Future Sustainability
Open Access Journal
ISSN 2995-0473

May 2024| Volume 02 | Issue 02 | Pages 20-26

Journal homepage: https://fupubco.com/fusus

https://doi.org/10.55670/fpll.fusus.2.2.3

Future Publishing LLC

Article

Analyzing meteorological parameters using Pearson correlation coefficient and implementing machine learning models for solar energy prediction in Kuching, Sarawak

Geoffrey Tan[1], Hadi N. Afrouzi[1]*, Jubaer Ahmed[2], Ateeb Hassan[1], Firdaus M-Sukki[2]

[1]Faculty of Engineering Computing and Science, Swinburne University of Technology, Sarawak, 93350, Kuching, Malaysia
[2]School of Engineering and Built Environment, Edinburgh Napier University, Merchiston Campus, 10 Colinton Road, Edinburgh, EH10 5DT, UK

ARTICLE INFO

Article history:
Received 03 January 2024
Received in revised form
02 February 2024
Accepted 12 February 2024

Keywords:
Energy modeling, Machine Learning,
Pearson Correlation Coefficient,
Regression techniques,
Solar energy prediction, Solar forecasting

*Corresponding author
Email address:
hafrouzi@swinburne.edu.my

DOI: 10.55670/fpll.fusus.2.2.3

ABSTRACT

Solar energy is one of the clean renewable energy sources that can offset the rising consumption of fossil fuels. However, the meteorological parameters, such as solar irradiance, ambient and solar module temperatures, relative humidity, etc., constantly change, and so does the solar power generation. Such variations cause instability in the power grid operation due to injecting an unpredicted amount of power. Hence, solar energy prediction models capable of learning from past weather data and predicting future energy generation are highly desired for grid operation and planning. The objective of this study is to determine the suitable meteorological parameters for the solar energy prediction model based on the Pearson correlation coefficient and to implement them in different machine learning models. It is found in this study that five meteorological parameters, namely Air temperature, cloud opacity, global tilted irradiance, relative humidity, and zenith angle, correlate highly with solar energy generation. Later, based on the correlations, four machine-learning models were implemented to predict the solar power for Kuching, Sarawak. The accuracy of the models is measured through standard matrices such as root mean square error, mean square error, mean absolute error, and R-squared value.

1. Introduction

The entire world is going through an energy transition due to replacing fossil fuels with renewable energy sources. According to Holechek et al. [1], the world's energy consumption is increasing quickly and is predicted to increase by around 60% by 2030; during that time, fossil fuels will continue to dominate the world's energy use. However, the emissions from massive fuel combustion have resulted in global warming and depletion of the ozone layer, which caused significant climate change across the world. Therefore, many countries are introducing alternative green & renewable energy sources to meet the energy demand. Solar energy is one of the preferred candidates, abundant in many parts of the world. Malaysia is a tropical country with diverse energy resources, including fossil fuels and various renewable energy sources [2]. According to Vaka et al. [3], just 8% of Malaysia's total energy is now produced from renewable sources, despite its commitment to reach 20% by 2025. Moreover, the Malaysian Investment Development Authority (MIDA) states that solar energy, hydroelectricity, and biomass are Malaysia's only flourishing renewable energy technologies. As it is in the equatorial zone, Malaysia has advantages in developing its solar energy technologies due to its abundance in this region [4]. Malaysia receives mean daily solar radiation of 4.7 to 6.5 kWh/m^2 for roughly 10 hours each day, which is a large amount of solar energy throughout the year [5]. The government has also introduced several new incentive schemes such as net energy metering (NEM), feed-in tariff (FIT), large-scale solar (LSS), and self-consumption (SELCO) after realizing the potential of solar in

the country. As shown in Figure 1, solar energy in Malaysia is around the higher spectrum in the radiation chart. However, the uncertainty in solar energy prediction due to its dependency on atmospheric parameters makes market penetration challenging [6].

