compared to other inputs, these inputs have higher eigen values. Data on solar radiation that is trailing dominate the significant input. Relative humidity is also important since it is connected to the dynamics of solar radiation measurements. All weather parameter measurements are, however, primarily influenced by the lagged solar radiation intensity.

The MLP ANN model was developed using an earlier version of the air temperature predicting model. In this study, the MLP ANN model estimation was used in the previous study to estimate the air temperature, and in the current study, the model is used to improve the estimate of the solar radiation data model. Table 8 provides specifics on the detained major inputs that are simulated for this model. This model is fully described in Table 9.

Table 9. The Strucutre of MLP Model

Model Version	Layer	Neuron	Activation Function
Roy (2020)	Layer 1	16	Relu
	Layer 2	16	Relu
	Layer 3	1	Linear

MLP model is encouraged from previous research about air temperature estimation that is applied for solar radiation data in this study. Adam optimizer is used to train and evaluate the models using 32 batches and 100 epochs [24]. Table 10 show about the result of MLP Model in time step and loss training.

Table 10. Time Step and Loss Training

Scenario	Data Splitting	Epoch	Time Step	Training Loss
1	70:30	100/100	862 us/step	0.0056
2	75:25	100/100	945 us/step	0.0054
3	80:20	100/100	1 ms/step	0.0055
4	85:15	100/100	962 us/step	0.0054
5	90:10	100/100	1 ms/step	0.0052

In order to test the MLP, 100 epochs were run with 32 batches, generating time and training loss data for each batch in various of scenarios. The proportion of testing data compared to training data is displayed for each scenario. From these scenarios, the first scenario is the fastest

time and has the highest number of loss training from the other scenarios. In addition, the fifth scenario has the higher value of time step and lowest number of loss training.

3.7. Estimation Model ARIMA-MLP

In the ARIMA-MLP hybrid method experiment, the algorithm calculation process was carried out by providing ACF and PACF plots on the residual values. Residual values are obtained from the ARIMA model after fitting the data. This is the difference between the actual value and the value predicted by the model. In Figure 9, it can be seen that each lag shows a significant difference in value, which means the data is stationary.

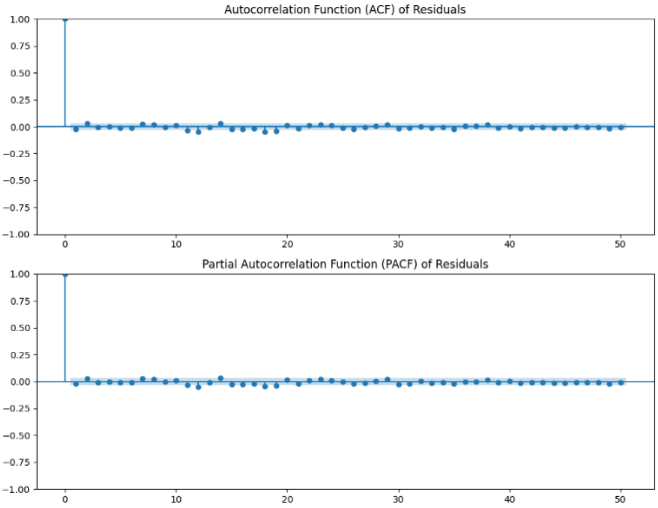

Figure 9. PACF and ACF Graphic of Residuals Data

The next stage is that the residual values in the ARIMA model results are plotted over a certain time period to see the trend and range of residual values obtained. Figure 10 shows the residual value against the zero value line for the time sample from 28 January 2023 to 31 January 2023.

