Knowledge-based Engineering and Sciences (2022), **3**, (1), 28–36
doi:10.51526/kbes.2022.3.1.28-36

RESEARCH PAPER

A Novel Machine Learning based Computing Algorithm in Modeling of Soiled Photovoltaic Module

Anas Faskari Shehu,*,1 Tolulope Falope,2 Gerald Ojim,1 SettingsYusuf Badamasi Abdullahi,1 and S. I. Abba3,4

1Department of Electrical and Computer Engineering, Faculty of Engineering, Baze University, Abuja, Nigeria
2Energy and Power, School of Water, Energy and Environment, Cranfield University, MK41 OAL.UK
3Department of Civil Engineering, Faculty of Engineering, Baze University, Abuja, Nigeria
4Interdisciplinary Research Center for Membrane and Water Security, King Fahad University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
*Corresponding author. Email: anas.shehu@bazeuniversity.edu.ng

(Received 30 November 2021; revised 20 March 2022; accepted 26 April 2022; first published online 30 April 2022)

Abstract

Soiling of Photovoltaic (PV) Panels is the accumulation of these dust particles on the panel surfaces. Soiling absorbs, scatters, and reflects a portion of incoming sunlight, lowering the intensity that reaches the solar cell's active part. Predicting expected power under soiling conditions is a difficult problem, and a variety of models have been proposed, each with a different set of inputs. Here, we report on a novel machine learning based computing algorithm in modeling of soiled photovoltaic module. The daily measurements (current, voltage, temperature, and wind speed), as well as a dependent variable (power expected) of PV performance loss due to soiling are available for 90 days, of which 80 days were considered for modeling. Two methods of cross validation were used to ensure that there was no overfitting or underfitting in the training and testing data. The data was split into two groups with the hold-out cross validation approach used. In this study, 75% of the data was used to train the models and 25% was used to test them. The efficacy of the proposed model was evaluated against the multilinear Regression (MLR) model during calibration, and validation periods based on Mean Absolute Error (MAE) and Nash–Sutcliffe efficiency (NSE). According to the results of appraisal, the models developed in this study performed good in estimating the predicted power of a soiled PV module with minimal error in Gaussian Process Regression Matern 5/2 GPR-M (MAE = 0.0784 and 0.0784 and high efficiency in NSE = 0.9745 and 0.8604) in training and testing respectively. Furthermore, the results indicate the better performance and suitability of the model with five input parameters in predicting the expected power.

Keywords: Artificial intelligence; expected power; machine learning; photovoltaic

1. Introduction

Solar energy has been identified as one of the most promising possibilities to replace non-renewable energy sources. This alternative clean energy source addresses some of the major issues that carbon-based fossil energy sources face, such as the depletion of existing reserves and the damage that greenhouse gas emissions bring to key environmental resources. [17]. The photovoltaic (PV) system converts the solar power obtained from the sun to electricity through semiconductor materials like silicon and cadmium telluride. Solar energy is abundant, clean and infinite energy [15]. Renewables presently account for approximately 25% of worldwide electricity generation, with solar photovoltaic

(PV) and wind alone accounting for 40% of the total by 2040. Solar energy production in the world is predicted to reach 7200 TWh by 2040. [8]. Many factors influence photovoltaic power output, including sun position, solar irradiation intensity, wind speed, temperature, tilt angle, array mismatch, distribution losses, inverter efficiency, PV panel soiling, and load demand [17, 15]. As a result, the dynamic response of PV systems for utility grid (UG) performance must be thoroughly assessed, as linking a PV system with a UG may cause instability. The PV performance models' uncertainty is still too large [11]. Much of the previous and ongoing research on this topic has been on analyzing module performance rather than system performance. The use of linear and non-linear models to forecast the performance of PV modules under a variety of natural situations has been used in recent studies on optimizing the efficiency of solar modules [10].