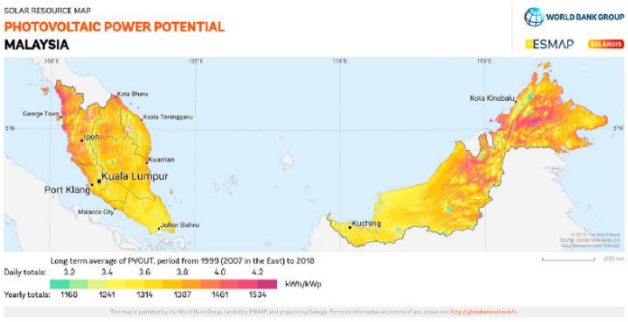


Figure 1. Solar Irradiance profile in Malaysia

The uncertainty in meteorological parameters such as solar irradiance, ambient temperature, solar hours, humidity, and cloud cover affect solar energy production significantly. Unexpected variations in a PV system's output may raise operating costs for the electricity system by increasing the need for primary reserves and posing potential risks to the reliability of the electricity supply [7]. Not only that, according to Alam et al. [8], one of the main impediments to integrating renewable energies into the grid is that their power supply is erratic and intermittent. This causes difficulties in grid management. Therefore, solar power forecasting becomes crucial for assuring the grid's reliability in addition to providing an ideal unit commitment and cost-effectiveness dispatch. According to Chakchak and Cetin [9], various techniques and tools, such as empirical models that use mathematical relationships between data measurements to estimate solar power, machine learning algorithms/models, and remote sensing tools, have been proposed in recent years to estimate solar power. Because the weather has a significant impact on solar energy generation and can have a variety of effects on it; therefore, it is essential to understand the link between various meteorological features before developing the energy forecast model [10]. It is difficult to predict from just one climate component because solar energy and power generation of solar panels depend on several weather parameters [11]. On the other hand, Jebli et al. [10] further state that it is vital to determine which aspects of the related meteorological variables include the most pertinent data to make reliable projections. Besides, meteorological parameters vary significantly from country to country. Thus, the ML model developed for one country would not be suitable for other countries, especially if they are not under the same climate conditions.

Several research studies have been conducted on studying the correlation between different meteorological parameters to better understand the relationship between each parameter. Hossain and Mahmood [12] used the Pearson Product-Moment Correlation Coefficient (PPMCC) to calculate the correlation coefficient for seven different weather indicators in the field of solar energy forecast. Similarly, Jebli et al. [10] used the Pearson correlation coefficient to identify a connection between 8 meteorological variables. On the other hand, Kumari and Toshniwal [6] conducted a study on the forecast of solar irradiance whereby they stated the significance of feature selection and analysis in machine learning, which not only improves the performance of the machine learning models on a high-dimensional dataset by lowering complexity and computing time but also aids in the elimination of unnecessary features. They identified the most important variables by analyzing eight weather parameters using the Pearson correlation coefficient. Meanwhile, in the study of solar radiation prediction, Huang et al. [13] used historical information from a continuous time series obtained by PPMCC to analyze the relationship between the sun radiation intensity and six other meteorological parameters. In another study, Ojo [14] considered air temperature, solar radiation, and wind speed data from the National Aeronautics and Space Administration and evaluated ten existing models. Besides, Zhu et al. [15], evaluated the correlations between solar irradiance intensity, atmospheric density, cloudiness, wind speed, relative humidity, and ambient temperature using the Pearson correlation coefficient. However, in the context of the Borneo region, Malaysia, no major study is found in the literature to examine meteorological parameters in depth to find the correlation with solar power generation. Thus, this study aimed to find suitable meteorological parameters that are highly correlated to solar power generation. Besides, this study also develops four ML algorithms based on the findings of highly correlated meteorological parameters. The remaining part of this paper is categorized into four sections. A summary of the preliminary steps and the methodology used to carry out this study, such as data preparation and pre-processing, are provided in sections 2 and 3. The results obtained will be presented and discussed in section IV before concluding this research paper by summarizing this study's key findings in section 4.

