Heliyon 10 (2024) e28898

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Machine learning forecasting of solar PV production using single and hybrid models over different time horizons

Shadrack T. Asiedu [a], Frank K.A. Nyarko [a], Samuel Boahen [a,*], Francis B. Effah [b], Benjamin A. Asaaga [a]

[a] Department of Mechanical Engineering, Kwame Nkrumah University of Science and Technology Kumasi, PMB, Kumasi, Ghana
[b] Department of Electrical Engineering, Kwame Nkrumah University of Science and Technology Kumasi, PMB, Kumasi, Ghana

ARTICLE INFO

Keywords:
Solar PV forecasting
Machine learning
Artificial neural network
Hybrid models

ABSTRACT

This study uses operational data from a 180 kWp grid-connected solar PV system to train and compare the performance of single and hybrid machine learning models in predicting solar PV production a day-ahead, a week-ahead, two weeks ahead and one month-ahead. The study also analyses the trend in solar PV production and the effect of temperature on solar PV production. The performance of the models is evaluated using R^2 score, mean absolute error and root mean square error. The findings revealed the best-performing model for the day ahead forecast to be Artificial Neural Network. Random Forest gave the best performance for the two-week and a month-ahead forecast, while a hybrid model composed of XGBoost and Random Forest gave the best performance for the week-ahead prediction. The study also observed a downward trend in solar PV production, with an average monthly decline of 244.37 kWh. Further, it was observed that an increase in the module temperature and ambient temperature beyond 47 °C and 25 °C resulted in a decline in solar PV production. The study shows that machine learning models perform differently under different time horizons. Therefore, selecting suitable machine learning models for solar PV forecasts for varying time horizons is extremely necessary.

1. Introduction

Electricity demand has risen globally, driven by population growth and industrialization [1,2]. This increased demand has primarily been met by electricity generated from fossil fuel sources. This has contributed to the increased greenhouse gas emissions and environmental degradation [3]. Renewable energy sources such as solar PV are increasingly being adopted as an alternative to non-renewable sources to reduce greenhouse gas emissions. Yet, their intermittent nature introduces uncertainty and instability into the power system [4]. As such, managing modern power systems with high penetration of solar PV has become more complicated, requiring accurate forecasts of solar PV production.

Many studies have proposed several physical and machine learning forecasting models to address this problem [5–9]. Machine learning models are, however, widely used due to their use of solar PV's historical data, which are a more accurate reflection of the actual system's performance. In addition, machine learning models can make forecasts without the design parameters of the solar PV system.

The Decision Tree algorithm with a tree depth of 7 was used in Ref. [10] to predict the output power of a solar PV plant with a Mean

* Corresponding author.
E-mail address: s.boahen@knust.edu.gh (S. Boahen).

https://doi.org/10.1016/j.heliyon.2024.e28898
Received 6 January 2024; Received in revised form 24 March 2024; Accepted 26 March 2024
Available online 29 March 2024

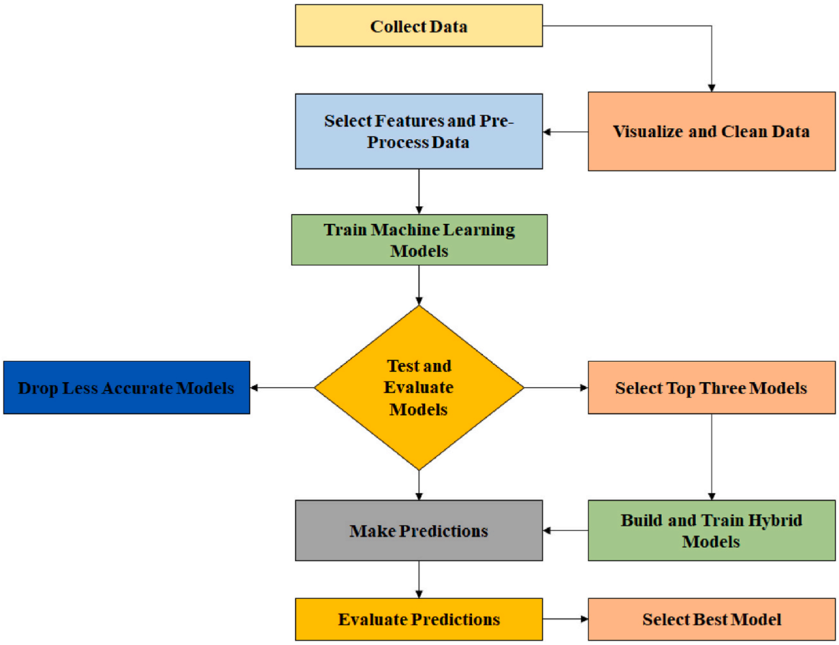


Fig. 1. Flowchart of research methodology.