It's worth noting that the literature includes work on various models and techniques for surface soiling modeling, such as ANFIS. For Instance, [8] worked on soiling losses as a barrier for energy security depends on photovoltaic power using traditional techniques. The result shows that an increase in the PV tilt angle, or high wind speed and heavy rain showers, reduces dust deposition. [15] Presented a comprehensive review of the soiling of photovoltaic. The work highlights the effect of soiling on PV cells, thus acknowledging the traditional manual and self-cleaning methods. [2] Presented multiple linear regression analysis (MLRA) and artificial neural networks (ANNs) as the tools for the mathematical analysis and Prediction of Ozone Concentration in Ambient Air. The meteorological parameters (the wind direction, the wind speed, air pressure, air temperature, solar radiation, and RH). The results shown that ANNs provide better estimates of ozone concentration on the monitoring site, whereas the multilinear regression model once again has proven to be less efficient in the accurate prediction of ozone concentration. [7] Developed an artificial neural network (ANN) model and a multiple linear regression (MLR) model for prediction of soiling with meteorological and environmental parameters as variables. The study employ a special algorithm, the Boruta algorithm implemented in the random forest technique, to determine all the important soiling variables from the selected list of parameters for use in predictive modelling. Also, soiling forecasting model that uses the selected daily meteorological and environmental parameters to predict the amount of soiling on a solar PV collector. Although, [17] they use linear regression and back propagation neural networks to predict the performance of soiled PV modules. The work uses satellite data of solar irradiance and ambient temperature as input variables with maximum power point tracking. Although, the result shows that it is possible to predict maximum power output of soiled PV modules at about 97% accuracy. [11] Worked on the environmental performance of solar photovoltaic systems. The work focus on four factors which are the accumulation of dust, water droplets, birds' droppings, and partial shading conditions using the traditional technique. The result shows that increasing the area of shading on a PV module surface by a quarter, half, and three quarters resulted in a power reduction of 33.7%, 45.1%, and 92.6%, respectively. However, dust accumulation reduced the power output by 8.80% and the efficiency by 11.86%. [6] Proposes a mathematical model utilizing a MLR to predict and estimate PV output power. The datasets using the MLR model are composed of solar irradiance, cell temperature and wind speed. The result shows the effect of global solar irradiance I has a significant effect in the model compared to cell temperature T and wind speed V. [13] propose a statistical methodology to estimate the energy losses due to soiling deposition on solar modules using multiple linear regression (MLR) model. The datasets are composed of environmental samples (solar radiance, environmental temperature, humidity, and wind speed). The test is carried out to evaluate whether the difference between the predicted values by the MLR model and the observed values is due to the soiling.

However, there are several emerging models which are yet to be explored for modeling soiling phenomena. In our paper, a nonlinear novel computing algorithm for modeling of soiled PV module with a machine learning approach is presented. The Study utilize nonlinear regression models based on the Gaussian Process Regression (GPR) and Multi Linear Regression (MLR) due to their robustness and simplicity than other statistical regression techniques. Also, the learning-based GPR and MLR approach are in a manner that ensures robust stability and recursive feasibility with respect to the estimation error.

2. Study Area

Nigeria is the most populous and economically powerful country in Africa, which explains the huge demand for electricity [14]. It is located between 4°N and 14°N latitude and as a result, it receives a lot of solar radiation all year. This energy could be put to excellent use in solar power system development [12]. Nigeria has a high solar electricity potential and worldwide irradiation [5]. Every day, Nigeria receives an average of 19.8 MJm2/day of solar energy and 6 hours of sunlight. The potential for concentrated solar power and photovoltaic output of the country is estimated to be over 427,000 megawatts (MW) [20].The current research was conducted in Abuja, Nigeria's capital city, which is located in the heart of the Niger-Benue river confluence. Abuja is located at 9° 4' 20.1504" N, 7° 29' 28.6872" E, and is 840 meters (2760 feet) above sea level [1], [3].

3. Proposed Methodology

The goal of this research is to see if using linear models and AI-based models to estimate the expected power (EP) from a PV module under soiling conditions is achievable. The proposed methodology is summarized in Figure 1.