2. Data sets and preprocessing

2.1 Meteorological dataset

The historical weather dataset for Kuching, Sarawak is obtained from Solcast [16]. Solcast provides historical data from 2007 to as little as seven days prior in the form of Time Series, Typical Meteorological Year (TMY), and Monthly Averages. The Time Series data is chosen as it is a historical record of weather data for a specified location whereby the data is recorded at an interval of 1 hour. It provides data on parameters such as global horizontal irradiance, direct normal irradiance, cloud opacity, temperature, wind speed, etc. The meteorological dataset obtained contains data for the location of the city of Kuching. Kuching, located between the coordinates of 1.5535°N and 110.3593°E, is the capital and the most populous city in Sarawak, Malaysia. The dataset was selected from 2007 to 2021.

2.2 Data preprocessing

It is necessary to clean and normalize the supplied data before identifying the correlation. The obtained correlation may be inaccurate since there could be some random and non-stationary components in the historical weather data brought on by changing weather conditions and uncertainty. Moreover, not all the data provided are useful, in this case referring to the data during night whereby the solar

irradiance is 0, which results in no solar power being generated. Hence, all the rows where the solar irradiance data equals 0 are removed since they are not required. The other values of irradiance and temperature are used to calculate the solar power generated using the single-diode model [17]. The flowchart of the coding process is shown in Figure 2. Based on the flowchart for the data pre-processing process, as shown in Figure 2, the libraries required are first imported before loading the dataset obtained. The PV module parameters are then initialized before the calculation process begins. For this project, the chosen PV module was Samsung SDI PV-MBA1BG247, and the specification is given in Table 1. To calculate how much solar energy could be produced from the given weather data, the global irradiance and the temperature was taken from the dataset.

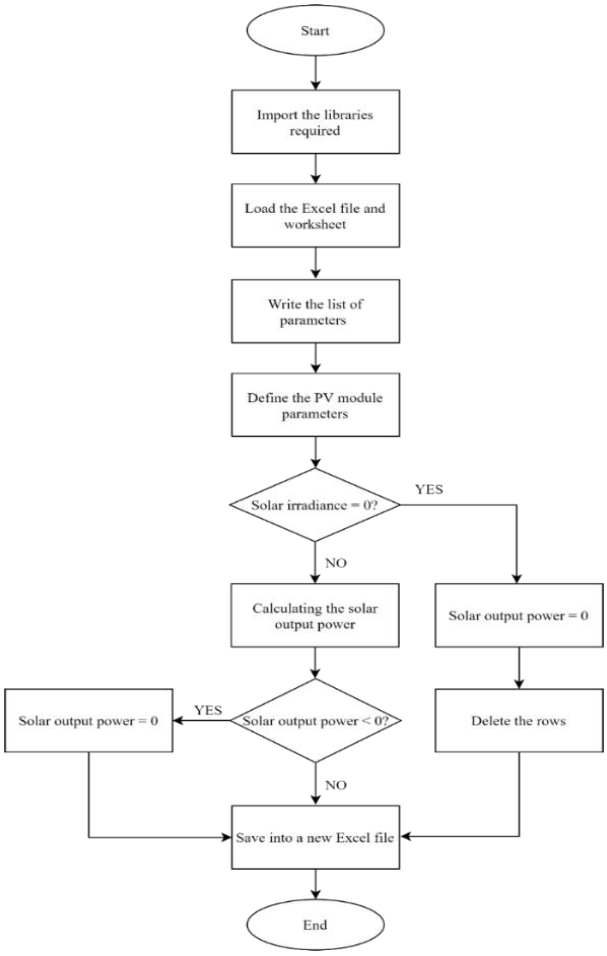


Figure 2. Data preprocessing flowchart

The single diode model equations for the PV module used to calculate the PV power generation are presented in Eq (1-7). The irradiance value from the dataset is used to find the value of the maximum current IMPP at MPP using Eq (1). Then, using the Irradiance values, photocurrent I_{ph} is calculated. The thermal voltage constant is found using the Eq (3) where the temperature value was taken from the dataset. Values from Eq (2) and (3) is placed in Eq (4) to find the open circuit voltage. It is found in the PV specification that VMPP is related to VOC by a multiplier of 0.8112. That relation is utilized to calculate the VMPP by using the V_{oc} from Eq (4). Then PMPP¬ is calculated using Eq (6). Using this process, a new column was added to the dataset, which represents solar power, which is the power produced under that hour of the dataset.