Absolute Percentage Error (MAPE) of 0.77%. A similar study by Ref. [11] used Random Forest to forecast the output power of Yarmouk University's PV solar system in Jordan. Their model successfully forecasted a total yield of 5548.96 MWh at a performance ratio of 95.73%. In Ref. [12], Support Vector Regression achieved a Root Mean Square Error (RMSE) of 318.4 for the hourly forecast of solar PV production.

A further study by Ref. [13] focused on the daily, hourly and 15-min ahead forecast of solar PV production. Observations from their study indicated that Random Forest performed better than Support Vector Regression, Linear Regression and Artificial Neural Network with an RMSE of 32. Further, in Ref. [14], Random Forest was used to predict solar PV production with an average Mean Absolute Error (MAE) of 1.0096 and a RMSE of 1.5878. Again, in Ref. [15], Random Forest performed better than Support Vector Regression and Decision Tree in forecasting solar PV production a day ahead. Their model obtained an RMSE of 95.32 and MAE of 50.21. The XGBoost algorithm's effectiveness in forecasting solar PV production was also explored in Refs. [16,17] for day-ahead and hour-ahead scenarios. Their studies observed significant improvement in the prediction accuracy.

Other studies also attempted to address this problem using artificial neural networks. For instance Ref. [18], compared the performance of an Artificial Neural Network with Random Forest, Decision Tree, XGBoost and LSTM in predicting the intra-day solar PV production. The findings showed that ANN performed better than all the other models, with an RMSE of 0.9988. An RMSE of 1.5565 was achieved by Ref. [19] when ANN was used to forecast solar PV production days ahead. Similarly, using solar irradiation, ambient temperature and model temperature as input variables, the study in Ref. [20] built an Artificial Neural Network that successfully predicted solar PV production an hour ahead and a day ahead. Further, a 24-h ahead forecast with ANN was experimented with by Ref. [21], where a MAPE of 10.6% was obtained on sunny days, and 18.89% was recorded on rainy days.

Researchers have also explored combining multiple models into hybrid models to improve the predicting performance of individual models. In Ref. [22], k-nearest neighbour (k-NN) was combined with an artificial neural network for short-term solar PV forecasts. The model showed an improvement in the prediction accuracy over k-NN. A Long-Short Term Model (LSTM) and AdaBoost were combined in Ref. [23] for a short-term solar PV forecast. The hybrid model performed better than the single LSTM model with an RSME of 0.0435 and an MAE of 0.0246. The authors in Ref. [24] also proposed a hybrid deep learning approach based on a convolutional neural network (CNN) and long-short-term memory recurrent neural network (LSTM) for forecasting PV output power.

While many studies have been conducted on using machine learning to forecast solar PV output, most have focused on specific time horizons such as an hour ahead, a day ahead and a month ahead. Studies in the literature also indicate that artificial neural networks and hybrid models improve prediction accuracy. However, only a few studies exist on how their performance compares with single models over different time horizons.

Therefore, this study seeks to fill this gap in the literature by comparing the performance of single, ensembles and hybrid machine learning models in predicting solar PV output power over four different time horizons. The study further analyses the trend in solar PV production and how it is influenced by ambient and module temperature. This study will contribute to selecting suitable machine learning forecast models and improve understanding of the effect of temperature on solar PV production.

Table 1
Characteristics of the solar PV module.

Module Type	JKM390M-72L-V	
	STC NOCT	
Maximum Power (Pmax)	390Wp	294Wp
Maximum Power Voltage (Vmp)	41.1 V	39.1 V
Maximum Power Current (Imp)	9.49A	7.54A
Open-circuit Voltage (Voc)	49.3 V	48.0 V
Short-circuit Current (Isc)	10.12A	8.02A
Module Efficiency STC (%)	19.67%	
Operating Temperature (°C)	−40 °C~+85 °C	
Maximum System Voltage	1500VDC (UL)/1500VDC (IEC)	
Maximum Series Fuse Rating	20A	
Power Tolerance	0~+3%	
Temperature Coefficients of Pmax	−0.37%/°C	
Temperature Coefficients of V_{oc}	−0.28%/°C	
Temperature Coefficient of I_{sc}	0.048%/°C	
Nominal Operating Cell Temperature (NOCT)	45 ± 2 °C	

Table 2
Descriptive statistics of dataset.

	PV production (kWh)	Irradiation (W/m2)	Ambient temperature (°C)	Module temperature (°C)
count	210235	208649	208648	208648
mean	1.67	184.61	26.05	32.50
std	2.47	263.99	2.73	9.10
min	0	0	20	19
25%	0	0	24	26
50%	0	7.03	25	28
75%	3.04	332	28	39
max	16.98	1344.11	54	72

2. Materials and methods

Fig. 1 depicts the research approach of the study. The data analysis, model training and evaluation were performed using Python 3.10 in Jupyter Notebook 6.5.4.