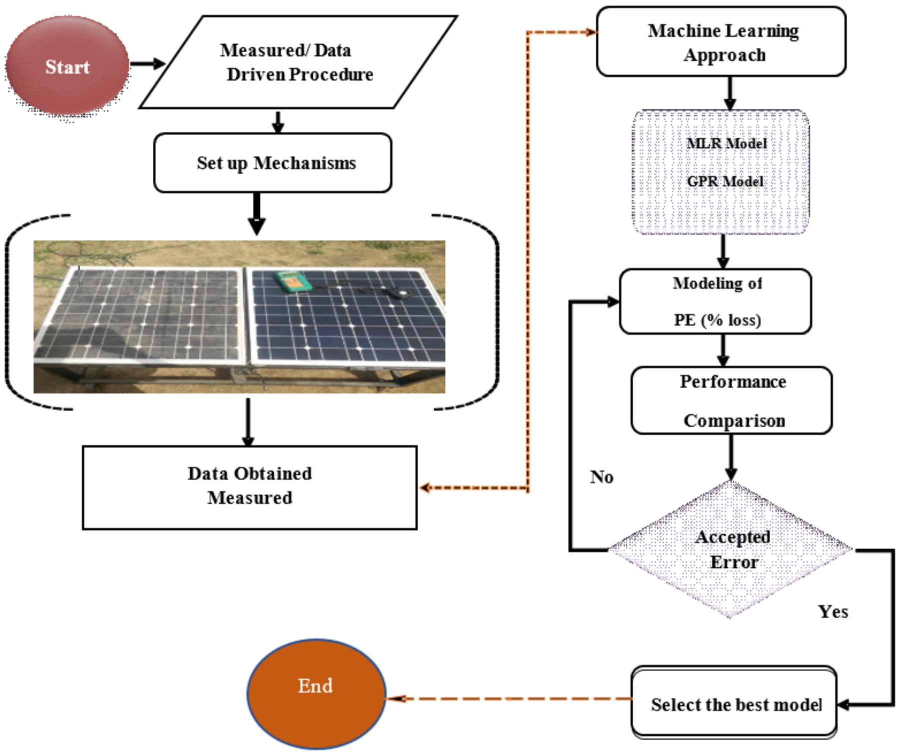

Figure 1: *Graphics of the proposed methodology.*

3.1 Multi Linear Regression

The Multi Linear Regression (MLR) model described in Eq (1) was used in this study because of its capacity to revise the weights and thresholds frequently in order to reduce the error function value [21].

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n, \tag{1}$$

where y is the dependent variable, a is the constant value or intercept of the regression line on Y-axis, b is the amount the response variable changes by the independent variables $x_i : x_1, x_2, \ldots, x_n$.

The vector A is determined such that the mean squared difference between the values of the linear regression predictions and the actual experimental data is minimized giving by Eq (2) [22].

$$\widehat{A} = Z^\top ZZ^\top Y \tag{2}$$

3.2 Gaussian Process Regression

The Gaussian Process Regression (GPR) is a set of random variables in which a limited number of them are integrated with Gaussian distributions [17]. GPR given as Eq (3), which is more useful in the prediction of many engineering problems due to its flexibility in giving uncertainty representation [18]. It also has a prior knowledge of functional dependency and data, GPR models are able to understand the predictive distribution corresponding to the test input. The models used under GPR were: Rational Quadratic (GPR-R), Squared Exponential (GPR-S) and GPR-M [9].

$$y_i = f(x_i) + \epsilon \tag{3}$$

In Eq (3), $f(x_i)$ stands for an arbitrary function that maps the inputs into the corresponding outputs, ϵ represents the regression error having an identically distributed Gaussian function with mean and variance values of zero and σ^2, respectively. The function $f(x)$ for any unobserved pair (x^*, f^*) in which f is the response and x is the explanatory parameter is obtained by:

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N_{n+1}\left(0, \begin{bmatrix} K(X,X) & k(X,x^*) \\ k(x^*,X) & k(x^*,x^*) \end{bmatrix}\right) \tag{4}$$

3.3 Data Collection and Processing

Data collection for this study was carried out in faculty of engineering, which is 8^o 58’ 38.4" N, 7^o 10’ 32.4" E, University of Abuja the capital city of Nigeria. The PV soiling effect was calculated based on field measurements at the test site. Concurrently, the environmental variables [15, 16] were also measured at the test site. In this study, we collected data on daily basis for three months spanning April 13, 2016 through July 13, 2016. The soiling ratio is the most frequent statistic for calculating soiling loss. The ratio between the performance of a soiled PV device in outdoor conditions and the performance of the same PV device without soiling is expressed by this measure, which is described in the IEC 61724-1 standard [19]. The measure of PV performance loss due to soiling measurements are available for 90 days, out of which 80 days were considered for modeling in which all selected variables’ measurements were available for the analysis. To ensure there was neither overfitting or underfitting in the training and testing data, two methods of cross validation were used. The hold-out cross validation approach was performed initially, with the data split. In this study, 75% of the data was utilized to train the models and 25% was used to test them. The Gaussian Process Regression approach was also tested using k-fold cross validation, where the data was separated into k equal numbers of sets using this validation procedure. The first set is utilized as test data on the first trial, while the remaining sets are used to train the model. The daily values after a cleaning or rain event were not included in this study.