$$I_{mpp} = (I_{MPP\_STC} \times \frac{G}{G_{STC}}) \quad (1)$$

$$I_{ph} = (I_{PV\_STC} + K_I \Delta T)\frac{G}{G_{STC}} \quad (2)$$

$$V_T = \frac{akT}{q} \quad (3)$$

$$V_{oc} = N_s V_T \ln\left(\frac{I_{ph}}{I_{SD}}\right) = N_s\left(\frac{akT}{q}\right)\ln\left(\frac{I_{ph}}{I_{SD}}\right) \quad (4)$$

$$V_{mp} = 0.8112 V_{oc} \quad (5)$$

$$P_{MPP} = V_{MPP} I_{MPP} \quad (6)$$

Table 1. PV module specifications

Parameters	Value
Light generated current, I_{PV_STC}	8.7912 A
Maximum power current, I_{MPP_STC}	8.21A
Open circuit voltage, V_{OC_STC}	37.5 V
Maximum Power Voltage, V_{MPP_STC}	30.1 V
Short circuit current coefficient, K_I	0.52753
Diode saturation and diffusion current, I_{SD}	2.5469e-10 A
Diode ideality factor, a	1.0032
Series resistance, R_s	0.34878 Ω
Shunt resistance, R_{sh}	273.0475 Ω
Number of cells, Ns	60
Solar irradiance at STC, G_{STC}	1000 W/m^2
The temperature at STC, T_{STC}	25°C or 298.15 K
Boltzmann constant, k	1.3807e-23 J/K
Electron charge, q	1.6022e-19 C

2.3 Data correlation

The correlation between different weather parameters is analyzed and studied using the Pearson correlation coefficient (PCC). PCC is a strong tool to understand the linear dependency between two variables. The Pearson correlation coefficient (PCC), assesses the degree and direction of a linear relationship between two random variables. Formally, the covariance of the two variables divided by the product of their standard deviations (which acts as a normalization factor) is what makes up the Pearson correlation coefficient of two variables, X and Y. It can be equivalently represented by the following formula:

$$r_{xy} = \frac{\sum(x_i-\bar{x})\sum(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \quad (7)$$

where $\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{N} x_i$ denotes the mean of x, and $\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{N} y_i$ represents the mean of y.

To linear transformations of either variable, the coefficient r_{xy} is invariant and ranges from -1 to 1. Additionally, the PCC indicates how strongly the two variables, x and y, are

associated linearly, with the correlation coefficient being positive for directly related variables and negative for inversely related ones. If $r_{xy} = 0$, the variables are regarded as being uncorrelated. The stronger the indicators of closeness to a linear relationship are, the closer $|r_{xy}|$ is to 1. The correlation coefficient's absolute value increases with the strength of the link between solar energy and weather variables, whereby a positive correlation shows a rising linear relationship and vice versa. The analyzed weather parameters are air temperature, surface pressure, cloud opacity, dew point temperature, precipitable water, global tilted irradiance, relative humidity, azimuth angle, 10m wind direction, 10m wind speed, and zenith angle.

3. Results and discussion

Prior to developing the machine learning models for solar energy prediction, correlation among different weather parameters was calculated using the Pearson correlation coefficient (PCC). The different weather parameters being investigated include surface pressure, air temperature, dew point temperature, 10m wind speed, global tilted irradiance, precipitable water, relative humidity, azimuth angle, 10m wind direction, cloud opacity, zenith angle, and solar power. A correlation matrix that shows the correlation between different weather parameters is presented in Figure 3.

