Pakistan Journal of Engineering and Technology
ISSN (Print): 2664-2042; ISSN (Online): 2664-2050
Received: 2024-04-09
Accepted: 2024-06-25
https://doi.org/10.51846/vol7iss2pp66-71

PAKJET
Pakistan Journal of Engineering & Technology

Article

Solar Irradiance Prediction for Zaria Town Using Different Machine Learning Models

Ibrahim Abdulwahab *, Sulaiman Haruna Sulaiman, Umar Musa, Ibrahim Abdullahi Shehu, Abdullahi Kakumi Musa, Ismaila Mahmud, Mohammed Musa, Abdullahi Abubakar, and Abdulrahman Olaniyan

Department of Electrical Engineering, Ahmadu Bello University Zaria, Zaria, Nigeria
* Correspondence: Ibrahim Abdulwahab (iabdulwahabb@gmail.com)

Abstract: The research is set to predict solar irradiation using various machine learning algorithms. This is done in order to construct and develop a high-efficiency prediction model that uses actual meteorological data to predict daily solar irradiance for the town of Zaria, Nigeria. To assist utilities working in various solar energy generation and monitoring stations in making effective solar energy generation management system decisions. Four machine learning models (artificial neural network (ANN), decision tree (DT), random forest (RF), and gradient boost tree (GBT).) were used to predict and compare actual and anticipated solar radiation values. The results reveal that meteorological characteristics (min-humidity, max-temperature, day, month, and wind direction) are critical in machine learning model training. The solar radiation prediction skills of multi-layer perceptron and decision tree models were low. In the prediction of daily solar irradiation, the ensemble learning models of random forest and gradient boost tree outperformed the other models. The random forest model is shown to be the most accurate in predicting solar irradiation.

Keywords: Machine Learning, Renewable Energy, Solar Irradiance, Weather Prediction.

1. Introduction

The demand for energy generation from renewable energy sources is at an all-time high, due to hot house effect, depletion and continuous change in the price of fossil fuel and environmental pollution [1, 2]. Wind, tidal streams ocean, biomass and solar are the mostly used renewable sources for generation of electrical energy [3, 4]. Being in abundance naturally makes solar energy, one of the most used forms of renewable energy source (RES) [5]. However, one of the major drawbacks of this form of energy is the intermittent nature of output power, which is mainly due to weather conditions [6, 7].

In the study and use of solar energy, solar radiation data is crucial. For modeling and designing of solar-based applications, it provides the necessary data about the solar energy impact on the earth's surface. Because of a lack of measurement tools or meteorological stations, solar radiation data are frequently unavailable in many developing nations. It is also true that many nations were unable to afford the pricey measurement equipment and methods required to determine sun irradiation [8]. The development of models that accurately quantify solar irradiance utilizing a number of meteorological characteristics, such as altitude, latitude, and longitude, as well as climatological parameters, such as humidity, daylight duration, temperature, pressure, etc. Variations in temperature and irradiation mostly determine the quality of produced solar energy. Also, the level of solar irradiance available at a particular time, determines the level of power output from the solar system. A large reduction in economic profit can be observed in large-scale solar farms due to power imbalance in photovoltaic system. Therefore, a correct forecast of solar irradiance is essential in removing the uncertainty and ensuring the integration of the photovoltaic systems in a smart grid [7].

Solar prediction models can be classified into three: The empirical models (which are simple but suffer from low accuracy), the radiative transfer models with inherent (not often used because of the complex parameters needed for the model to function effectively), and the artificial neural network models [8].

Several researches have been done to improve the accuracy of prediction techniques for solar irradiance due to the continuous increase in the installation of solar systems. In [9], they developed an artificial neural network technique based on the feed-forward multilayer perception model for solar radiation prediction in Malaysia. The basis for the ANN model is the feed-forward multilayer perception model including four inputs and one output. Inputs are latitude, longitude, day number, and sunshine ratio; the output is the clearliness index. Data from 28 weather stations were used in this work; of them, 23 were used for network training and 5 for testing the network. The results obtained were presented using MAPE performance criterion. In [10], a forecasting technique for short-range solar prediction using ANN was developed. In order to offer 15-min average clearness index projections for lead durations of 15 min, 60 min, 120 min, and 180 min, the short-range solar irradiance forecasting method in this work is cloud regime dependent. In order to reduce energy prices and offer good power quality in electrical power networks with

solar photovoltaic generations, the prediction of solar irradiance is crucial to the task. While considering the dependency between related hours of the same day, long short-term memory (LSTM) networks were trained to develop the proposed prediction model. The experimental results show that the suggested approach performed satisfactorily on a dataset gathered in Santiago, Cape Verde when compared to another competitive algorithm for single output prediction. Reference [7] proposed a scheme that used ensemble learning models for short-time prediction of solar irradiance data. Due to its beneficial properties, ensemble learning models are used more frequently than traditional single learners since they combine a number of weak regressors to produce predictions with higher accuracy [8]. The results obtained from the scheme showed that reliable and consistent prediction can be obtained from ensemble models when applied to data from different locations.