4. Performance Evaluation Criteria

Several evaluation matrices were employed in assessing the efficiency of a model. Although, this study uses Nash-Sutcliffe efficiency (NSE), a measure which quantitatively describe the accuracy of model output [4] and Mean Absolute Error (MAE), is a measure of error between paired observations expressing the same phenomenon. It is appropriate for applications with linear cost functions. The equations for NSE and MAE are presented as:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(O_i - O_m)^2}, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \bar{y}_i, \tag{6}$$

where P_i is the predicted data, O_i is the measured data, O_m is the mean value of the reference data, N is the number of observations, y_i is the observed data, $\bar{y}_i$ is the forecasted output (or predicted time series), and y is the observed data.

5. Results and Discussion

The GPR and MLR regression models were employed to train and test an input combination from our measured dataset in this study. The dataset contains many daily readings of four independent variables (current, voltage, temperature, and wind speed), as well as a dependent variable (power expected) related to PV module soiling effects. Within the range of -1 and 1, the data was normalized. As different independent variables have similar effects on the behavior of the networks regardless of their magnitudes, this will ensure changes in the signals. One of the most important advantages of the GPR model over other machine learning methods is its ability to capture the uncertainty of the model directly. It is flexible, easy to implement, and gives the ability to add prior knowledge and specification about the shape of the model by selecting different types of kernels. GPR directly gives a distribution for the prediction value, rather than just one value as the prediction. All the models (Rational quadratic, squared exponential and Matern5/2) were developed using MATLAB software 2021a.

Table 1: *Results of the performance evaluation of the developed models.*

Model	Training phase		Testing phase	
	NSE	MAE	NSE	MAE
GPR-R	0.9745	0.0796	0.8604	0.0796
GPR-S	0.9745	0.0796	0.8604	0.0796
GPR-M	**0.9745**	**0.0784**	**0.8604**	**0.0796**
MLR	0.9738	0.0750	0.8564	0.0750

The modeling results, in terms of performance evaluation criteria (NSE and MAE) are presented in Table 1, which indicates the superiority of the GPR-Matern52 with the highest NSE (0.9745 and 0.8604) and lowest MAE (0.0784 and 0.0796) in both training and testing phases, respectively. Besides, the Fig. (2) a-d illustrated the correlation results of the predicted functions in comparison to the observed data points in the time series plot for both training and testing modes.

The NSE correlation between the four models using the radar plot is shown in Fig. 3 A radar plot is a multivariate analysis tool that can use quantitative indicators to better reflect qualitative problems. The Fig depict the superiority of GPR-M as it moves close to unity over the other models. Moreover, the bar plot as described in Fig. 4 proves that MLR for MAE (0.07502 and 0.07502) in training and testing is superior with minimal error over GPR-R, GPR-S, and GPR-M.

Fig. 5 shows the MAE correlation between the four models (GPR-R, GPR-S, GPR-M and MLR) using the radar plot. A radar plot is a multivariate analysis method that can be used to better represent