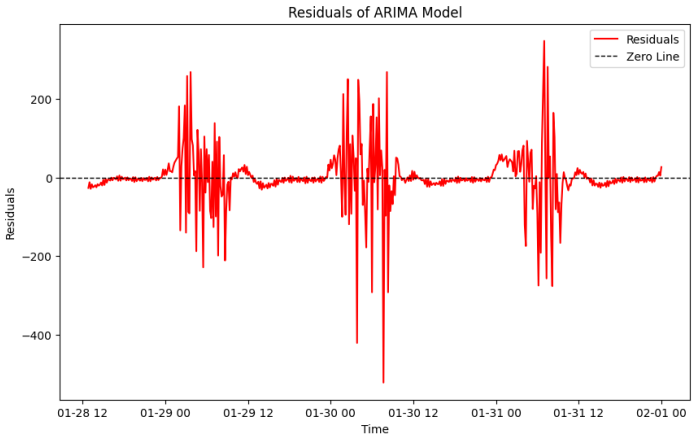

Figure 10. Plotting of Residuals of ARIMA Model

After obtaining the residual samples, ARIMA test results were also obtained for the best p, q and d order values with the smallest AIC values. Table 11 explains that the best ARIMA model (10,0,0) was obtained with an AIC value of 47635.767.

Table 11. AIC Value for ARIMA-MLP Model

ARIMA Model	AIC
ARIMA(1, 0, 0)	47667.993
ARIMA(8, 0, 0)	47653.631
ARIMA(9, 0, 0)	47640.917
ARIMA(10, 0, 0)	47635.767

After obtaining the residues that have been tested with ARIMA, these residues are used as input to the MLP model. The test results are in the form of a time table and training loss in each batch. Table 12 shows the ARIMA-MLP results in the epoch training process.

Table 12. Time Step and Loss Training

Scenario	Data Splitting	Epoch	Time Step	Training Loss
1	70:30	100/100	862 us/step	0.0056
2	75:25	100/100	945 us/step	0.0054
3	80:20	100/100	1 ms/step	0.0055
4	85:15	100/100	962 us/step	0.0054
5	90:10	100/100	1 ms/step	0.0052

Results show training time steps and losses on the ARIMA-MLP model for five different data splitting scenarios. The first scenario has the fastest time step and the largest training loss compared to the other scenarios. Furthermore, the fifth scenario has a longer time than the other scenarios but the training loss value is smallest than the others.

3.8. Evaluation Performance of Estimation Model

Evaluation performance of estimation models using the R^2 parameter to shows the contribution value of the independent variable in the estimation model to the dependent variable and RMSE to measure the level of accuracy of the estimation results of a model. Table 13 shows the performance evaluation results on ARIMA, MLP, and hybrid ARIMA-MLP models

Table 13. Evaluation Performance of ARIMA, MLP, and Hybrid ARIMA-MLP

Scenario	Data Splitting	ARIMA		MLP		Hybrid ARIMA-MLP	
		R^2	RMSE	R^2	RMSE	R^2	RMSE
1	70:30	0.997	470.988	0.943	80.716	0.975	52.874
2	75:25	0.996	471.052	0.936	530871036.09	0.983	43.407
3	80:20	0.996	449.946	0.941	431165.02	0.978	46.647
4	85:15	0.997	468.237	0.939	526761612.97	0.984	41.366
5	90:10	0.997	490.579	0.951	77.514	0.988	37.715

The estimate model's performance evaluation findings demonstrate that while the ARIMA model cannot outperform other models in terms of accuracy, it can produce a higher coefficient of determination value than the MLP and Hybrid ARIMA-MLP models. However, in contrast to the ARIMA model, MLP is unable to raise the determination coefficient value, and the accuracy value does not produce the anticipated outcomes due to the presence of extremely high values in the second, third, and fourth scenarios. It is believed that this is the result of the data not being standardized. Nonetheless, it is evident that the ARIMA-MLP hybrid model performs better because to its low RMSE value and high R^2 value. This demonstrates how hybrid ARIMA-MLP can enhance accuracy performance in the estimate model of solar radiation data. This result also improves the previous evaluation of the R^2 value which was 0.8896 in the prediction model for diffuse solar fraction data in one hour and one day intervals [25], then compared with the hybrid ARIMA-MLP model which had R^2 values recorded at 0.975, 0.983, 0.97, 0.984, and 0.988 in the training data ratio. and testing data 70:30, 75:25, 80:20, 85:15, and 90:10. This research provides benefits to the BMKG in having an accurate model in estimating solar radiation sensor data to support the analysis and prediction of meteorological drought phenomena and to the public in getting accurate information regarding early warning of extreme climate of droughts.