2. Materials and Methods

This section has three subsections. Firstly, the information regarding the collection of the data and its pre-processing methods is given. A highlight of the machine learning methods is provided. The performance metrics used to validate the models are briefly given.

2.1. Data Collection

The study area is Zaria, a city in Kaduna State, Nigeria. The study area is described as Köppen climate classification which is also known as a savanna climate (tropical wet and dry seasons) that has substantial variations in rainfall and a consistent high temperature throughout the year as a distinct characteristic. In Zaria, average high temperatures hover well above 28.3°C even in the coolest months. The warmest month typically sees averages soaring to around 38.6°C. Contrarily, during the coldest months, the average low temperatures usually do not fall below 14.8°C. This confirms a generally high-temperature profile in Zaria throughout the year.

Relative humidity fluctuates significantly across the year. The dry season, which coincides with lower humidity levels, witnesses an average relative humidity as low as 16%, while the wet season sees a spike to averages as high as 82%. Wind speeds average around 10 km/h throughout the year, with few variations from month to month.

The datasets were collected from the Samaru Automated Data Station located at the Institute of Agricultural Research (IAR), faculty of Agricultural Science, Ahmadu Bello University Zaria. Normally, climatological datasets are made up of the average daily weather for the thirteen months deemed usual for the location in question. The datasets contain records of solar irradiance, lowest and highest temperatures, lowest and highest humidity, dew point, rainfall, day, month, year, wind speed, wind gravity, and wind direction. The datasets of 13 months which span within the year 2020-2021, were collected. Given the duration of the data collection period and the inherent inaccuracies in instrument-based observations, data quality control was critical. Daily data records totaling 420 rows and 13 columns were obtained.

2.2. Machine Learning Methods

Machine-learning (Mahadevan) models are types of artificial intelligence. To have an acceptable accuracy, there is a need to combine machine-learning models with other algorithms to achieve the desired goal set. The following are the main steps in the building of an ML model: data preparation, feature selection, data pre-processing, and model development. In this study, four different ML algorithms were used. These are artificial neural networks (ANN), decision trees (DT), random forests (RF), and gradient boost trees (GBT).

2.2.1. Artificial Neural Network (ANN)

ANN is a machine learning technique influenced by biological brain systems that can process non-linear relationships, data classification, pattern prediction, optimization, clustering, and simulation. Usually, as shown in Fig. 1, an ANN model consists of three layers: the input layer, the hidden layer, and the output layer.

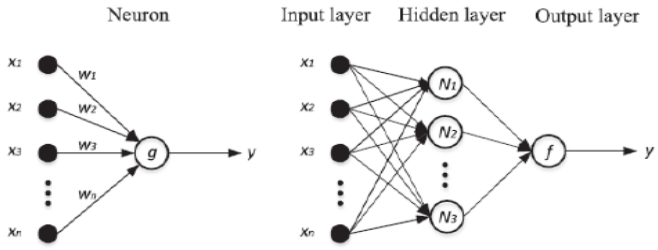

Figure 1. A typical artificial neural network model.

$$y = g\left(\sum_{i=1}^{n} w_i x_i\right) \quad (1)$$

where
x_i = input
w_i = input's weight
g = transfer function

$$Y = f(W, X) \quad (2)$$

where
X = input vector
W = weight vector
$f(.)$ = function relating the input and output vector.

2.2.2. Decision Tree

Due to its recursive divide-and-conquer character, decision trees are a rule-based learning method with a straightforward core principle. Prediction from inputs X1, Xp must result in a response or class Y. A binary tree is grown to do this. In forming a branch, the entire set of features is taken into account. Using each of the independent variables, a model is created. Mean squared error is utilized for each of the various variables to determine the best split. A test to one of the inputs, let's say Xi, is applied at each node in the tree. The direction

(left or right) of the branch of the tree is chosen depending on the results of the test. A forecast is made after the tree reaches a leaf node. To obtain this prediction, at the leaf stage, the training data points could have been used entirely or averaged.