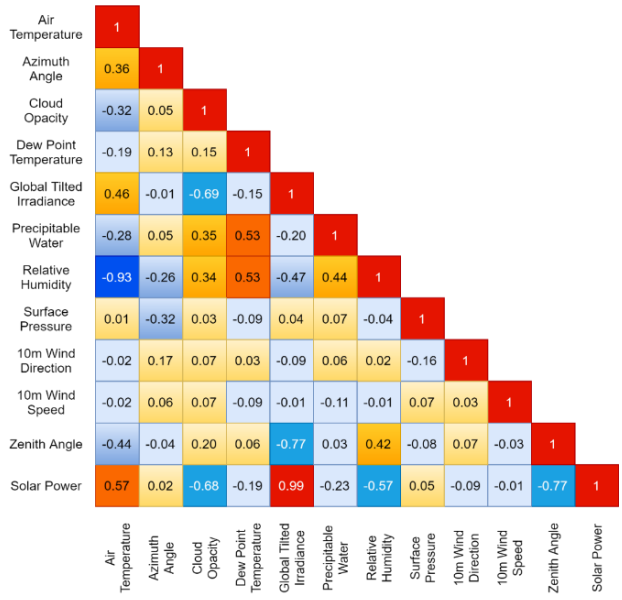


Figure 3. Correlation matrix among the weather parameters

3.1 Correlation matrix

The correlation matrix provided valuable insight into which parameters have a strong correlation with solar power generation.

- Air temperature and global tilted irradiance strongly correlate with solar power.
- Cloud opacity, relative humidity, and zenith angle strongly correlate negatively with solar power.
- Global tilted irradiance strongly correlated negatively with the cloud opacity and zenith angle.
- Dew point temperature positively correlated with precipitable water and relative humidity.
- Relative humidity illustrated a strong negative correlation with air temperature.

Some of the selected correlation diagrams are illustrated in Figure 4.

3.2 Prediction models and comparisons

After analyzing the correlation between different weather parameters and solar power, five parameters are chosen to develop the ML models. These are namely air temperature, global tilted irradiance, cloud opacity, relative humidity and zenith angle. Four different MLA was developed namely Multiple Linear Regression (MLR), Random Forest Regression (RFR), Support Vector Regression (SVR), and Lasso Regression (LR). The reason for selecting these four ML models is that they are frequently seen as more interpretable since it is simpler to comprehend how the model came to a specific prediction.

On the other hand, the correlations between meteorological parameters and solar power output can be difficult to explain using Support Vector Machine (SVM) and Artificial Neural Networks (ANN) because they tend to be more "black boxes". Moreover, SVM and ANN can be prone to overfitting, especially when dealing with noisy or insufficient data. In contrast, simpler models like MLR and LR are less likely to overfit, and RFR also provides a level of regularization that helps mitigate overfitting. The graphs presented in Figure 5 illustrate the predicted and actual values of solar power from each machine-learning model. The red plot represents the predicted values, whereas the blue plot represents the actual values of solar power.

The RMSE, MSE, MAE, and R-squared values for each machine learning model are listed in Table 2 to summarize their performance. It can be observed that the results obtained from the MLAs, the MLR, RFR, and LR presented a good accuracy in terms of R-squared values of 99.99%, 99.42%, and 99.42%, respectively. SVR was the least accurate among these models, with an R-squared value of 98.78% and a longer prediction time than other models. Regarding prediction errors, RFR was the most accurate model with minimal errors and less than 0.5W of power output. In contrast, the SVR has the highest prediction error with RMSE, MSE, and MAE values of 8.72W, 76.05W, and 5.91W power output, respectively.

Since the dataset utilized in this study contains intricate and non-linear linkages between datasets, the RFR and SVR are both very good at handling these connections. Nevertheless, the PCC's adoption to identify the highly correlated meteorological indicators is warranted because weather conditions cause their non-stationary and unpredictable processes. Furthermore, it is worth noting that the MLR and LR present identical results in each aspect. The MLR identifies the line of best fit for the five parameters selected from the Pearson correlation coefficient. In contrast, the LR identifies the regression coefficients of the parameters to produce the most accurate prediction with minimal errors.

