Real time photovoltaic power forecasting and modelling using machine learning techniques.

Rita Mwende [1*], *Sebastian* Waita[1], *Geoffrey* Okeng'o [2]

[1] Condensed Matter Research Group, Department of Physics University of Nairobi, Kenya
[2]Astrophysics & Space Science Research Group, Department of Physics University of Nairobi, Kenya

Abstract. Photovoltaic (PV) system installations have increased in recent years partly due to growing energy needs from a rising population. Such PV systems producing electricity contribute in promoting green energy. However, solar energy is highly intermittent and uncontrollable due to its high spatial and temporal variations of atmospheric conditions. With such variability, PV power forecasting is therefore crucial for full integration of solar energy into the grid. In this study, Support Vector Regression (SVR) and Random Forest Regression (RFR) models were built and used to forecast real-time PV power output of a 1.5kW solar PV system installed at the Department of Physics, University of Nairobi in Kenya. SVR model outperforms RFR model with root mean square (RMSE) of 43.16 adjusted R^2 of 0.97 and mean absolute error (MAE) of 32.57 on the validation. Dataset compared to RMSE of 86, adjusted R^2 of 0.90, MAE of 69 were obtained for RFR model. A real time power forecast application based on the SVR model was successfully built using the Shiny application in R software. This shows that SVR model is more robust than RFR and has capabilities of reducing errors during computations. **Keywords**: Photovoltaic system; Power forecasting; Support Vector regression; Random Forest Regression.

1 Introduction

Kenya's economic growth has led to a rise in the demand for electricity from 1802 MW in 2018 to 1912MW in November 2019, rising steadily by 3.6% annually[1] 74.5% of Kenya's energy demand is provided by wind, hydropower, solar and geothermal power which are all renewable energy sources with fossil fuel only supplying 25.5% to the energy mix. Majority of the power is derived from hydropower supplying approximately 677MW followed by geothermal 670 MW of the total 2.7GW installed capacity [1] However, hydropower capacity is adversely affected by long periods of drought which have been experienced since 2015. On the other hand, geothermal power has great potential of providing up to 10 GW power [2].

However, rising investment charges, land disputes, lack trained personnel, huge grid infrastructural investment hinder its full exploitation [3] The focus of renewable energy has shifted to solar energy due to its abundance and availability.
Due to her location at the equator, Kenya receives an abundance of solar energy averaging between 5-7 sunshine hours and 4-6 kw/m^2 insolation daily [4]. Photovoltaics are highly popular source of solar energy because they require low maintenance, silent and clean energy[5].

However, solar power generation is heavily dependent on the variation of weather parameters such as temperature, relative humidity, dust accumulation and wind speed [6][7]. This inherent fluctuating nature of solar energy poses a major challenge in the quest to fully integrate solar energy power plants into existing power grids without compromising on the stability of the power output.
Hence, proper energy budgeting and planning, requires the development of reliable predictive and forecasting models able to provide accurate performance forecasts and modelling information for PV solar systems power output.
PV forecasting methods include physical, statistical and hybrid methods. Statistical methods have become popular because they are much simpler to implement, require less input data than traditional methods hence have low computational costs [8] Solar PV power forecasting models were based solely on the use of historical solar irradiance data on assumption that that solar irradiance is the only parameter influencing the performance of PV system [9]
However, this assumption led poor forecast, hence led researchers to include other weather parameters as inputs in PV predictive models. Models that used parameters such as wind speed, ambient temperature, solar irradiance, relative humidity yielded highly accurate predictions [7-10].
They also study the relationships between the weather variables and can determine variable importance.

* Corresponding author: mwenderita74@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Predictive models that used feature selections techniques such as principal component analysis (PCA) to reduce dimensionality which improves the performance of the models

This work aimed at creating interactive models using Support Vector machines and Random forest regression that accurately predict PV power output using real-time observations and weather data using support vector machine and random forest.

1.1 Support vector machines

Support vector machines (SVM) is used to perform both classification and regression. It involves the construction of a separation hyperplane or collection of hyperplanes to execute regression on high dimensional data. When the algorithm gets labeled training data it forms the optimum hyperplane which separates new sample data with main the goal being to find a hyperplane $f(x)$ that maximum error (ε) from the training data and should be as flat as possible [11].

