SOLAR IRRADIANCE FORECASTING FOR MALAYSIA USING MULTIPLE REGRESSION AND ARTIFICIAL NEURAL NETWORK

Poh-Leng Yew & Yih Hwa Ho*

Centre for Telecommunication Research & Innovation, Fakulti Kejuruteraan Elektronik & Kejuruteraan Komputer, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia

*Email: yihhwa@utem.edu.my

ABSTRACT

The installed capacity of solar photovoltaic (PV) globally continues to rise. In Malaysia, the monthly average daily solar radiation is 4,000-5,000 Wh/m², with the average daily sunshine duration ranging from 4 to 8 h. However, the output of solar energy is related to solar irradiance, which lacks stability due to weather variation. Therefore, solar irradiance forecasting has become an important resource for network grid operators to control the output of solar PV energy. Weather forecasting data, such as temperature, dew point, humidity, pressure and wind speed, are widely available from local meteorological organisations. However, solar irradiance forecasting data is often unavailable. In this paper, multiple regression (MR) and artificial neural network (ANN) models are used to forecast solar irradiance using weather forecasting data. The correlation of each weather parameter with solar irradiance is investigated. It is evident that the ANN model is able to improve the accuracy in terms of root mean square error (RMSE) by 18.42% of its as compared to the MR model.

Keywords: *Solar energy; solar irradiance; forecasting; multiple regression (MR); artificial neural network (ANN).*

1. INTRODUCTION

In the terms sustainability, there is a consideration for solar energy in the fields of environment, economic and social. Traditional power generation is based on fossil fuel, such as coal, petroleum, and natural gas, which is also known as non-renewable energy resource (Kumar, 2020). The depletion of fossil fuel can cause serious problems, in particular an energy crisis (Manieniyan *et. al.*, 2009). This is the most important reason to expand renewable energy, such as solar energy. The decreasing cost of solar energy deployment has made it a better choice for clean energy generation. To this end, the installed capacity of solar photovoltaic (PV) globally continues to rise (Fraas, 2014). Furthermore, solar energy is suitable to be expanded in Malaysia as compared to other renewable energy as it is located in the equatorial region, which has hot climate throughout the year. In Malaysia, the monthly average daily solar radiation is 4,000-5,000 W/m², with the average daily sunshine duration ranging from 4 to 8 h (Aziz *et al*., 2016).

The output of solar energy is related to solar irradiance, which lacks stability due to weather variation. Solar irradiance forecasting techniques can help to stabilise the production of electricity based on solar energy and sustain its integration with power generation based on fossil fuel (Akhter *et al*., 2019). There are different kinds of forecasting techniques. For instance, regression is a statistical approach, while artificial neural network (ANN) is a machine learning approach. Abuella & Chowdhury (2015), Kumar *et al*. (2016), Jeon & Kim (2020), Anthony & Ho (2021) and Khan *et al*. (2022) used ANN for estimation of solar radiation. On the other hand, Mekparyup *et al*. (2013), Nalina *et al*. (2014) and Massidda & Marrocu (2017) used of multilinear and multivariate regression for prediction of solar irradiance.

There is a lack of solar irradiance models for Malaysia despite its suitable geography with high monthly average daily solar radiation and average daily sunshine duration. This paper uses multiple regression (MR) and ANN to forecast solar irradiance using weather forecasting data, including temperature, humidity, wind speed and pressure.

2. METHODOLOGY

2.1 Multiple Regression (MR)

Regression is a process of modelling between a dependent variable, and one or more independent variables (Ul-Saufie *et al.,* 2011; Ostertagová, 2012). Regression with more than one independent variable is known as MR. Linear regression is a linear form, while quadratic regression is a non-linear form (Akinwande *et al*., 2015). The effect of multicollinearity makes the coefficients of regression insignificant when there are many similar independent variables. In order to avoid multicollinearity, variance inflation factor (VIF) can be used to detect the correlation of independent variables (Daoud, 2017). The range of VIF is as shown in Table 1. For instance, independent variables with VIF of above 5 should be removed.

Table 1: The range of VIF (Daoud, 2017).

VIF	Correlation between independent variables
1	Not correlated
Between 1 and 5	Moderately correlated
Greater than 5	Highly correlated

The goal of regression is to find the best fitted line (can be linear or quadratic) using the method of least-squares fit with estimated coefficients. Sinha (2013) suggested the general equations for multiple linear and quadratic regressions, as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (1)$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon \quad (2)$$

where:

β_0 - intercept
$\beta_1$and β_2 - linear coefficients
$\beta_{11}$and β_{22} - quadratic coefficients
β_{12} - interaction coefficient
X_1 and X_2 - independent variables
ε - random error that follows normal distribution with mean 0.