2.1. Description of the solar PV plant

The 180 kWp solar PV plant chosen for this study is an industrial fuel facility located at Tema in Ghana. The site's geographical location is at latitude 5.7348° N and longitude 0.0302° E in Ghana, where the average annual solar irradiation is 5.1 kWh/m^2/day [25]. The solar PV plant comprises 462 Mono Perc Diamond cell solar modules, each rated at 390 W. The manufacturer of the module is Jinko Solar. The setup has 27 strings of 17 modules per string. However, 3 other strings have 18 modules per string. The strings are connected in three groups to one of the eight 20 kVA- 3 phase Fronius SYMO inverters with 97.9% efficiency. The characteristics of the solar PV module are presented in Table 1.

2.2. Data collection and analysis

The data is collected from the web application of Fronius Solar, the manufacturer of the smart inverters used for this solar PV system. The data which is not publicly available was extracted from the company's webpage [26] with permission from Tino Solutions Ltd, a Ghanaian solar PV company based in Accra.

The PV production, irradiation, module temperature, and ambient temperature data were selected from April 1st' 2021, to March 31st' 2023, in 5-min timesteps. A similar data for the month of April 2023 was further extracted for model evaluation. Table 2 presents the descriptive statistics of the dataset.

2.3. Data cleaning

A total of 1590 data points containing missing inputs for any of the variables were excluded from the model training. The presence of outliers was checked and excluded from the training. All data points recording more than 10 W/m^2 solar irradiation without solar PV production were excluded from the model training. Data points with less than 1 kWh for more than 250 W/m^2 solar radiation were also removed. These assumptions were made based on descriptive statistics where a mean of 184.61 W/m^2 radiation produces 1.67 kWh of solar PV energy. Other isolated cases of outliers were neglected since their numbers were too small compared to the overall dataset.

2.4. Correlation and feature selection

Before model training, a Pearson correlation heatmap was generated to determine how the selected variables explain the changes in solar PV production. More importantly, this technique helps to detect multicollinearity among the explainable variables. The correlation formula is shown in equation (1) [18].

$$r = \frac{\sum_{i=1}^{n} (x_1 - \overline{x}\ )(p_i - \overline{p_i})}{\sqrt{\sum_{i=i}^{n} (x_i - \overline{x_i}) \sum_{i=i}^{n} (p_i - \overline{p_i})^2}} \tag{1}$$

Where x_i is the target value, $\overline{x}_i$ is the mean of the target value, p_i is the predicted value, $\overline{p_i}$ is the mean of the predicted value, and n is the number of datasets.

2.5. Data preprocessing

The date and time were engineered using Label Encoder into year, month, day, hour and minute components. The input data for the Artificial Neural Network was transformed using the 'MinMaxScalar' from the 'Sklearnpreprocessing' module. The data was then split into 80% training and 20% testing to train and evaluate the machine learning models.

2.6. Model training

The cleaned and pre-processed data were used as inputs to train the Linear Regression model, Ridge Regression, Lasso Regression, k-nearest neighbour, Random Forest, AdaBoost, XGBoost and Artificial Neural Network. The input parameters used for the training were solar irradiation, ambient temperature, module temperature, year, month, day, hour and minutes of the dataset.

The three best-performing models were stacked to generate hybrid models to improve the prediction performance. The prediction performances of the models were compared using coefficient of determination (R^2 score), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), and the overall best model was selected for the final predictions.

2.6.1. Linear regression

The linear regression model makes predictions based on a linear function relating the independent variable to the feature variables of the data. The goal of machine learning when training a linear regression model is to learn from the data to generate the parameters for the linear function that accurately predicts the output variable for each input value. The best-fit model is achieved by error minimization to reduce the variance between the predicted values and the actual values of the output variables to the barest minimum. The relationship between the dependent variable Y and the independent variable X is determined by equation (2) [12].

$$\mathrm{Y} = \beta_o + \beta_1 \mathrm{X} + \epsilon \tag{2}$$

where:

$\boldsymbol{\beta}_0$ is a constant (the point where it intersects the y-axis),

$\boldsymbol{\beta}_1$ is the regression coefficient (the slope of the regression line),

$\in$ is the error term to minimize for best-fit model.

Real-life regression problems involve several predictive variables (x_1, x_2, x_3, …, x_n). This, in a number of cases, results in multicollinearity, where some of the independent variables linearly depend on themselves. This can introduce distortions into the regression model. Instances also exist where the presence of less relevant feature variables in the dataset results in overfitting of the model. These anomalies in the regression model are addressed by introducing a regularization term to shrink the parameters of the colinear and less relevant feature variables to zero or a lower value. This regularization technique is achieved by performing techniques such as LASSO Regression or Ridge Regression.

