POWER OUTPUT PREDICTIONS OF PHOTOVOLTAIC SYSTEM USING MACHINE LEARNING

Siyasanga I. May[1], Lawrence Pratt[1], Kittessa Roro[1], and Pitshou Bokoro[2]

[1]Council of Scientific and Industrial Research, Pretoria, South Africa, CSIR-Energy Centre, E-mail: smay@csir.co.za.

[1]CSIR Energy Centre, Pretoria, South Africa E-mail: lpratt@csir.co.za

[1]CSIR Energy Centre, Pretoria, South Africa, E-mail: kroro@csir.co.za

[2]University of Johannesburg, Faculty of Engineering and Built Environment, Department of Electrical and Electronics Engineering Technology, E-mail: pitshoub@uj.ac.za

Abstract

This work focuses on developing prediction models for the power output of multiple PV technologies installed at the outdoor test facility on the Pretoria campus of the Council for Scientific and Industrial Research. Random Forest (RF) and Adaboost machine learning models are trained with historic time-series data sets (measured meteorological and PV electrical parameters) to predict historical output power of the photovoltaic (PV) system. Sub-hourly measured data from January 2019 until November 2019 was averaged to hourly intervals for training and testing. The data undergo a pre-processing step where outliers are identified and removed. A very strong correlation ($r^2 \sim 0.99$) was calculated between Isc and PV output because PV output is largely determined by the plane of array irradiance and the resulting current generation. A strong correlation between PV output and plane of array ($0.89 < r^2 < 0.99$) and between PV output and module temperature ($0.62 < r^2 < 0.72$) are also calculated, depending on the module type. The models are then trained on the datasets and the accuracy is quantified based on the root mean squared error (RMSE) between the actual measured PV output and the predicted PV output of different PV technologies. RF generally outperformed the Adaboost regression. Both regression models achieved minimal RMSE on predictions for the thin film module technologies with maximum RMSE of 0.2 W for Adaboost and 1.2 W for the Random Forest. In future work, the trained models will be used to forecast future electricity production from PV plants using only forecasted weather data as inputs.

Keywords: Photovoltaic module; Random Forest; Adaptive Boosting; Power output predictions

1. Introduction

Solar Photovoltaic (PV) installations have been leading the renewable energy industry in the past few years, with the total installed global capacity of 627 GW by the end of 2019 [1]. The main driver for the evolution of this renewable technology has been dramatic reductions in cost and significant technological advancements [1], [2]. However, the variability in renewable resources brings challenges to the power system operator. These daily and seasonal variations in the grid-tied PV systems threaten the stability and reliability of the power network [3]. The ability to predict PV power output offers better system preparedness [4]. The application of artificial intelligence (AI) and supervised machine learning have been topics of interest in the PV space [5]–[7]. In PV predictions, these models have proven to be more reliable and economical than traditional methods on both PV generation and weather predictions. The computer algorithms are convenient in meeting specific aspects in the PV domain, be it in PV power prediction or weather related influences. [5], [7]–[12]. Supervised learning models work by finding the complex hidden data patterns between given inputs and map output values with great accuracy [13].

This study uses supervised machine learning (ML) approach to predict PV power output based on multiple inputs. Two machine learning models are considered in defining the prediction models: Random Forest regression and Adaptive boosting (Adaboost) regression. Initially, the dataset is pre-processed to prune out the outliers and lessen the training time and modelling errors resulting from an unfiltered dataset. The relative importance between parameters is assessed and correlations quantified. The prediction performances of both models are examined using regression metrics and various plots.

2. Methodology

2.1. System set up

The outdoor test facility stationed at the rooftop of building 34 at the Council for Scientific and Industrial Research (CSIR), Pretoria campus was utilized in this study. The system was built on a flat rooftop and hosts seven (7) pairs of different PV modules mounted on a fixed tilt rack facing true north (0^0 azimuth) at 25^0 tilt. The PV modules under test include Bi-facial PERC 270 Watt peak (Wp), Bi-facial n-type c-Si 280 Wp, Mono-facial mono-crystalline 275 Wp, Mono-facial mono-crystalline 330 Wp, Mono-facial poly-crystalline 315 Wp, Thin film 105

Wp and Thin film 175 Wp. Five of the seven module types are shown in Fig. 1 below

Fig. 1. Modules under test at the outdoor test facility on Building 34 in the CSIR Pretoria campus

Each PV module is connected to an individual maximum power point tracker (MPP) system coupled with the electronic load (EL) and is configured in accordance with the manufacturer rating label. The EL and MPP equipment are shown in Fig**.** 2 below.