The machine learning models' performance is being compared with the models designed by Markovics and Mayer. They designed four different models, namely Linear Regression (LR), Random Forest Regression (RFR), Support Vector Regression (SVR), and Artificial Neural Network (ANN), to predict solar energy in Morocco and Brazil.

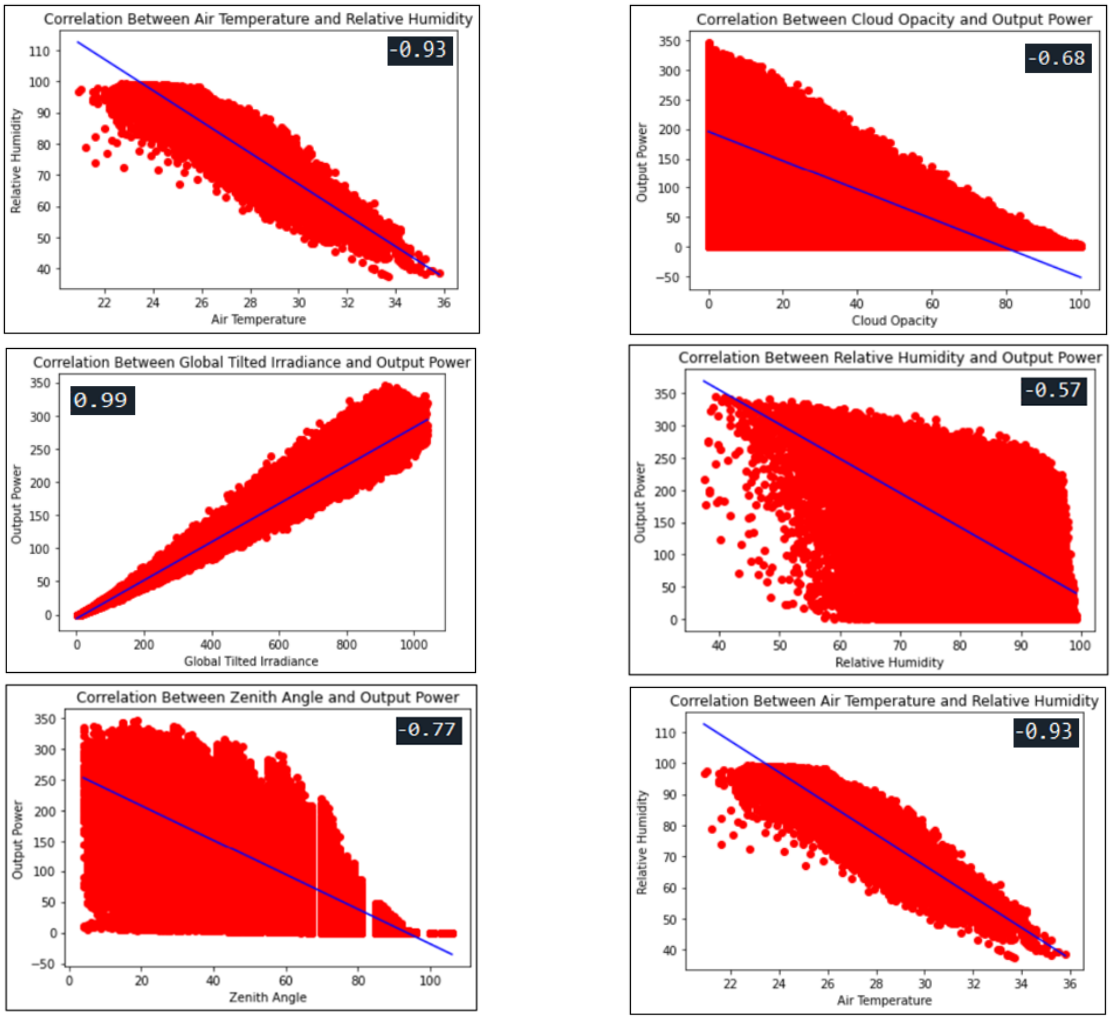


Figure 4. Selected correlation figures between different meteorological parameters