2.2 Artificial Neural Network (ANN)

ANN is a machine learning algorithm that allows neurons to learn like a human brain. A neural network consists of input, hidden and output layers. For each layer, there is a neuron (also known as node) that is connected in between multi-layer networks (Massidda & Marrocu, 2017). Mathematically, the output of a neuron can be calculated using Equation 3. The structure of a neuron is shown in Figure 1. In this study, weather parameters, including temperature, humidity, wind speed and pressure, are used as the input, *x* to the ANN, while solar irradiance is used at the output, *y*.

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

where:

x	-	input
w	-	weight
b	-	bias
y	-	output.

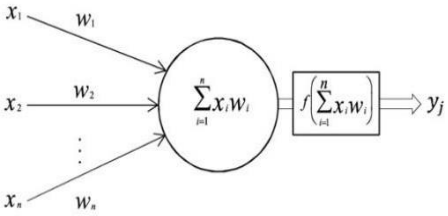

Figure 1: The structure of a neuron (Khatib *et al.*, 2012).

Feed forward neural network (FFNN) is a type of neural network with the application of forecasting. FFNN is based on the backpropagation learning algorithm, which is a type of` supervised learning. The function of backpropagation is to update weights to minimise error. By adjusting the weights, minimum error between the actual and predicted outputs of the ANN can be achieved (Grossi & Buscema, 2007).

3. RESULTS AND DISCUSSION

The solar irradiance and weather data are obtained from a weather station located in Melaka, Malaysia (N 2.314100, E 102.318353). In this study, three months of data (March to May 2020) is used for training the models to forecast irradiance for June 2020.

3.1 Data Analysis for the Dependent and Independent Variables

Solar irradiance is identified as the dependent variable, while temperature, humidity, wind speed and pressure are identified as the independent variables. The data analysis is separated into two parts. The first part is a correlation analysis among the independent variables. The second part is the correlation analysis between the dependant and independent variables.

Figure 2 shows that the data analysis between the independent variables. The relationship between temperature and humidity is the most significant, which is inversely related to each other. Furthermore, it is difficult to identify any relationship for the other independent variables.

Table 2 shows the correlation coefficient (R) and VIF among the independent variables. The relationship between temperature and humidity has the highest VIF of 4.1519, which indicates moderate correlation. For the rest of the independent variables, the values of VIF are close to 1, which indicates no correlation. Based on this, all the independent variables are included for modelling as they have VIF of less than 5.

Figure 3 shows that the data analysis between the dependent and independent variables. The left column shows the overall data view, while the right column shows the zoomed in view from the overall data for better view of the data analysis. It is found that there is high solar irradiance when the temperature is high. Besides that, humidity is the inverse of temperature and thus solar irradiance decreases as humidity increases. Wind speed is not steadily related to solar irradiance. For the pressure, it can be observed that the solar irradiance increases at the moment when pressure drops from its peak.

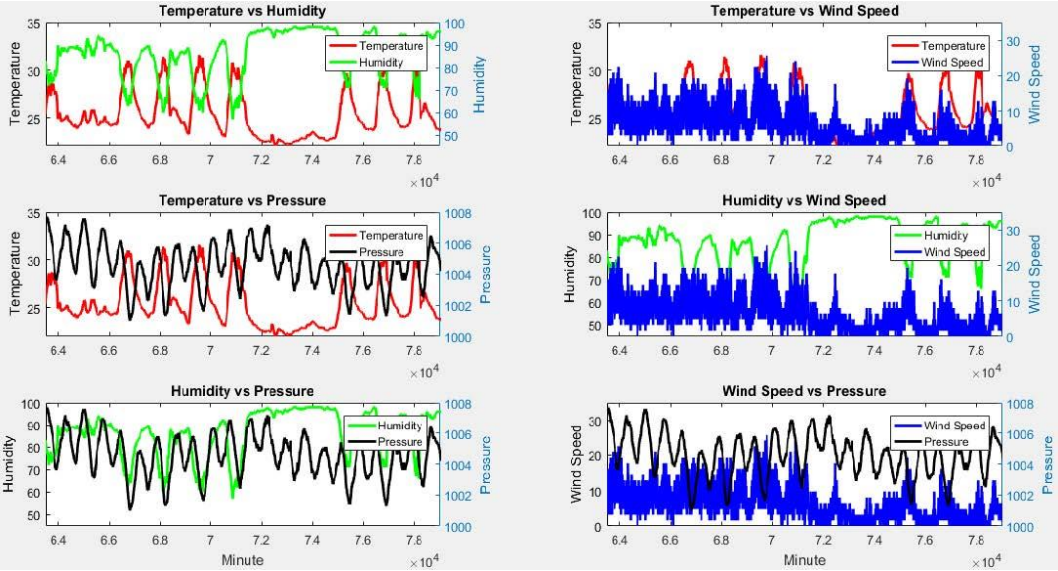

Figure 2: The data analysis between the independent variables (part of data).