Hyperplane $f(x)$ is expressed by the linear equation

$$f(x) = w_i x_i + b \quad (1)$$

where b is the slack variable

In SVR, the set absolute error or deviation from the hyperplane should be less or equal to the specified margin called the maximum error ε whose value parameter can be tuned to achieve high accuracy in a model. To ensure the flatness, one has to ensure w is has small as possible this is done by optimizing the problem to give [11]

$$Min\frac{1}{2}|w|^2 \quad (2)$$

subject to

$$\begin{aligned} y_{i\_}w_i x_i + b \le \varepsilon \\ w_i x_i + b - y_i \le \varepsilon \end{aligned} \quad (3)$$

where ε is the maximum error , Min is minimize.

Most case errors may occur beyond the ε we denote the deviation from the margin as $\xi_{i\_}$, Equation (3) now expressed as shown below [11]

$$Min\frac{1}{2}|w|^2 + C\sum_{i=1}^{l} \xi_i \quad (4)$$

Constraints

$$\begin{aligned} y_i - wx_i - b \le \varepsilon + \xi_i \\ w_i x_i + b - y_i \le \varepsilon + \xi_i \end{aligned} \quad (5)$$

where $C > 0$ is the penalty parameter of the error term. When C increases the tolerance for points outside the ε also increases and as C approaches zero the tolerance approaches zero [11].

SVR has gained popularity because it can effectively classify non-linear data by mapping inputs into high-dimensional feature spaces even when the datasets are small. Kernel function enables one to locate a hyperplane in the higher dimensional space without elevating computational cost. Increase in the dimension of data leads to a rise in the computational cost. When dimension increases and the separating hyperplane is not found in a particular dimension, a kernel is expected to shift the data to a higher dimension support vector classifier. This is achieved by adding a kernel trick which transforms the classes into a higher dimensional space, where classes can be linearly separated [11]. Kernels are classified into linear, polynomial, radial basis function kernels. Function (ϕ) maps training vectors (x_i) into higher dimensional space, this is known as the kernel trick $K(x_i, x_j)$ expressed by the equation [11]

$$K(x_i, x_j) \equiv \phi(x_i)^T (x_j). \quad (6)$$

Furthermore, SVR is less prone to overfitting issue, after training the prediction phase is rapid and works well with high dimensional data..

1.2 Random Forest Regression

Random forest regression (RFR) involves growing of trees depending on random vector Θ_k such that the tree predictor $h(x, \Theta)$ takes a numerical value. A random forest is built by taking an average over k of the trees to reduce the variance hence finding a balance between the two extremes which is expressed as [12]

$$RFR = \{h(x, \Theta_k), \quad k = 1, \ldots\} \quad (7)$$

where $\{\Theta_k\}$ is the random vector and $h(x, \Theta)$ is the tree predictor

Random vectors $\{\Theta_k\}$ are independently identically distributed and each tree selects the most popular class at input x vectors [12]. The mean squared generalization error (GE) for predictor is $h(x)$ is given by the equation

$$GE = E_{xy}(Y - h(x))^2 \quad (8)$$

where E_{xy} is expected value

The GE for forests converges as to a limit as the number of trees increases. For an accurate RFR model low correlation between residuals and low error trees are key [12]. The more the number of trees the more robust the forest becomes. The RFR do not over fit data as more tress are added but GE is produced.

1.3 Accuracy Metrics for the evaluation of prediction models

Several metrics are used to determine the accuracy solar (PV) prediction models based on ML techniques. They include mean squared error (MSE), coefficient of determination (R^2), Adjusted R^2 and mean absolute error. The MSE measures an average value of the squares of errors, expressed in the equation [13]

$$MSE = \frac{1}{B}\sum_{i=1}^{B}(y_i - y_p)^2 = RMSE^2 \quad (9)$$

where y_i is the i-th actual value, $y_p$is the predicted value for y_i, B is the number of samples, and RMSE is the square root of MSE.