Table 2. Comparison of four ML models

Model	RMSE (W)	MSE (W)	MAE (W)	R-Squared (%)
MLR	6.00	35.99	4.19	99.42
SVR	8.72	76.05	5.91	98.78
RFR	0.37	0.14	0.19	99.99
LR	6.00	35.99	4.19	99.42

Since ANN is not included in this research, the performance of the remaining three models is reviewed. In Morocco, the RF is the most accurate model, and LR is the least accurate model, whereas in Brazil, the RF is the most accurate model, and SVR is the least accurate model. The performance of the models designed in this research agrees with the models designed by Markovics and Mayer for prediction in Brazil. However, like feature extraction and identification, there are no complete right or wrong results as this process depends on the collected data and how the models are being trained to make predictions.

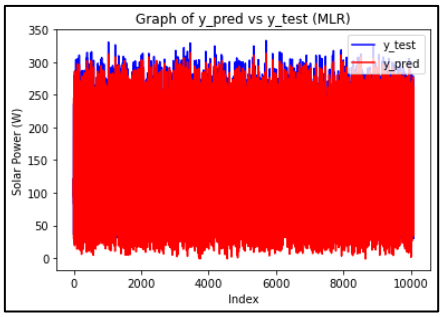
4. Conclusions

This work identified the relationship between different meteorological parameters and solar power using the Pearson correlation coefficient. Analysis of these relationships reveals that various meteorological parameters demonstrate various correlations with one another, resulting in the following observations:

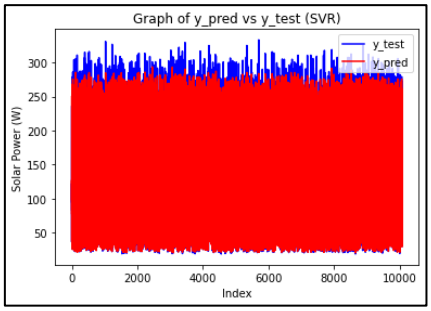
- Air temperature and global tilted irradiance strongly correlate positively with solar power.
- Cloud opacity, relative humidity, and zenith angle strongly correlate negatively with solar power.

Based on such correlations, four different machine learning models: MLR, RFR, SVR, and LR, were implemented, and their

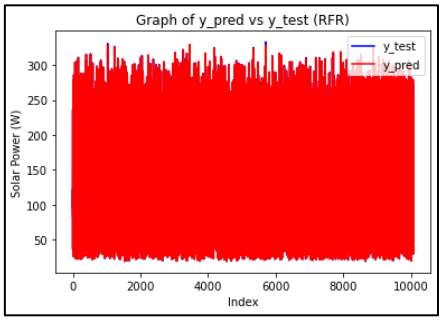
prediction accuracy was determined using performance metrics such as Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared value. It was concluded that from all machine learning models, MLR, RFR, and LR presented good accuracy in R-squared values of 99.99%, 99.42%, and 99.42%, respectively in a Malaysian context. Regarding prediction error, RFR was the most accurate model with minimal errors, which are less than 0.5W of power output. In contrast, the SVR has the highest prediction errors with RMSE, MSE, and MAE values of 8.72W, 76.05W, and 5.91W power output, respectively. In a nutshell, RFR outperformed all the other models with the highest prediction accuracy and the lowest prediction errors in the context of Kuching, Sarawak, Malaysia.



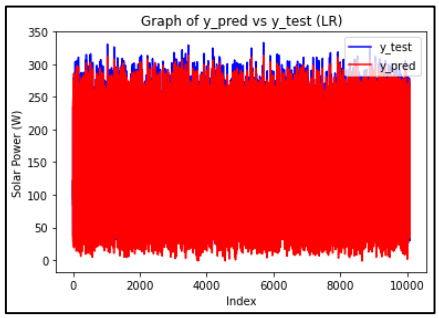
(a) Solar power graph with predicted and measured values using MLR



(b) Solar power graph with predicted and measured values using SVR



(c) Solar power graph with predicted and measured values using RFR.