Fig. 2. MPPT and Electronic Loads (EL)

The maximum power point (Pmpp) measurements for both thin film and crystalline Silicon (c-Si) technologies are measured every 1 minute and 10 minutes, respectively. Current-Voltage (IV) sweeps are carried out at every 2-minutes interval for thin films and 10-minutes for c-Si. Fig. 3 below shows the weather station also situated on Building 34. The measured meteorological datasets from the weather station includes a plane of array irradiance (PoA), temperatures (ambient and PV module temperatures), and wind speed all recorded at 1 minute intervals.

Fig. 3. CSIR weather station in Pretoria campus

2.2. Data Exploration

The hourly averaged PV system and meteorological data for the period from January 2019 to November 2019 is used for this research. The data set of seven (7) PV module pairs measured from 5 AM to 6 PM sun hours. The dataset features and parameters include open-circuit voltages (Voc), short-circuit current (Isc), maximum power point voltages (Vmpp), maximum power point current (Impp), maximum power (PV output), plane of array irradiance (PoA), wind speed, module temperatures and ambient temperatures. One-week hourly data of BYD module during March equinox (20th March 2019) is plotted in Fig. 4 below to show the measured electrical and weather parameters.

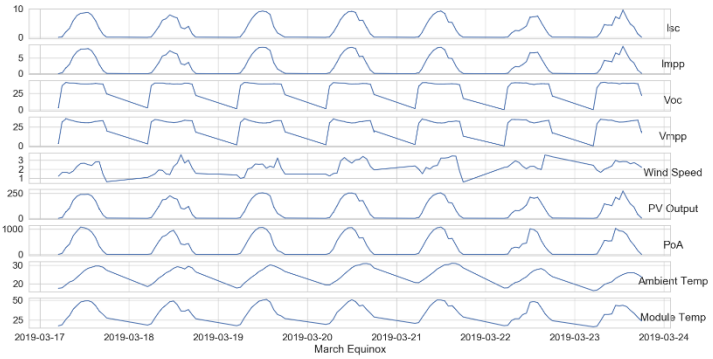

Fig. 4. PV and weather parameters near March equinox.

In this work, python programming language is used to train and evaluate the chosen ML models. All the measured data are used as input variables to the developed models except for the measured Pmp which is chosen as output variable in the study as shown in Fig. 5 below.

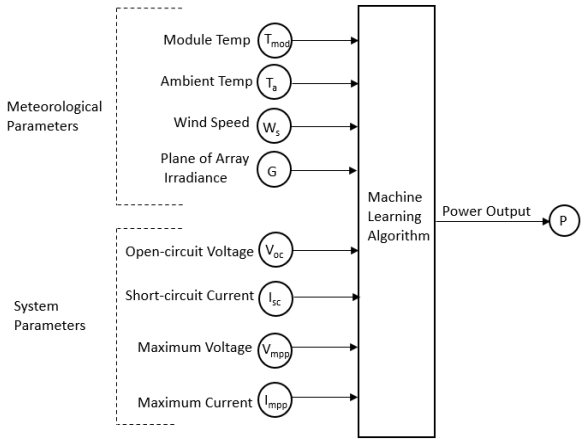

Fig. 5. Inputs and Output parameters

The data undergo a pre-processing step where possible invalid values are identified and removed. The outliers consist of missing and invalid values arising from maintenance and system faults. The relationship among the many parameters in the dataset is explored to assess the distributions and the correlations.

Fig. 6 below shows the univariate distributions of each electrical performance parameter along the diagonal and the pairwise scatterplots in the off the diagonal for all seven (7) PV module types, coloured by module name. The figure shows only the data from the test dataset. The linear correlation between Isc and PV output is strong, as expected. The square of the linear correlation coefficient (r^2) is greater than or equal to 0.99 for all module types, meaning 99% of the variability in the PV Output can be explained by the variability in Isc. The slope of each line varies depending on the module technology. The linear correlation between Voc and PV Output is weak ($r^2 < 0.31$) for all module types, so not as useful in predicting PV Output as the Isc.