When the RMSE decreases the predictive model's performance increases. Mean absolute error (MAE) is the average difference between the predicted and real values, it is computed using the equation

$$MAE = \frac{1}{B}\sum_{B}|y_i - y_p| \quad (10)$$

The MAE shows measure of errors between the predicted values and the real values but does not indicate the direction of the error. Coefficient of determination (R^2) is the proportion of the variance of the dependent variable which the independent variables describe as expressed in the equation [13]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_p)^2}{\sum_{i=1}^{n}(y_i - y_m)^2} \quad (11)$$

where y_m is the mean of the actual values of y_i.

The adjusted R-squared ($Adj\ R^2$) is a modification of R-squared and only increases when the independent variable is significant. It is expressed in the equation [13]

$$Adj\ R^2 = 1 - (1 - R^2)\frac{B-1}{B-p-1} \quad (12)$$

where p is the total number of independent variable.

2 . Materials and Method

2.1 Experimental setup

This study was on a 1.5kW PV string installed at the Department of Physics, Chiromo Campus, University of Nairobi, Kenya. It consisted of six 250W Polycrystalline Solinc solar panels connected in series. The solar panels were first cleaned using a clean cloth and plain water before commencing with the measurements. Solar irradiance was measured using a HT304N Reference Cell while the PV module temperature was measured using a PT300N temperature sensor. A current-voltage (IV) analyzer was used to measure the current and voltage of the solar PV system. The data was collected from 10:00 a.m. to 3:00 p.m. EAT at 30 minutes' interval for 21 days. Data analysis was done using R–software and Origin 9.1 software. The data was then pre processed using Principal component analysis. The preprocessed data was divided into training, testing and validation dataset using the ratio 60%, 20% and 20% respectively. Support Vector regression (SVR) predictive model using radial basis kernel model and Random Forest Regression (RFR) model of 20 trees were built in the R software Cross validation (CV) was done using the *leave one out cross validation (LOOCV)*, *k -fold* and random resampling on the training dataset, in order to prevent the models from overfitting. The performance of the trained model was evaluated using the mean absolute error (MAE), root mean square error (RMSE) and the coefficient of determination (R^2) on the training and testing dataset validation dataset and finally tested using the test set.

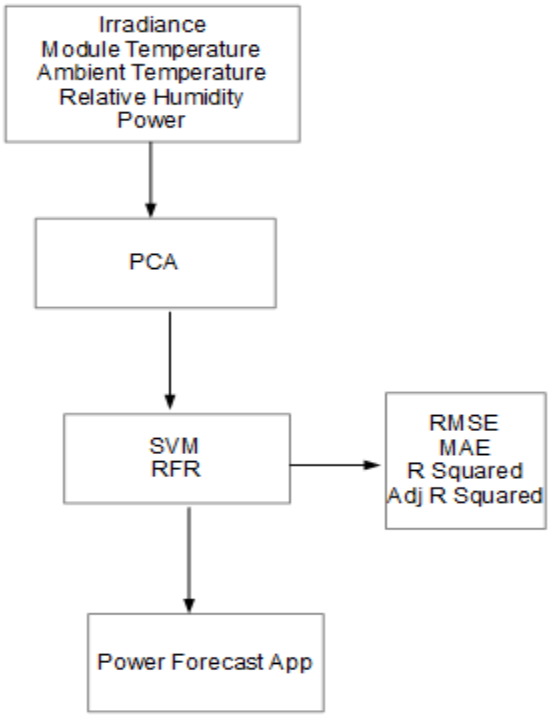


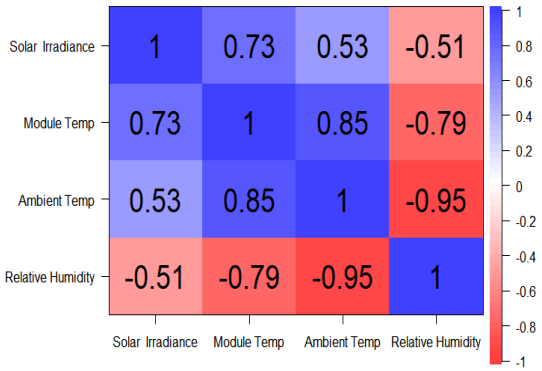
Figure 1:Block diagram showing a summary of the methodology

3 .RESULTS AND DISCUSSION

3.1 Principal component analysis

Exploratory data analysis was done to discover trends and relationships between 231 observables and 5 variables collected for a period of three weeks.