(d) Solar power graph with predicted and measured values using LR.

Figure 5. Test and the predicted data using four ML models

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] J. L. Holechek, H. M. Geli, M. N. Sawalhah, and R. Valdez, "A global assessment: can renewable energy replace fossil fuels by 2050?," Sustainability, vol. 14, no. 8, p. 4792, 2022.

[2] M. Wong, H. N. Afrouzi, A. Hassan, E. Jayamani, J. Tavalaei, J. Sunarso, & K. Mehranzamir, "Design and Techno-Economic Analysis of a Hydrogen-Based Micro Hydro-Solar Hybrid Energy System for Sustainable Energy Access: A Case Study in Sri Aman, Sarawak." IJEETC, vol. 13, no. 1, 2024.

[3] M. Vaka, R. Walvekar, A. K. Rasheed, and M. Khalid, "A review on Malaysia's solar energy pathway towards carbon-neutral Malaysia beyond Covid'19 pandemic," Journal of cleaner production, vol. 273, p. 122834, 2020.

[4] F. S. M. Chachuli, N. A. Ludin, M. A. M. Jedi, and N. H. Hamid, "Transition of renewable energy policies in Malaysia: Benchmarking with data envelopment analysis," Renewable and Sustainable Energy Reviews, vol. 150, p. 111456, 2021.

[5] Cheng, B. W. Z., Mehranzamir, K., Afrouzi, H. N., & Hassan, A. (2022). Feasibility analysis and economic viability of standalone hybrid Systems for Marudi Electrification in Sarawak, Malaysia. Future Energy, 1(2), 28-45.

[6] P. Kumari and D. Toshniwal, "Deep learning models for solar irradiance forecasting: A comprehensive review," Journal of Cleaner Production, vol. 318, p. 128566, 2021.

[7] M. Sun, C. Feng, and J. Zhang, "Probabilistic solar power forecasting based on weather scenario generation," Applied Energy, vol. 266, p. 114823, 2020.

[8] M. S. Alam, F. S. Al-Ismail, A. Salem, and M. A. Abido, "High-level penetration of renewable energy sources into grid utility: Challenges and solutions," IEEE Access, vol. 8, pp. 190277-190299, 2020.

[9] J. Chakchak and N. S. Cetin, "Investigating the impact of weather parameters selection on the prediction of solar radiation under different genera of cloud cover: A case-study in a subtropical location," Measurement, vol. 176, p. 109159, 2021.

[10] I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," Energy, vol. 224, p. 120109, 2021.

[11] D. Markovics and M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," Renewable and Sustainable Energy Reviews, vol. 161, p. 112364, 2022.
[12] M. S. Hossain and H. Mahmood, "Short-term photovoltaic power forecasting using an LSTM neural network and synthetic weather forecast," Ieee Access, vol. 8, pp. 172524-172533, 2020.
[13] C.-J. Huang, Y. Ma, and Y.-H. Chen, "Solar Radiation Forecasting based on Neural Network in Guangzhou," in 2020 International Automatic Control Conference (CACS), 2020: IEEE, pp. 1-5.
[14] O. S. Ojo, "Evaluation of photovoltaic solar power using the different operating temperature models over a tropical region," Energy Systems, pp. 1-23, 2023.
[15] H. Zhu, H. Chen, W. Zhu, and M. He, "Predicting Solar cycle 25 using an optimized long short-term memory model based on sunspot area data," Advances in Space Research, vol. 71, no. 8, pp. 3521-3531, 2023.
[16] "https://solcast.com/data-for-researchers." (accessed 15 June, 2021).
[17] J. Ahmed and Z. Salam, "An Enhanced Adaptive P&O MPPT for Fast and Efficient Tracking Under Varying Environmental Conditions," IEEE Transactions on Sustainable Energy, vol. 9, no. 3, pp. 1487-1496, 2018, doi: 10.1109/TSTE.2018.2791968.