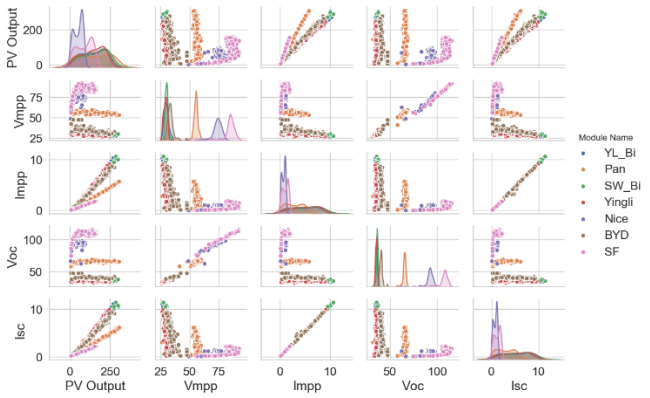

Fig. 6. Scatterplots showing the correlation among electrical parameters for the PV modules under test

Fig. 7 below shows the univariate distributions of the measured weather parameters along the diagonal and the correlation with PV Output, coloured by module name. The correlation between PV Output and PoA is strong ($0.89 < r^2 < 0.99$) depending on the module. The correlation between PV Output and module temperature is moderate ($0.62 < r^2 < 0.72$). Based on these correlations, the forecasting models should include the PoA irradiance and module temperature as important inputs when forecasting PV output.

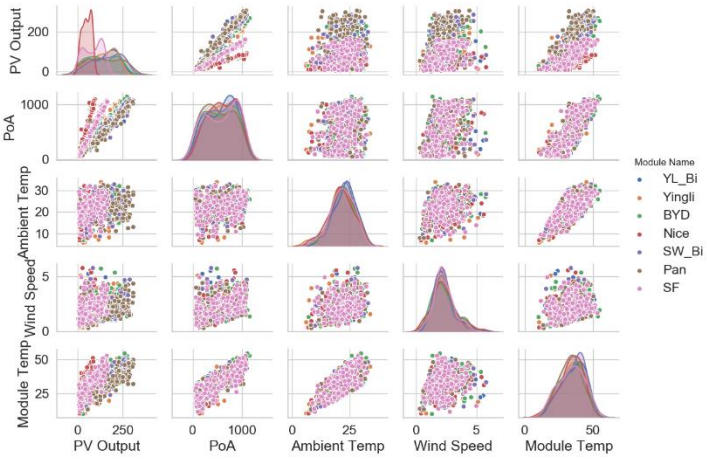

Fig. 7. Scatterplots showing the correlation among weather parameters and PV Output for the PV modules under test

2.3 Training and Testing

The cleaned data sets of each of the seven (7) module pairs are randomly sliced into training and testing subsets. The training data is used to optimize the model parameters and the test data is used to quantify the error in the predictions. Table 1 below summarizes the total number of observations and parameters in each data set.

Table 1. Data set allocation and sizes

Data subsets		Data Size		
		Rows	Columns	Size [%]
Original dataset	X	60203	8	100%
	Y	60203	1	
Training	X-Train	48162	8	80%
	Y-Train	48162	1	
Testing	X-Test	12041	8	20%
	Y-Test	12041	1	

Prior to training, each parameter in the data sets is normalized using the z-score as expressed in Equation 1. The z-score transforms the raw data from the measured units to standardized values for each parameter resulting in a mean of zero (0) and a variance value of one (1). This process helps to speed up the

training model completeness [16], [17].

$$z = \frac{x-u}{s} \quad (1)$$

Where u represents the mean value of the training sample and s represents the standard deviation.