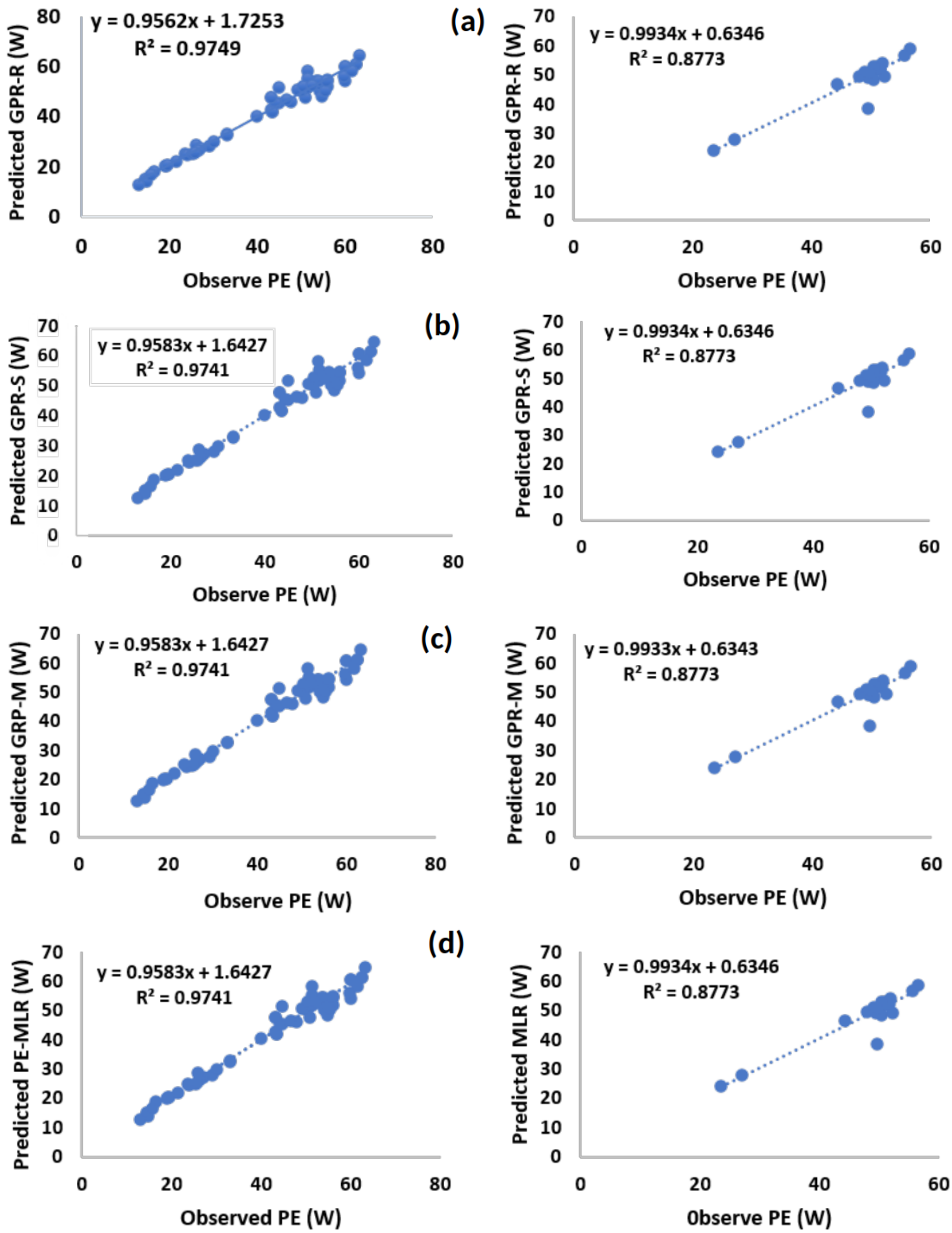

Figure 2: *Correlation Results of predicted and observed values of; (a) GPR-R (b) GPR-S (c) GPR-M and (d) MLR models in training and testing mode.*

qualitative issues using quantitative data. The graph depicts GPR-M's dominance over the other models as it approaches unity. Furthermore, the bar plot in Fig. 6 reveals the superiority of GPR-R, GPR-S, GPR-M as against MLR for NSE in training and testing is with small error.

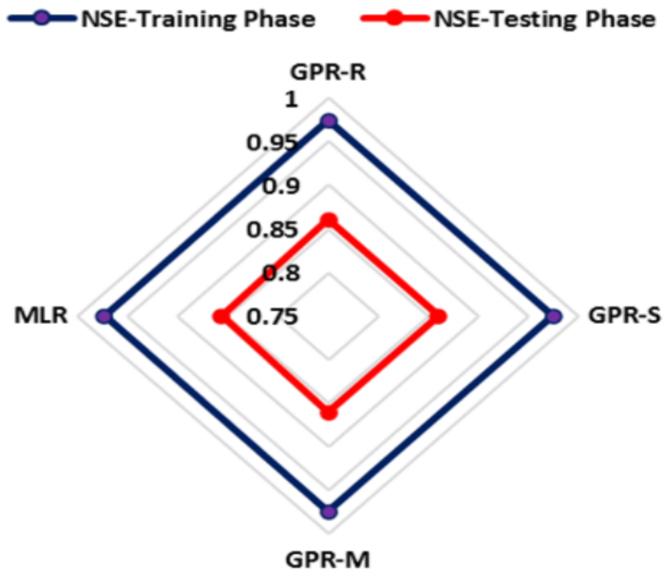

Figure 3: *NSE Radar plot for training and testing; GPR-R, GPR-S, GPR-M and MLR models.*