2.2.3. Random or Bagged Forests

Random Forest is an ensemble ML algorithm that is used in both regression and classification problems, to improve the accuracy of its base ML algorithm. Additionally, the method lessens variation and aids in avoiding over-fitting. It can be utilized with any kind of learning method, even though it is typically applied to decision tree approaches. A specific instance of the model averaging approach is bagging. Bagging is very helpful when used with tree models because these models are very sensitive to changes in the training data. Bagging is frequently used in conjunction with another concept when applied to tree models: subspace sampling is the process of building each tree from a distinct random subset of the features. This promotes variation in the ensemble even more and has the added benefit of speeding up each tree's training. Random forests refer to the ensemble technique that was produced.

2.2.4. Gradient Boost Tree

This ensemble model constructs the model stage-by-stage by optimizing a loss function, using decision trees as weak learners. Boosting was developed as a method of creating a powerful "committee" by merging multiple weak classifiers. Bad classification is given more weight over time through an iterative process. Performance in categorization is dramatically improved by a straightforward method. Using a so-called weak learner, a learner with a significant bias relative to variance, the residuals from the recursive linear forecasts are applied to a boosting auto-regression technique at each horizon.

3. Performance Criteria

In order to forecast how well a model will generalize to out-of-sample data, we used two evaluation measures in this work under the train/test split model evaluation approach. The experiment used a 75% training data split combined with 25% test data.

Two statistical measures—coefficient of determination (R^2) and root mean square error—that are frequently used in the literature to compare the efficacy of ML models were employed as evaluation metrics in the current study.

The coefficient of determination, or R^2, is a metric that sheds light on how well a model fits the data. It accepts numbers in the range 0 to 1. The model performs better the larger this indicator is. Therefore, the model's capacity to estimate values is improved by the R^2's proximity to unity. This indicator can be quantitatively stated as:

$$R^2 = 1 - \frac{\sum(y_a - y_p)^2}{\sum(y_a - \bar{y}_p)} \quad (3)$$

where:

n = number of observations
y_a = actual value
y_p= predicted value
$\bar{y}_p$= mean of predicted values

RMSE compares each actual alteration that occurs between the estimated and measured values, which in turn gives information regarding the performance index. It resembles the mean absolute deviation error to some extent. The mathematical formula for RMSE is:

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (y_{i\,e} - y_{i\,m})^2} \quad (4)$$

where:
n = number of observations
$y_{i\,e}$= the i[th] estimated value
$y_{i\,m}$= the i[th] measured value

Three phases make up the experiment: preparation of data, model creation and prediction. As illustrated in Fig. 2, three phases comprised the data preprocessing: data quality, data partition control and variable selection.

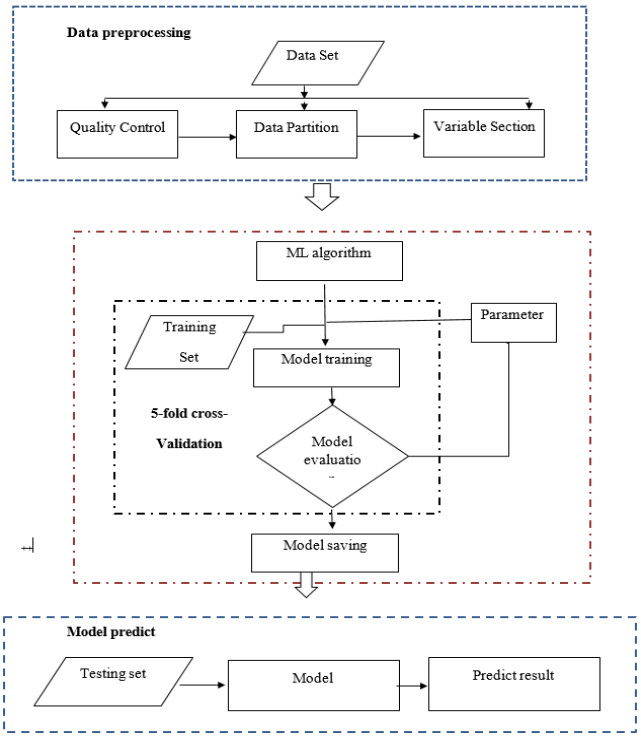

Figure 2. Flow chart for solar irradiance prediction using machine learning models.