2.4. Random Forest Regression Model

The random forest works on a crowd wisdom theory. It is centred on the principle where the group decisions carry more weight than individual decisions. This means during prediction each tree casts a vote and the majority vote wins. RF regression then calculates the average of all votes received to generate a great estimate of what the expected value should be. RFs fall in both supervised learning algorithms and ensemble algorithms. In supervised learning, inputs and resultant outputs are grouped into a training set where the model learns the hidden relationship between inputs and output features. The trained model is tested on the test data where new inputs are given and the model uses learned skill and predicts the unseen output. It ensembles a great number of decision trees into its final prediction [14]–[17].

2.5. Adaptive Boost (Adaboost) Regression Model

The Adaptive Boost (Adaboost) ensemble works on fitting sequences of weak learners that are modified or repeated to become better. This is done to minimize the loss function. The decision trees with single splits are called weak learners. The weighted sum of all predictions results in the final predictions. The number of weak learners are managed by n-estimators and the learning rate-parameters manage the contribution of the weak learners in the final combination [13], [18]. In practice, a uniform weight is allocated for each training dataset to determine its significance. When the assigned weights are high, that set of training data points have great influence on the training set. In the same way, when assigned weights are low, their influence in the training set is low. In Adaboost, the feature of importance in are:

- base_estimator: The weak learners used to train the model.
- n_estimators: Total number of weak learners to train in each iteration.
- learning_rate: It adds to the weights of weak learners and uses 1 as a default value.
- base_estimator: It is a weak learner used to train the model.

In both models, the architectural design of both ML models consist is as follows:

- Total number of inputs considered is eight (8) in both models.
- Then the number of n_estimators is one thousand five hundred (1500) in both models.
- The number of output or target value is one (1) in both models.
- The random state is set to forty two (42). This is used as a seed to the random generator to ensure that the model is deterministic and reproducible in each execution.

2.6. Model evaluation metrics

The developed machine learning models are evaluated using the conventional model evaluation metrics to assess the prediction performances of each model.

2.6.1 Mean Squared Error

The mean squared error is the quadratic error or loss calculated as:

$$MSE\ (y_i, \hat{y}_i) = \frac{1}{n_{sample}} \sum_{i=0}^{n_{sample}-1} (y_i - \hat{y}_i)^2 \quad (2)$$

Where y_i represents the actual values and $\hat{y}_i$ represents the predicted values of the i-th sample. The squared error is calculated for each hourly measurement in the data set. Then the MSE is calculated from the hourly errors for each time interval of interest. The calculated errors from MSE are watt squared (W^2) values. In this case, we consider RMSE on a monthly basis for each module.

2.6.2 Root Mean Squared Error

The root mean squared error is described as:

$$RMSE\ (y_i, \hat{y}_i) = \sqrt{\frac{1}{n_{sample}} \sum_{i=0}^{n_{sample}-1} (y_i - \hat{y}_i)^2} \quad (3)$$

Where y_i represents the actual values and $\hat{y}_i$ represents the predicted values of the i-th sample. The RMSE is calculated by taking a square root of the obtained results from Equation 2 above. This will results in the same units as the target variable while the MSE results in squared units.

3. Results

3.1. Evaluation of prediction errors per PV module

The trained Adaboost (AD) and Random Forest (RF) regression models are tested on the randomly selected 20% as shown earlier in Table 1 above. The predictions are inclusive of all the seven (7) PV module pairs. From the available test dataset, a single (sun hours) record of 2nd February 2019 is shown in Fig. 8 below to visualize the model predictions. The profiles for each module are very similar in shape since each module is exposed to the same PoA and mounted on the same tilt. The predicted Pmp by both models on each module type follow a similar trend and are

closely matched to the measured actual Pmp.

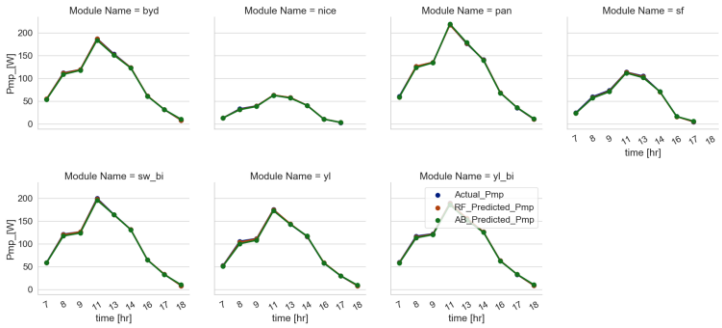

Fig. 8. Hourly actual Pmp and predicted Pmp of each PV module type

The monthly RMSE predictions of all module pairs from the test dataset are presented in Fig. 9 below. The RMSE values are calculated from the hourly records of each module and presented in monthly values. Then the RMSE results for each module type over the entire measurement period in hourly sums are also presented in Table 2.