6. Conclusion

Predicting the power yield of PV modules is critical for supplying steady solar energy in micro-grids and grid integration. The feasibility of a nonlinear novel computing algorithm for the simulation of power expected with a machine learning approach is presented in this research. All of the models described here have the advantages of being simple to use and apply, which increases confidence in their use. The models' prediction skills were assessed using metrics such as the NSE and MAE. The results revealed that the nonlinear models developed in this work performed good in estimating the predicted power of a soiled PV module with minimal error in GPR-M (MAE = 0.0784 and 0.0784 and high efficiency in NSE = 0.9745 and 0.8604) in training and testing respectively. As a result, it can be stated that GPR-M is a promising method for predicting expected power that can also be used to predict other essential PV module attributes. However, it was discovered that the choice of

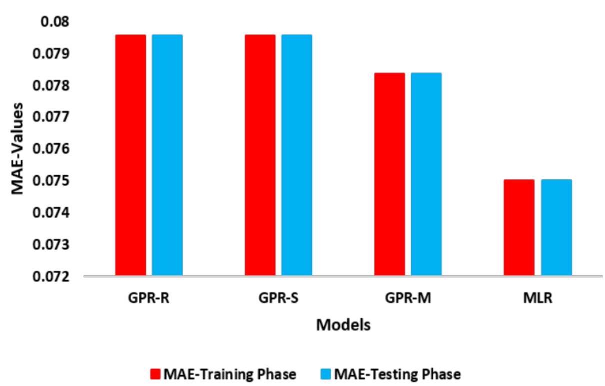

Figure 4: *Bar plot comparing the MAE values of; GPR-R, GPR-S, GPR-M and MLR models in training and testing mode.*

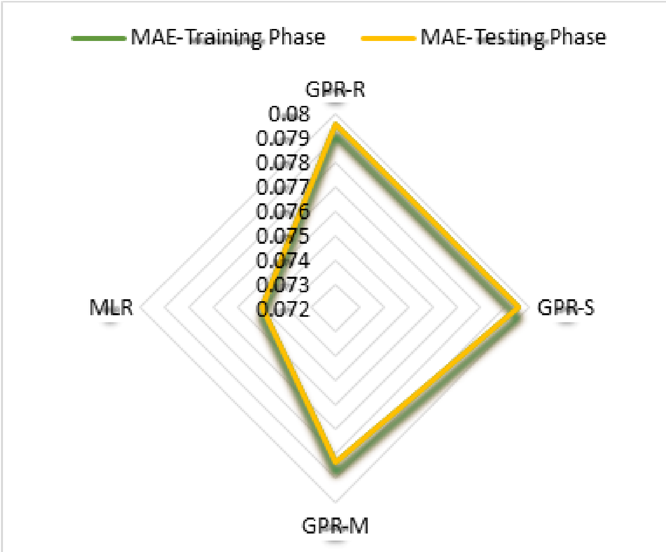

Figure 5: *MAE Radar plot for training and testing; GPR-R, GPR-S, GPR-M and MLR.*

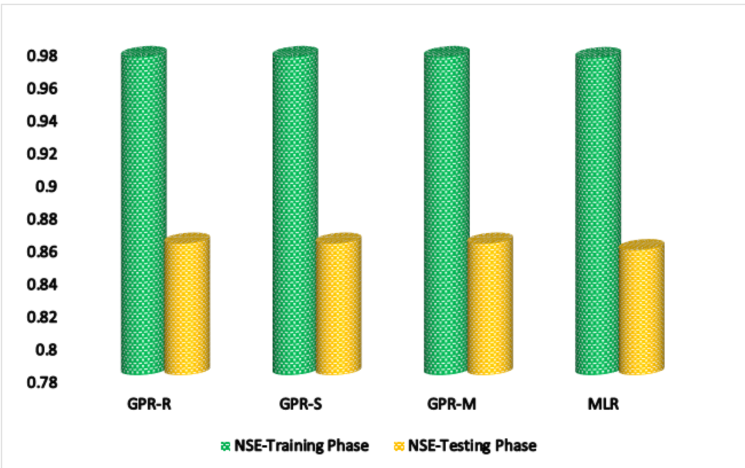

Figure 6: *Bar plot comparing the NSE values of; GPR-R, GPR-S, GPR-M and MLR models in training and testing mode.*

input parameters had an impact on the model. As a result, sensitivity analysis should be performed to assess the importance of input parameters, which may aid in the appropriate selection of input elements for improved predictive model performance.