Table 1 :Correlation matrix showing the degree of correlation between the measured weather parameters



A correlation matrix containing Pearson's coefficient correlation values between solar irradiance, ambient temperature, module temperature and relative humidity was obtained as shown in Table 1

The correlation coefficients shown in Table 1 are greater than 0.70 indicate high correlation between variables hence indicate redundant information in the data which often decrease accuracy of a predictive model. Principal component analysis was used to remove redundant information from the measured weather parameters resulting into four uncorrelated principal components (PC's).The first (PC1), second (PC2), third (PC3) and fourth principal (PC4) components explained 79.86%, 15.53%, 3.70% and 1.113% of variance of the data respectively. The PCA results indicate that the first two principal components (PC1 and PC2) account for majority (95.19 %) of the variability of the dataset.

3.2 Support Vector Regression

SVR predictive model was built and its performance evaluated. Table 2 shows that the model trained using the LOOCV yielded the lowest RMSE ,40.40 compared to k-fold (3), and CV random resampling are 40.70 and 47.96 respectively. The performance is the built trained models on new tested dataset shows that the LOOCV and to k-fold technique yielded same value of 45.10 while the CV resampling technique yielded highest RMSE of 50.30.

Table 1:Performance evaluation of SVR training data set and test dataset based on k-fold, "LOOCV" and CV (Random resampling) employed

	Root mean square error (RMSE)		**Coefficient of determination (R^2)**		**Mean absolute error (MAE)**	
Cross Validation technique	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
k − fold(	40.70	45.10	0.98	0.97	30.40	29.01
LOOCV	40.40	45.10	0.98	0.97	29.01	29.27
CV (Random resampling)	47.96	50.30	0.97	0.96	31.72	32.06

3.3 Random Forest Regression

The leave cross validation yielded the best model with highest R^2 of 0.96 and the lowest MAE and RMSE of 51 and 65 on training dataset as shown in Table 3. However, on the testing dataset *k-fold(3) cross* validation technique outperformed the *LOOCV* and CV random resampling with lowest RMSE and MAE of 84.4 and 62.2 respectively and the highest R^2 of 0.90 as shown in Table 3

Table 2:Performance evaluation of RFR training data set and test dataset based on k-fold, "LOOCV" and CV (Random resampling) employed

	Root Mean Absolute Error (RMSE)		**Coefficient of determination (R^2)**		**Mean Absolute Error (MAE)**	
Cross Validation techniqu e	Train	Test	Train	Test	Train	Test
k − fold(3	76W	84.4	0.94	0.90	58.1	62.2
LOOCV	65W	94	0.96	0.87	51.8	68
CV (Random resampling)	79W	88.89	0.94	0.88	63.98	64.5

The evaluation of the performance of the SVR *LOOCV* and RFR *k-fold* (3) trained was evaluated further using a validation dataset.

The results showed that SVR yielded highest R^2 of 0.97 compared to 0.90 of RFR. While the MAE was at 32.57 for SVR model against 69 for the RFR model performance on the validation dataset as shown in Table below 4.

Table 3:Comparison of performance of RFR and SVR based on performance on validation dataset

ML technique	**RMSE**		**Adj R^2**		**MAE**	
	Train	**Valid**	**Train**	**Valid**	**Train**	**Valid**
RFR *k-fold (3)*	76	86	0.95	0.90	58.1	69
SVR *(LOOCV)*	40.4	43.16	0.98	0.97	29.01	32.57

3.4 Power forecast application based on the 1.5kW PV Solar system

The power forecast application was successfully built using the Shiny application in R environment based on the SVR model using the "LOOCV" technique. The application is equipped with input buttons namely; solar irradiance, module temperature, ambient temperature and relative humidity and then outputs the real time predicted PV power output based on the trained SVR model as shown in Figure 2.The figure 2 shows a demonstration of random input parameters feed into the application and real time PV power forecast.