Table 2: Values of *R* and VIF between the independent variables.

Parameters	*R*	VIF
Temperature & Humidity	-0.8713	4.1519
Temperature & Wind Speed	0.2452	1.0639
Temperature & Pressure	-0.2544	1.0692
Humidity & Wind Speed	-0.3773	1.1659
Humidity & Pressure	0.1994	1.0414
Wind Speed & Pressure	0.1071	1.0116

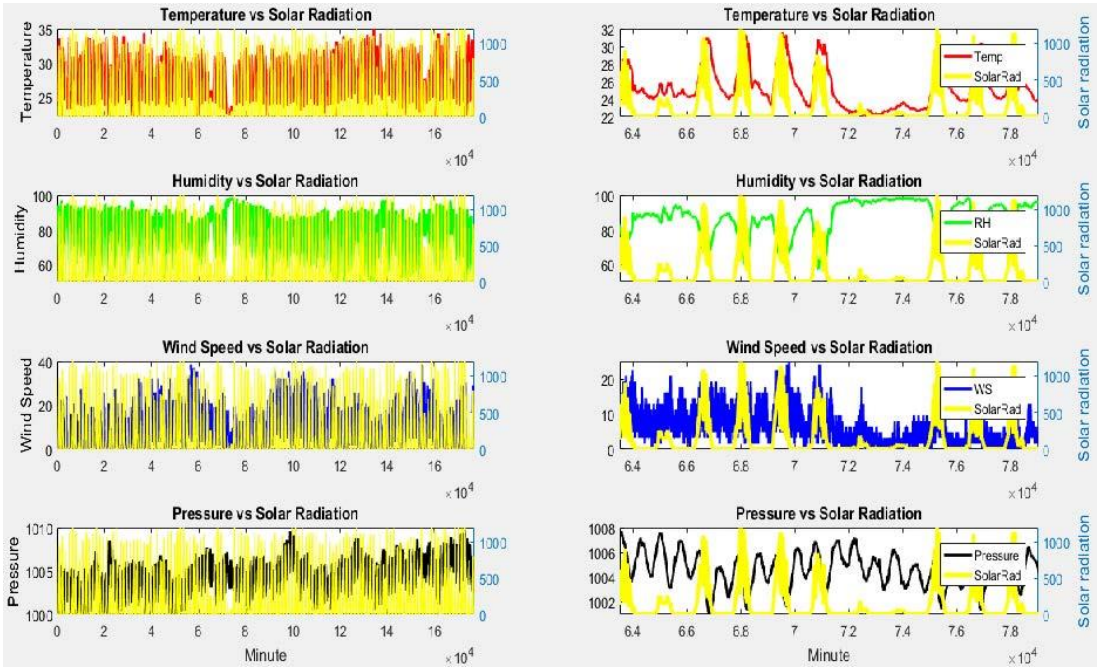

Figure 3: The data analysis between the independent variables and dependent variable.

From Table 3, the relationship between temperature and solar irradiance indicates the highest *R* of 0.6884. This is because strong solar irradiance causes higher temperature. In addition, the lowest *R* of $8.01e^{-4}$ and the highest root mean square error (RMSE) of 865.9054 is found for the relationship between pressure and solar irradiance.

Table 3: Values of *R* and RMSE for the dependent and independent variables.

Parameters	*R*	RMSE
Temperature & Solar Irradiance	0.6884	327.2544
Humidity & Solar Irradiance	-0.6878	313.7555
Wind Speed & Solar Irradiance	0.3552	336.6058
Pressure & Solar Irradiance	$8.01e^{-4}$	865.0954

3.2 Result Validation for the MR Model

The Statistics and Machine Learning Toolbox in MATLAB is used for the results validation for the MR model, as shown in Table 4. With the increase of parameters used, *R* is increased while RMSE is decreased. There is also improvement from 1st order (linear) to 2nd order (quadratic) regression. As the result, *R* is increased from 0.7642 to 0.8815 and RMSE is decreased from 185 to 135 for the model including all the parameters.

Table 4: Validation for MR model with different parameters.