The main models-building techniques were variable selection, algorithm choice, model development, and model preservation. In the parameter selection step, the 5-fold cross-validation method was utilized. These experimental steps were used:

1) data collection;
2) data preprocessing;
3) Selection of machine learning algorithm to predict solar radiation;
4) If the best predictive ability is achieved, save the model;
5) return to step (2) and select another machine learning model;
6) input the preprocessing dataset;
7) save and analyze predicted results.

3.1. Variable Selection

This selection process is vital when constructing ML models. Genetic algorithm, the Tabu search, particle swarm optimization, and the random forest algorithm are some of the mostly used variable selection algorithms. Data variable selection is accomplished using the random forest algorithm. The random forest model was built and trained using normalized daily data, and the model's significance was computed using them as well. The experiment proceeded as follows:

1) The dataset is partition into training set to testing set ratio;
2) The training set is used in training the model;
3) then, the testing set is used in computing the performance index of the model

3.2. Model Building

Built using Python 3.6 with external libraries including NumPy, Pandas, and the scikit-learn ML package (Sklearn), the model was Four machine learning techniques were applied in construction of the models. The starting parameter values of every method were set depending on their characteristics. For instance, empirical computations and neural network design ideas helped to define the hidden layer and neuron counts in a neural network model. The pertinent selection ranges of the adjustment parameters and other parameters were then established using the methods of parameter adjustment applied for different machine learning algorithms. Each of the four machine-learning models had parameters selected using Sklearn's GridSearchCV approach; the best model was subsequently stored.

4. Results and Discussion

This section focuses on the results obtained during the training of the different machine learning models. When available, daily global solar radiation is the most common available measure with a sufficiently lengthy period of data. However, daily data are required for suitable solar system sizing and evaluation. The ensemble model was trained, as explained in the methodology section, to generate daily solar radiation. The daily min-temperature, max- temperature, min. humidity, max-humidity, wind direction, wind speed, wind gravity, dew, day, month, and year are the main inputs.

Various combinations of these parameters have been suggested as matrix input (stimuli) for the four (4) machine learning models considered in this research, with solar irradiance as the target output (label). For each combination, different configurations of the 4 machine-learning models were tested by changing the model parameters. Training was done on three different machine learning models so that the best configuration was chosen. Table 1 shows the best results for training through the RMSE and coefficient of correlation R2. The training was done over the data of 13 months.

The result shows that the random forest model outperforms the competition. This suggests that the proposed model, based on the RMSE and R2 value, has a greater ability to forecast future data. Additionally, it is more effective in forecasting daily sun radiation.

4.1. Performance Comparison for the Models Developed

From Table 1 it clearly shows the RMSE and R^2 values for the 4 machine learning models. The random forest model gave the best prediction result with R^2 value of 0.771 and gradient boost tree also displayed a good prediction performance followed by the decision tree and multi-layer perceptron. The random forest model has the lowest RMSE value of 529.064, reflecting the highest precision of all the models.

4.2. Discussion

One can predict solar irradiance at any location in the world if past meteorological data is available to train the model. Four machine learning models were predicted using daily data; Table 1 shows that random forest models had greater prediction ability than the other models.

Table 1. Performance indices of the models.

Model	RMSE	R^2
Decision Tree	1045.117	0.569
Random Forest	529.064	0.771
Gradient Boost Tree	565.747	0.748
Multi-layer Perceptron	1001.751	0.521

Figs. 3, 5, 6 and 7 show the plotted predicted and actual data of a daily solar irradiance. It is clearly shown from these plots that there is good agreement between predicted and actual data especially random forest and gradient boost tree performed better with correlation coefficient of 0.771 and 0.746, respectively, while decision tree and multi-layer perceptron performed poorly. The variations that occurred between predicted and actual data are due to high fluctuations in weather conditions in Nigeria.

Fig. 4 illustrates in the final random forest model the significance of the input variables as predictors. Min- humidity shown as the most critical feature, followed by max-temperature, month, day, max-humidity, min-temperature, rainfall, wind direction, wind speed, dew, wind gravity and year.