Conflicts of Interest: The authors have no conflict of interest to any part.

References

[1] I. R. Abubakar. "Abuja city profile". In: *Cities* 41 (2014), pp. 81–91.

[2] M. Arsic et al. "Prediction of ozone concentration in Ambient air using multilinear regression and the artificial neural networks methods". In: *Ozone Sci. Eng* 42.1 (2020), pp. 79–88.

[3] S. A. Ayegba et al. "Assessment Of global solar radiation: A case study Of Abuja, Nigeria". In: *Int. J. Innov.* 3.13 (2016), pp. 342–346.

[4] G. Ayzel et al. "Streamflow prediction in ungauged basins: Benchmarking the efficiency of deep learning". In: *E3S Web Conf.* Vol. 163. 2020.

[5] O. Bamisile et al. "Application of deep learning for solar irradiance and solar photovoltaic multi-parameter forecast". In: *Energ. Source Part A* 00.00 (2020), pp. 1–21.

[6] A. Al Bashir et al. "Analysis of effects of solar irradiance, cell temperature and wind speed on photovoltaic systems performance". In: *Int. J. Energy Econ.* 10.1 (2020), pp. 353–359.

[7] K. Chiteka, R. Arora, and S. N. Sridhara. "A method to predict solar photovoltaic soiling using artificial neural networks and multiple linear regression models". In: *Energy Syst.* 11.4 (2018), pp. 981–1002.

[8] A. Ghosh. "Soiling losses: A barrier for India's energy security dependency from photovoltaic power". In: *Challenges* 11.1 (2020), p. 9.

[9] M. Jamei et al. "On the assessment of specific heat capacity of nanofluids for solar energy applications: Application of Gaussian process regression (GPR) approach". In: *J. Energy Storage* 33 (2021), p. 102067.

[10] M. Li and J. Wang. "An empirical comparison of multiple linear regression and artificial neural network for concrete dam deformation modelling". In: *Math. Probl. Eng.* 2019 (2019), p. 7620948.

[11] R. J. Mustafa et al. "Environmental impacts on the performance of solar photovoltaic systems". In: *Sustain.* 12.2 (2020), pp. 1–17.

[12] A. A. Osinowo et al. "Analysis of global solar irradiance over climatic zones in Nigeria for solar energy applications". In: *J. Sol. Energy* 2015 (2015), p. 819307.

[13] K. Ribeiro, R. Santos, and R. Rajagopal. "A statistical methodology to estimate soiling losses on photovoltaic solar plants a statistical methodology to estimate soiling losses on photovoltaic solar plants". In: *J. Sol. Energy Eng.* 143.6 (2021), p. 064501.

[14] S. Salisu et al. "Solar radiation forecasting in Nigeria based on hybrid PSO-ANFIS and WT-ANFIS approach". In: *Int. J. Electr. Comput. Eng.* 9.5 (2019), pp. 3916–3926.

[15] A. Shaju and R. Chacko. "Soiling of photovoltaic modules - Review". In: *IOP Conf. Ser. Mater. Sci. Eng.* 396.1 (2018), p. 012050.

[16] S. Shapsough, R. Dhaouadi, and I. Zualkernan. "Using linear regression and back propagation neural networks to predict performance of soiled PV module". In: *Procedia Comput. Sci.* 155 (2019), pp. 463–470.

[17] S. Shapsough, R. Dhaouadi, and I. Zualkernan. "Using linear regression and back propagation neural networks to predict performance of soiled PV modules". In: *Procedia Comput. Sci.* 155.2018 (2019), pp. 463–470.

[18] F. H. Sinz et al. "Learning depth from stereo". In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. Vol. 3175. 2004, pp. 245–252.

[19] G. P. Smestad et al. "Practical ANN model for predicting the PSS of two-way reinforced concrete slabs". In: *Sci. Rep.* 10.1 (2020), pp. 1–14.

[20] *Solar Report, Nigeria*. 2021.

[21] H. Sun and Y. Song. "Research on predictive control of students' achievements in physical education class and physical education class universities based on mathematical model analysis of multiple linear regression equations". In: *ACM Int. Conf. Proceeding Ser.* 2021, pp. 1420–1424.

[22] V. L. Tran and S. E. Kim. "Practical ANN model for predicting the PSS of two-way reinforced concrete slabs". In: *Eng. Comput.* 37.3 (2021), pp. 2303–2327.