5.0 CONCLUSION

In this research, the solar radiation sensor data estimation model can be carried out using ARIMA, MLP and hybrid ARIMA-MLP models. The test results of the estimation model based on the coefficient of determination value show that the ARIMA model can provide a high coefficient of determination value in each data splitting scenario, for variations in differences in data splitting it has no significant effect. The MLP estimation model shows a coefficient of determination value that is lower than the ARIMA model. On the other hand, MLP is able to improve the RMSE value in the ARIMA model in 70:30 and 90:10 splitting data but in other splitting data it shows large and unexpected values. Furthermore, the ARIMA-MLP hybrid estimation model is able to improve the RMSE value of the ARIMA and MLP models even though the coefficient of determination value is not as good as the ARIMA model but still shows quite good values. This research shows that the ARIMA-MLP hybrid model is able to contribute to

increasing the accuracy value in RMSE compared to the ARIMA and MLP models in estimating solar radiation sensor data. The research results of the solar radiation estimation model using hybrid ARIMA-MLP increase the R^2 evaluation value of previous research on the diffuse solar fraction prediction model at intervals of one hour and one day. This research provides benefits to BMKG by providing an accurate model to estimate solar radiation sensor data to support predictions of meteorological drought phenomena and provide appropriate information to the public regarding early warning of extreme climate droughts. Further research can utilize the data normalization stage before testing using the MLP model or similar algorithms such as artificial neural networks.

REFERENCES

[1] S. Oyelami, S. 1 Oyelami, N. A. Azeez, S. A. Adedigba, O. J. Akinola, and R. M. Ajayi, "A Pyranometer for Solar Radiation Measurement-Review," *Adeleke Univ. J. Eng. Technol.*, vol. 3, no. 1, pp. 61–68, 2020, [Online]. Available: https://www.researchgate.net/publication/349210517

[2] K. C. Ho, B. H. Lim, and A. C. Lai, "Recovery of the Solar Irradiance Data using Artificial Neural Network," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1755-1315/721/1/012006.

[3] A. Forstinger *et al.*, "Expert quality control of solar radiation ground data sets," in *Proceedings - ISES Solar World Congress 2021*, International Solar Energy Society, 2021, pp. 1037–1048. doi: 10.18086/swc.2021.38.02.

[4] S. Gupta *et al.*, "Estimation of Solar Radiation with Consideration of Terrestrial Losses at a Selected Location—A Review," *Sustainability (Switzerland)*, vol. 15, no. 13. Multidisciplinary Digital Publishing Institute (MDPI), Jul. 01, 2023. doi: 10.3390/su15139962.

[5] G. Chhabra, J. Ranjan, and V. Vashisht, "A Review on Missing Data Value Estimation Using Imputation Algorithm," 2019. [Online]. Available: https://www.researchgate.net/publication/334695903

[6] M. Bayray *et al.*, "Temporal and spatial solar resource variation by analysis of measured irradiance in Geba catchment, North Ethiopia," *Sustain. Energy Technol. Assessments*, vol. 44, Apr. 2021, doi: 10.1016/j.seta.2021.101110.

[7] Z. Gao, W. Cheng, X. Qiu, and L. Meng, "A Missing Sensor Data Estimation Algorithm Based on Temporal and Spatial Correlation," *Int. J. Distrib. Sens. Networks*, vol. 2015, 2015, doi: 10.1155/2015/435391.

[8] E. E. Başakin and M. Özger, "Missing Data Imputation for Solar Radiatıon by Deep Neural Network," *Eur. J. Sci. Technol.*, May 2022, doi: 10.31590/ejosat.1085022.

[9] E. Chodakowska, J. Nazarko, Ł. Nazarko, H. S. Rabayah, R. M. Abendeh, and R. Alawneh, "ARIMA Models in Solar Radiation Forecasting in Different Geographic Locations," *Energies*, vol. 16, no. 13, Jul. 2023, doi: 10.3390/en16135029.