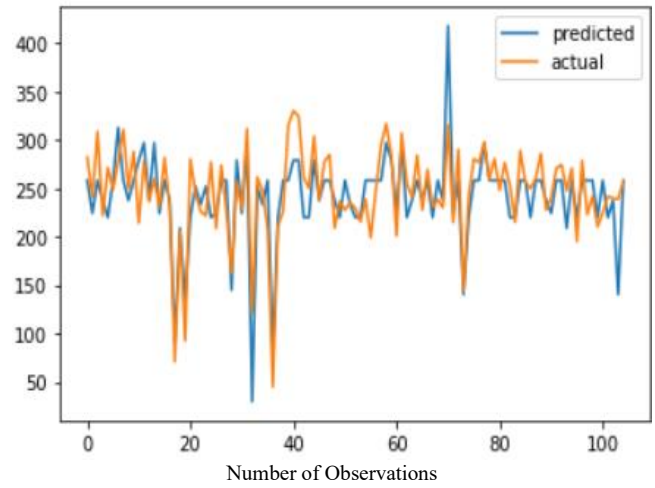

Figure 3. Predicted vs actual solar irradiance data for decision tree.

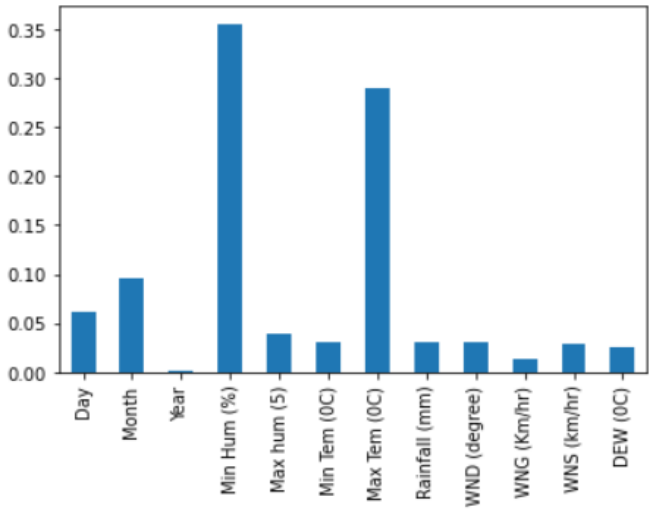

Figure 4. Feature extraction of variable selection.

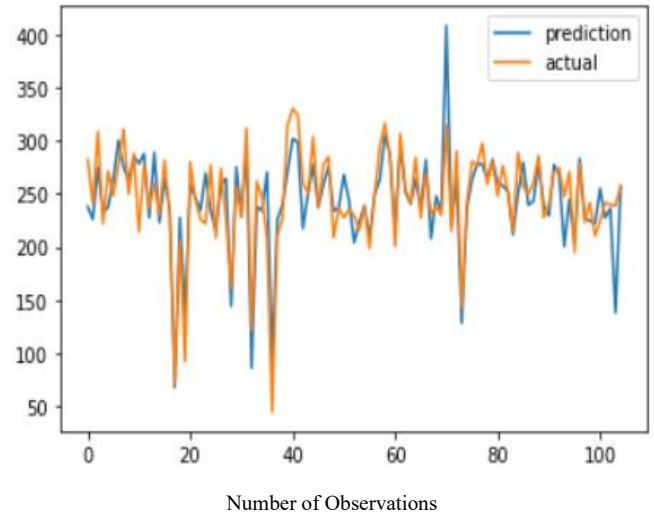

Figure 5. Solar irradiance data for random forest.

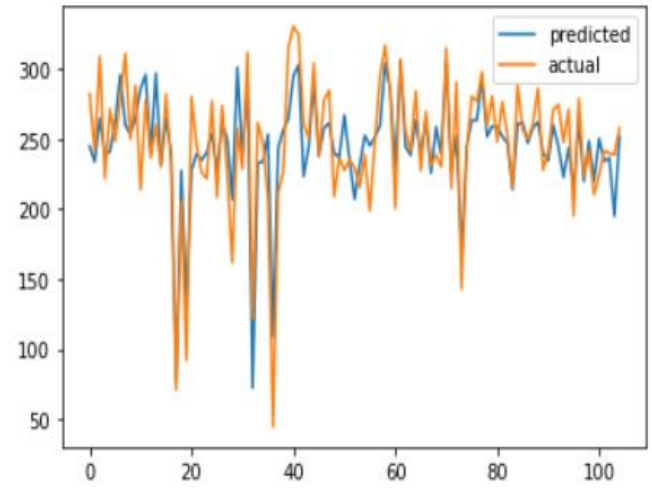

Figure 6. Solar irradiance data for gradient boost tree.