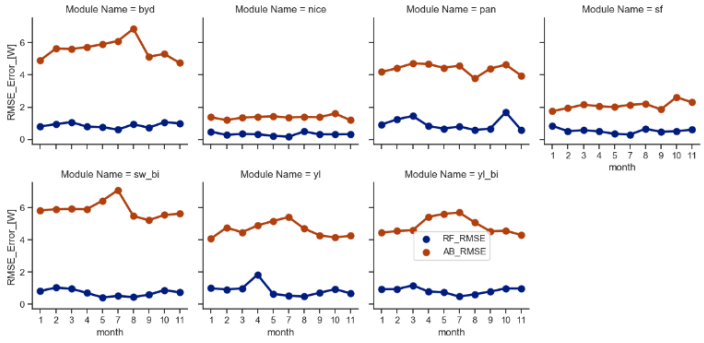

Fig. 9. RMSE of Pmp of each PV module manufacturer

On module level, RF prediction errors outperformed the AB prediction errors. Both the thin-film PV modules; Solar Frontier and Nice (SF& Nice) show minimal RMSE on both models with a maximum of 0.2 W and 1.2 W respectively.

In RF prediction, the highest RMSE record of 1.8 W is recorded in April by the Yingli (YL) PV modules followed by the Panasonic (Pan) PV modules on October with RMSE of 1.7 W. In AB predictions, the highest RMSE is on July for SolarWorld bifacial (SW-Bi) PV modules and August on BYD module with RMSE records of 7.1 W, 6.7 W respectively. Nice PV module is observed to have the minimum RMSE prediction errors on both RF and AB regression predictions with maximum RMSE values of 0.2 W and 1.2 W. Both models are observed to be sensitive to missing data resulting in higher prediction errors in monthly records.

In total sums for the entire measurement period (Table 2), the AB records the highest RMSE error across all module types with a maximum value of 58.9 W in the SW_Bi module and minimum of 13.7 W in the Nice module. The RF maintained the lowest RMSEs on total sums with maximum RMSE of 9.4 W in the Pan module and minimum of 3.2 W in the Nice module.

Table 2. Averaged actual, predictions and RMSE powers.

PV Module	Actual Power [Wh]	RF Pred. [Wh]	AB Pred. [Wh]	RF RMSE [W]	AB RMSE [W]
BYD	135222.1	135178.5	139211.7	8.7	55.9
Nice	43680.5	43669.6	44329	3.2	13.7
Pan	162153.2	162116.1	163506.6	9.4	43.8
SF	76576	76596.1	77699.6	5.3	21.1
SW-Bi	148072	148084.1	151071.1	7.1	58.9
Yl	130640.8	130666.4	132149.5	8.7	46.1
Yl-Bi	139945.0	139935.3	142518.5	8.4	48.8

5. Conclusion

In this work, the Adaboost and Random Forest machine learning models are trained and evaluated for predicting the PV output of different PV modules in the CSIR outdoor test facility. Analysis of the weather data shows the plane of array irradiance and module temperature have the strongest correlation with the PV output, so they must be included as inputs to the prediction models. The models are evaluated based on the RMSE of measured versus predicted power. The Random Forest algorithm achieved the lowest RMSE with no exceptions. On a technology level, all the RMSEs for thin film PV modules are lower compared to crystalline silicon PV modules for both machine learning models. The prediction error values measured in this paper clearly indicate that the Random Forest algorithm is superior to the Adaboost algorithm.

Acknowledgements

The author acknowledges the great support and assistance received from both CSIR, Energy Supply colleagues and University of Johannesburg. Also, personally thankful for the support received from my Wife.