Parameters	First order (Linear)		Second order (Quadratic)	
	R	RMSE	*R*	RMSE
Temp	0.6885	208	0.7148	200
Temp, RH	0.7113	201	0.7197	199
Temp, RH, WS	0.7246	197	0.7589	187
Temp, RH, WS, Pressure	0.7403	193	0.8000	172
Temp, RH, WS, Pressure, Time	0.7642	185	0.8815	135

*Note: Temp - Temperature; RH - Humidity, WS - Wind Speed

Based on the estimated coefficients, the equation of for the MR model is as follows:

$$\begin{aligned} y = {} & 4005X_1 + 1185X_2 - 1005.1X_3 - 186.6X_4 - 4586X_5 - 0.43917X_1X_2 \\ & + 2.8699X_1X_3 - 3.9952X_1X_4 - 104.36X_1X_5 \\ & + 0.0072604X_2X_3 - 1.1856X_2X_4 + 23.461X_2X_5 \\ & + 0.93005X_3X_4 - 1.2566X_3X_5 + 46.991X_4X_5 \\ & + 2.7276{X_1}^2 + 0.027146{X_2}^2 - 0.12369{X_3}^2 \\ & + 0.18551{X_4}^2 - 8.55.27{X_5}^2 \end{aligned} \quad (4)$$

where:

X_1 - temperature
X_2 - humidity
X_3 - wind speed
X_4 - pressure
X_5 - time

3.3 Result Validation for the ANN Model

The Deep Learning Toolbox in MATLAB is used for the results validation for the ANN model, as shown in Table 5. The ANN is set to only one hidden layer that consists of five neurons. It is found that the model gets better as the number parameters used is increased. With only the parameter of temperature used, the ANN only achieves *R* of 0.7278 and RMSE of 195.3103. The ANN that includes all the weather parameters as inputs has the highest *R* of 0.9067 and lowest RMSE of 121.0106.

Table 5: Validation for the ANN model with different parameters.

Parameters	*R*	RMSE
Temp	0.7278	195.3103
Temp, RH	0.7298	194.9912
Temp, RH, WS	0.7709	181.4396
Temp, RH, WS, Pressure	0.8308	159.8094
Temp, RH, WS, Pressure, Time	0.9067	121.0106

*With only one hidden layer that consists of five neurons

**Note: Temp - Temperature; RH - Humidity, WS - Wind Speed

Table 6 shows the validation of the ANN model for all the weather parameters with different number of hidden layers. Each hidden layer has the same number of neurons, which is five neurons. The ANN with three hidden layers achieved the highest *R* of 0.9173 and lowest RMSE of 114.1820. This indicates that with more hidden layers, the ANN becomes deeper and provides better results.

Table 6: Model validation with different number of hidden layers.

Parameters	No. of Hidden Layers	No. of Neuron	*R*	RMSE
Temp, RH, WS, Pressure, Time	1	5 (for each hidden layer)	0.9067	121.0106
	2		0.9153	116.3766
	3		0.9173	114.1820

3.4 Discussion

From the results obtained, it is found that the ANN model performed better as compared to the MR model. This is as the ANN model is based on the backpropagation algorithm, while the MR model is based on the best fit line that is regressed by all the parameters.

The comparison of *R* and RMSE values for the MR (second order quadratic) and ANN (three layers with five neurons for each layer) models is shown in Table 7.

Table 7: Comparison of *R* and RMSE values for the MR and ANN models.

Model	*R*	RMSE
MR (second order quadratic)	0.8815	135
ANN (three layers with five neurons for each layer)	0.9173	114.1820

Figure 5 shows that final structure of the ANN model consisting of three hidden layers. With each hidden layer having five neurons, the model has a total of 15 neurons. As it has the highest *R* and lowest RMSE, it is the best model to be used to forecast solar irradiance.

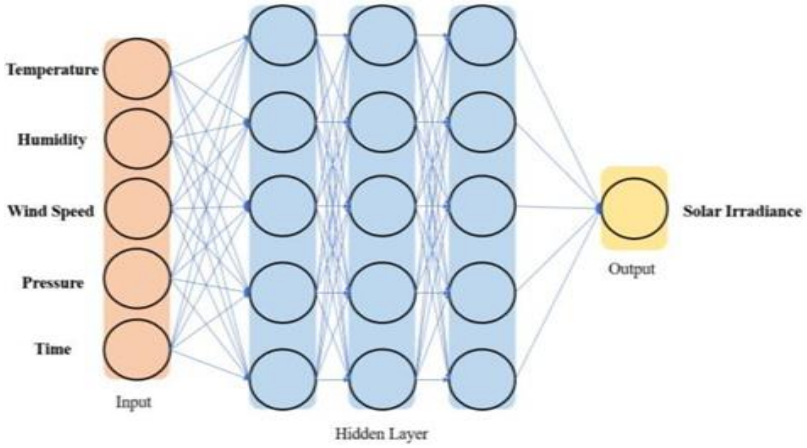

Figure 5: The final ANN structure.