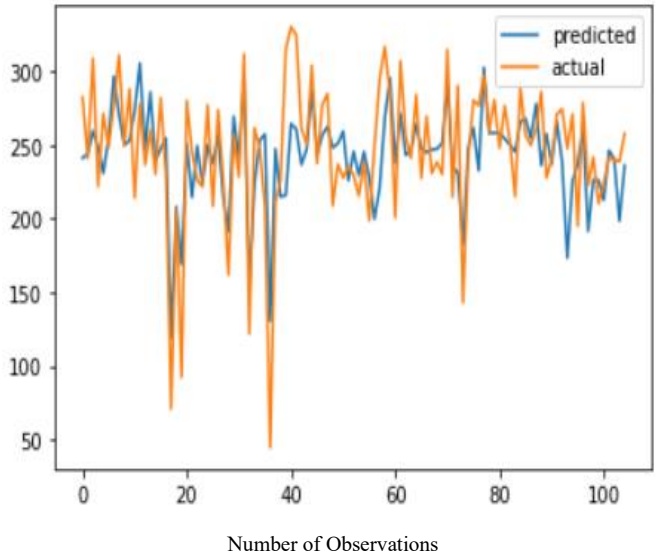

Figure 7. Solar irradiance data for multi-layer perceptron.

5. Conclusion

Data preprocessing and variable selection for Zaria were accomplished using meteorological elements and solar radiation data from 2020 to 2021. Then, using Python (Sklearn), four machine-learning models were created. By comparing and evaluating the predictive ability of the 4 machine learning models using the RMSE and R^2 indices, the random forest model was selected due to their performance is better than other models. It was concluded that the random forest model performed well over other machine learning models.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] I. Abdulwahab, A. S. Abubakar, A. Olaniyan, B. O. Sadiq, and S. A. Faskari, "Control of Dual Stator Induction Generator Based Wind Energy Conversion System," in *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, 2022, pp. 1-5: IEEE.

[2] M. U. Iliyasu, M. Ozohu, S. S. Yusuf, I. Abdulwahab, A. Ehime, and A. Umar, "Development of an Optimal Coordination Scheme for Dual Relay Setting In Distribution Network Using Smell Agent Optimization Algorithm,"*Covenant Journal Of Engineering Technology,* 2022.

[3] Hafis, A., Adamu, A. S., Jibril, Y., & Abdulwahab, I. (2023). An Optimal Sizing of Small Hydro/PV/Diesel Generator Hybrid System for Sustainable Power Generation. *Journal of Engineering Science Technology Review, 16*(6)

[4] A. Babaita, A. Mati, Y. Jibril, A. Kunya, and I. Abdulwahab, "Development of a load frequency control scheme for an autonomous hybrid microgrid,"*Zaria Journal of Electrical Engineering Technology,* vol. 11, no. 1, 2022.

[5] John, R., Mohammed, S. S. & Zachariah, R. Variable step size Perturb and observe MPPT algorithm for standalone solar photovoltaic system. 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017. IEEE, 1-6.

[6] I. Abdulwahab, S. A. Faskari, T. A. Belgore, and T. A. Babaita, "An Improved Hybrid Micro-Grid Load Frequency Control Scheme for an Autonomous System," FUOYE Journal of Engineering and Technology, vol. 6, no. 4, Dec. 2021, doi: https://doi.org/10.46792/fuoyejet.v6i4.698

[7] J. Lee, W. Wang, F. Harrou, and Y. Sun, "Reliable solar irradiance prediction using ensemble learning-based models: A comparative study," *Energy Conversion and Management*, vol. 208, p. 112582, Mar. 2020, doi: https://doi.org/10.1016/j.enconman.2020.112582

[8] L. Huang, J. Kang, M. Wan, L. Fang, C. Zhang, and Z. Zeng, "Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events," *Frontiers in Earth Science*, vol. 9, Apr. 2021, doi: https://doi.org/10.3389/feart.2021.596860.

[9] T. Khatib, A. Mohamed, K. Sopian, and M. Mahmoud, "Solar Energy Prediction for Malaysia Using Artificial Neural Networks," *International Journal of Photoenergy*, vol. 2012, pp. 1–16, 2012, doi: https://doi.org/10.1155/2012/419504.

[10] T. McCandless, Sue Ellen Haupt, and G. S. Young, "A regime-dependent artificial neural network technique for short-range solar irradiance forecasting," *Renewable Energy*, vol. 89, pp. 351–359, Apr. 2016, doi: https://doi.org/10.1016/j.renene.2015.12.030.