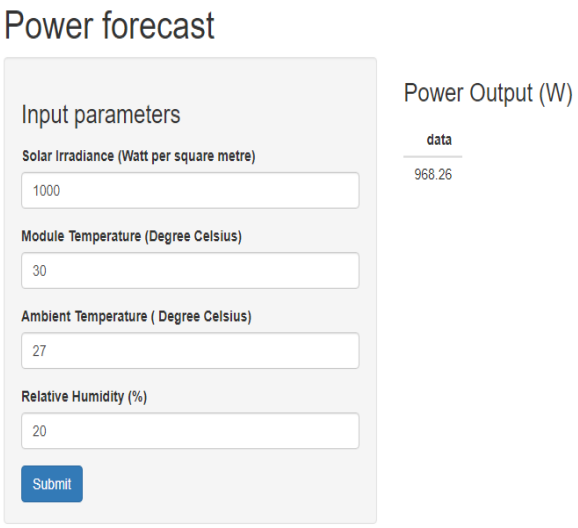


Figure 2 Power Forecast application user interface with input buttons, submit button and PV Output power.

4 CONCLUSION

An experimental setup was successfully installed to collect performance data of the 1.5kW PV system in varying weather conditions for a period of three weeks. PV power forecasting models coupled with PCA were built using SVR and RFR and were successfully trained, validated and tested to forecast real-time PV power output. SVR model. SVR model outperforms RFR model on the validation. dataset. Power forecast application was also built in R shiny based on the SVR LOOCV model built. This shows great potential on the development of site-specific and dynamic solar PV forecasting models.

Data collection for a longer period is recommended to ensure a large dataset is used for both training and testing hence increasing model accuracy. Data be collected from different sites with varying weather conditions to ensure inclusivity in the interactive predictive app built.

We acknowledge the support from the Department of Physics, University of Nairobi for providing the equipment used in this research, and the University of Nairobi postgraduate scholarship for the funding towards this research.

REFERENCES

[1] Africa. Energy. Series, "Kenya Special report 2020," (2020).

[2] P. F. Achieng, B. Davidsdottir, and I. Birgir, "Potential contribution of geothermal energy to climate change adaptation : A case study of the arid and semi-arid eastern Baringo lowlands , Kenya ☪," *Renewable and Sustainable Energy Reviews*, **vol. 16**, no. 6, pp. 4222–4246, 2012, doi: 10.1016/j.rser.2012.01.081.

[3] D. Samoita, C. Nzila, and P. A. Østergaard, "Barriers and Solutions for Increasing the Integration of Solar Photovoltaic in Kenya ' s Electricity Mix," (2020).

[4] Solargis, "Kenya_PVOUT_mid-size-map_156x220mm-300dpi_v20191015." 2019.

[5] A.M.K. El-Ghonemy, "Photovoltaic Solar Energy : Review," *International Journal of Scientific & Engineering Research*, **vol. 3**, no. 11, pp. 1–43, 2012, [Online]. Available: https://www.ijser.org/researchpaper/Phot

[6] F. Touati, M. A. Al-Hitmi, N. A. Chowdhury, J. A. Hamad, and A. J. R. San Pedro Gonzales, "Investigation of solar PV performance under Doha weather using a customized measurement and monitoring system," *Renewable Energy*, **vol. 89**, pp. 564–577, (2016), doi: 10.1016/j.renene.2015.12.046.

[7] A. Khandakar *et al.*, "Machine Learning Based Photovoltaics (PV) Power Prediction Using Di ff erent Environmental Parameters of Qatar," (2019).

[8] M. G. De Giorgi, P. M. Congedo, and M. Malvoni, "Photovoltaic power forecasting using statistical methods : impact of weather data," no. September (2015), doi: 10.1049/iet-smt.2013.0135.

[9] C. Wu and Y. Lou, "Predicting solar generation from weather forecasts," pp. 528–533, (2011).
[10] M. Madhiarasan and S. N. Deepa, "Review of Forecasters Application to Solar Irradiance Forecasting," **vol. 2**, no. 2, pp. 26–30, (2017).
[11] A. J. Smola and B. S. C. H. Olkopf, "A tutorial on support vector regression ∗ ¨," pp. 199–222, (2004).
[12] L. Breiman, "Random Forests," pp. 1–33, (2001).
[13] 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, (2019), doi: 10.3390/su11051501.