References

[1] "Renewable Capacity Statistics 2020," */publications/2020/Mar/Renewable-Capacity-Statistics-2020.* /publications/2020/Mar/Renewable-Capacity-Statistics-2020 (accessed Apr. 13, 2021).

[2] "Snapshot of Global PV Markets - 2020," p. 20.

[3] S. Impram, S. Varbak Nese, and B. Oral, "Challenges of renewable energy penetration on power system flexibility: A survey," *Energy Strategy Rev.*, vol. 31, p. 100539, Sep. 2020, doi: 10.1016/j.esr.2020.100539.

[4] IRENA, “Future of Solar Photovoltaic - Executive Summary,” p. 8.
[5] E. Lorenz, J. Hurka, D. Heinemann, and H.-G. Beyer, “Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2009, doi: 10.1109/JSTARS.2009.2020300.
[6] W. Omran, M. Kazerani, and M. M. A. Salama, “Investigation of Methods for Reduction of Power Fluctuations Generated From Large Grid-Connected Photovoltaic Systems,” *Energy Convers. IEEE Trans. On*, vol. 26, pp. 318–327, Apr. 2011, doi: 10.1109/TEC.2010.2062515.
[7] A. Asrari, T. X. Wu, and B. Ramos, “A Hybrid Algorithm for Short-Term Solar Power Prediction—Sunshine State Case Study,” *IEEE Trans. Sustain. Energy*, vol. 8, no. 2, pp. 582–591, Apr. 2017, doi: 10.1109/TSTE.2016.2613962.
[8] S.-G. Kim, J.-Y. Jung, and M. Sim, “A Two-Step Approach to Solar Power Generation Prediction Based on Weather Data Using Machine Learning,” *Sustainability*, vol. 11, no. 5, p. 1501, Mar. 2019, doi: 10.3390/su11051501.
[9] F.-V. Gutierrez-Corea, M.-A. Manso-Callejo, M.-P. Moreno-Regidor, and M.-T. Manrique-Sancho, “Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations,” *Sol. Energy*, vol. 134, pp. 119–131, Sep. 2016, doi: 10.1016/j.solener.2016.04.020.
[10] A. K. Yadav and S. S. Chandel, “Solar radiation prediction using Artificial Neural Network techniques: A review,” *Renew. Sustain. Energy Rev.*, vol. 33, pp. 772–781, May 2014, doi: 10.1016/j.rser.2013.08.055.
[11] M. Mishra, P. Byomakesha Dash, J. Nayak, B. Naik, and S. Kumar Swain, “Deep learning and wavelet transform integrated approach for short-term solar PV power prediction,” *Measurement*, vol. 166, p. 108250, Dec. 2020, doi: 10.1016/j.measurement.2020.108250.
[12] S. Park, Y. Kim, N. J. Ferrier, S. M. Collis, R. Sankaran, and P. H. Beckman, “Prediction of Solar Irradiance and Photovoltaic Solar Energy Product Based on Cloud Coverage Estimation Using Machine Learning Methods,” *Atmosphere*, vol. 12, no. 3, Art. no. 3, Mar. 2021, doi: 10.3390/atmos12030395.
[13] Y. Zhang, *New Advances in Machine Learning*. BoD – Books on Demand, 2010.
[14] J. Brownlee, “How to Develop a Random Forest Ensemble in Python,” *Machine Learning Mastery*, Apr. 19, 2020. https://machinelearningmastery.com/random-forest-ensemble-in-python/ (accessed Mar. 12, 2021).
[15] “Random forest vs SVM,” *The Kernel Trip*. /statistics/random-forest-vs-svm/ (accessed Apr. 19, 2021).
[16] J. Brownlee, “Random Search and Grid Search for Function Optimization,” *Machine Learning Mastery*, Mar. 07, 2021. https://machinelearningmastery.com/random-search-and-grid-search-for-function-optimization/ (accessed Mar. 11, 2021).
[17] “Machine Learning | An Introduction.” https://www.c-sharpcorner.com/blogs/machine-leaning-an-introduction (accessed Apr. 28, 2021).
[18] “A Guide To Understanding AdaBoost,” *Paperspace Blog*, Feb. 23, 2020. https://blog.paperspace.com/adaboost-optimizer/ (accessed May 09, 2021).