[10] Enung, H. Kasyanto, R. R. Sari, and M. F. Lubis, "Application of Multilayer Perceptron (MLP) Method for Streamflow Forecasting (Case Study: Upper Citarum River, Indonesia)," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics, 2023. doi: 10.1088/1755-1315/1203/1/012032.

[11] M. Khashei and Z. Hajirahimi, "A Comparative Study of Series ARIMA/MLP Hybrid Models for Stock Price Forecasting," *Commun. Stat. - Simul. Comput.*, vol. 48, no. 9, pp. 2625–2640, Oct. 2019, doi: 10.1080/03610918.2018.1458138.

[12] M. Qureshi, M. Daniyal, and K. Tawiah, "Comparative Evaluation of the Multilayer Perceptron Approach with Conventional ARIMA in Modeling and Prediction of COVID-19 Daily Death Cases," *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/4864920.

[13] World Meteorological Organization, *Guide to Climatological Practices 2018 edition*, no. WMO-No. 100. 2018.

[14] S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrpides, "CheckV assesses the quality and completeness of metagenome-assembled viral genomes," *Nat. Biotechnol.*, vol. 39, no. 5, pp. 578–585, 2021, doi: 10.1038/s41587-020-00774-7.

[15] S. Good *et al.*, "Benchmarking of automatic quality control checks for ocean temperature profiles and recommendations for optimal sets," *Front. Mar. Sci.*, vol. 9, no. February, pp. 1–25, 2023, doi: 10.3389/fmars.2022.1075510.

[16] Kedeputian Bidang Inskalrekjarkom, "INOVASI," *BMKG*, Jakarta, 2022.
[17] J. Wang, T. Ji, and M. Li, "A combined short-term forecast model of wind power based on empirical mode decomposition and augmented dickey-fuller test," *J. Phys. Conf. Ser.*, vol. 2022, no. 1, 2021, doi: 10.1088/1742-6596/2022/1/012017.
[18] G. M. Tinungki, "The analysis of partial autocorrelation function in predicting maximum wind speed," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 235, no. 1, 2019, doi: 10.1088/1755-1315/235/1/012097.
[19] U. A. Yakubu and M. P. A. Saputra, "Time Series Model Analysis Using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for E-wallet Transactions during a Pandemic," *Int. J. Glob. Oper. Res.*, vol. 3, no. 3, pp. 80–85, 2022, doi: 10.47194/ijgor.v3i3.168.
[20] I. Syahrini, R. Radhiah, W. F. Damanik, N. Sciences, U. S. Kuala, and B. Aceh, "Application of the Autoregressive Integrated Moving Average (ARIMA) Box-Jenkins Method in Forecasting Inflation Rate in Aceh," *Transcendent J. Math. Appl. ISSN*, vol. 2, no. 1, pp. 27–33, 2023.
[21] H. S. Wicaksana *et al.*, "Air Temperature Sensor Estimation on Automatic Weather Station Using ARIMA and MLP," vol. 14, no. 2, 2022.
[22] V. Rajalakshmi and S. G. Vaidyanathan, "Hybrid Time-Series Forecasting Models for Traffic Flow Prediction," *Promet - Traffic - Traffico*, vol. 34, no. 4, pp. 537–549, 2022, doi: 10.7307/ptt.v34i4.3998.
[23] 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 Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.
[24] P. S. Saketh, R. Rohit, and B. Suneetha, "Weather Forecasting using Machine Learning," *6th Int. Conf. Inven. Comput. Technol. ICICT 2023 - Proc.*, pp. 13–18, 2023, doi: 10.1109/ICICT57646.2023.10134218.
[25] B. Ihya, A. Mechaqrane, R. Tadili, and M. N. Bargach, "Prediction of hourly and daily diffuse solar fraction in the city of Fez (Morocco)," *Theor. Appl. Climatol.*, vol. 120, no. 3–4, pp. 737–749, May 2015, doi: 10.1007/s00704-014-1207-y.