4. CONCLUSION

In this study, MR and ANN methods were used to forecast solar irradiance using weather parameters, including temperature, humidity, wind speed and pressure. It is evident that the ANN model is able to improve the accuracy in terms of by 18.42% as compared to the MR model. The ANN structure with three layers and five neurons for each layer provided highest accuracy with *R* of 0.9173 and RMSE of 114.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM) and the Ministry of Higher Education Malaysia for funding this research under the Fundamentals Grant Scheme (FRGS/2018/FKEKK-CETRI/F00358).

REFERENCES

Abdullah, W.S.W., Osman, M., Kadir, M.Z.A.A. & Verayiah, R. (2019). The potential and status of renewable energy development in Malaysia. *Energies*, **12:** 2437.

Abuella, M. & Chowdhury, B. (2015). Solar power forecasting using artificial neural networks. *Proc. 47th Annual North Am. Power Symp.*, Charlotte, North Carolina, US., 4 - 6 October 2015.

Akhter, M.N., Mekhilef, S., Mokhlis, H. & Mohamed Shah, N. (2019). Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. *IET Renew. Power Gener.* **13**: 1009-1023.

Akinwande, M. O., Dikko, H.G. & Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open J. Stat.* **5**: 754–767.

Anthony, A. & Ho, Y.H. (2021). Solar Irradiance forecasting using Global Positioning System (GPS) derived total electron content (TEC). *Def. S T Tech. Bull.* **14**: 91 – 100.

Aziz, P.D.A., Wahid, S.S.A., Arief, Y.Z. & Aziz, N.A. (2016). Evaluation of solar energy potential in Malaysia. *Tr. Bioinformatics*. **9**: 35-43.

Daoud, J. I. (2017). Multicollinearity and regression analysis. *J. Phys.: Conf. Ser*. **949**: 012009.

Fraas, L.M. (2014). *Low-Cost Solar Electric Power*, Springer Nature, New York, US.

Grossi, E. & Buscema, M. (2007). Introduction to artificial neural networks. *Eur. J. Gastroenterol. Hepatol*. **19**: 1046–1054.

Jeon, B.K. & Kim, E.J. (2020). Next-day prediction of hourly solar irradiance using local weather forecasts and lstm trained with non-local data. *Energies* **13(**20): 5258.

Khan, W., Walker, S. & Zeiler, W. (2022). Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy*, **240**: 122812.

Kumar, M. (2020). Social, economic, and environmental impacts of renewable energy resources. *In* Okedu, K.E., Tahour, A. & Aissaou, A.G. (Eds), *Wind Solar Hybrid Renewable Energy System*. IntechOpen, London, UK.

Kumar, R. Pathania, S., Gupta, A., Sekhar, R. & Aggarrwal, R. K. (2016). Artificial neural network model for precise estimation of global solar radiation. *Int. J. Curr. Res.* **8**: 31119 – 31124.

Manieniyan, V., Thambidurai, M. & Selvakumar, R. (2009). Study on energy crisis and the future of fossil fuels. *Proc. SHEE 2009*, Annamalai University, Chidambaram, India, 11–12 December 2009.

Massidda, L. & Marrocu, M. (2017). Use of multilinear adaptive regression splines and numerical weather prediction to forecast the power output of a PV plant in Borkum, Germany. *Sol. Energy*. **146**: 141–149.

Mekparyup, J., Saithanu, K. & Dujjanutat, J. (2013). Multiple linear regression equation for estimation of daily averages solar radiation in Chonburi, Thailand. *Appl. Math. Sci*. **7**: 3629–3639.

Nalina, U., Prema, V., Smitha, K. & Rao K. U. (2014). Multivariate regression for prediction of solar irradiance. *Proc. Int. Conf. Data Sci. Eng. 2014 (ICDSE 2014)*, Kochi, India, 26-28 August 2014.

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Eng.*, **48**: 500–506.

Sinha, P. (2013). Multivariate polynomial regression in data mining: methodology, problems and solutions. *Int. J. Sci. Eng. Res*. **4**: 962–965.

Ul-Saufie, A.Z., Yahya, A.S., Ramli, N.A. & Hamid, H.A. (2011). Comparison between multiple linear regression and feed forward back propagation neural network models for predicting pm 10 concentration level based on gaseous and meteorological parameters. *Int. J. Appl. Sci. Tech*, **1**: 42–49.