2.6.2. k-Nearest Neighbour (KNN) regression

The k-nearest neighbours (k-NN) regressor is a non-parametric supervised machine learning approach for forecasting continuous numerical values. In k-nearest neighbour regression, the algorithm forecasts the target variable of a new data point by considering the values of its k-nearest neighbours in the training dataset [27]. These neighbours are chosen based on their proximity to the newly added data point [28].

Without assuming specific data distribution, k-NN captures complex relationships in the data and makes predictions based on the values of the nearest data points. The value of the predicted data point is the average value of the k-nearest neighbours [22]. The choice of k, the number of nearest neighbours considered for the prediction, determines the bias-variance trade-off. A more flexible model with lower bias with higher variance is obtained when the value of k is small, while larger k values result in smoother predictions with higher bias but lower variance. For this study, the 3 nearest neighbours were chosen.

2.6.3. Decision tree regression

The decision tree is a machine learning algorithm for predicting continuous numerical values. It functions by building a model that resembles a tree and partitions the data recursively according to the values of input features. The tree's leaf nodes reflect the projected output values, whereas each interior node represents a choice based on a feature. The best feature and split point optimizing the reduction in variance or mean squared error are chosen as part of the decision tree construction process. As a result, partitions are made that reduce prediction error and guarantee the tree includes the most informative features [18]. For this study, the default values of 2 and 1 were selected for the minimum split node and minimum leaf sample. The maximum leaf node was not explicitly set to allow the tree to keep growing until no further split improves the model's performance.

2.6.4. Random Forest regression

Random Forest uses an ensemble of decision trees as base learners to make predictions. Through bootstrap sampling, the trees are randomly selected for training. The randomly selected trees' prediction average is calculated as the ensemble model's predicted value. For this reason, Random Forest Regressor gives marginally different values for each round of training. As a result, the random state is set to a given value to have repeated results for the same input parameters. Equation (3) presents the underlying equation of this ensemble model.

$$y(x) = \frac{1}{N}\sum_{i=i}^{n} h_i(x) \tag{3}$$

where N is the total number of decision trees in the ensemble and h_i is the ith base learner [29].

Random Forest algorithm, however, incorporates randomness and ensemble techniques to reduce overfitting and improve the model's predictive power as compared to individual decision trees. In this study, 200 decision trees (n estimators = 200) were used and the trees were allowed to grow until they reach their maximum depth potential.

2.6.5. XGBoost and AdaBoost

XGBoost combines the predictions of weak decision trees using an optimised objective function to generate a strong predictive model. XGBoost minimises the error function of the weak decision trees by employing a gradient boosting technique and iteratively adding decision trees that focus on reducing the errors of previous trees.

The underlying mathematical equation of XGBoost is given in equation (4).

$$y_i = y_o + \eta \sum_{k=1}^{n} f_k(U_i) \tag{4}$$

Where: y_i is the predicted value of the parameter vector Ui, y_o is the average of the parameters in the training data, n denotes the number of estimators, and η represents the learning rate of the model [29].

XGboost further introduces a regularization term to reduce the complexity of the model to prevent overfitting. A total number of 200 decision trees were used for this model with a default depth of 6.

While XGboost sequentially adds decision trees to the ensemble model to correct the errors of the previous model, AdaBoost applies a weighting technique in iteration to each weak decision tree added to the ensemble. The weighting is assigned based on the performance of each decision tree in predicting the desired outcome. The weight is then adjusted in each iteration to give more emphasis to the data points that were poorly predicted. This technique seeks to improve the overall prediction of the ensemble of weak models [7].

During model training, a weighted voting system based on the performance of each decision tree is used to generate the final prediction. The final prediction is then obtained by the mean of the weighted prediction of all the decision trees.

2.6.6. Artificial Neural Network (ANN)

An Artificial Neural Network is a biology-inspired machine learning algorithm that learns complex relationships in datasets to make predictions. ANN consists of interconnected nodes called neurons. These neurons are arranged into input layers, single or multiple hidden layers and an output layer. Each neuron receives input signals, assigns weights, and computes the weighted sum using an activation function. Based on the value of the weighted sum, the activation function chooses the neuron's output. The number of feature variables determines the number of neurons on the input layer.

Similarly, the number of output variables corresponds to the number of neurons in the output layer. The underlying equation that governs the computations performed by the neurons is presented in equation (5), and the activation function is shown in equation (6) [29].

$$y = \sum_{i=1}^{n} (w_i x_i) + b \tag{5}$$

$$F(y) = f\left(\sum_{i=1}^{n} (w_i x_i) + b\right) \tag{6}$$

ANN performs several rounds of learning to improve the outcome of the predicted value. Each prediction phase is followed by backpropagation to modify the weight and biases of the connections and nodes to minimize the error difference between the predicted

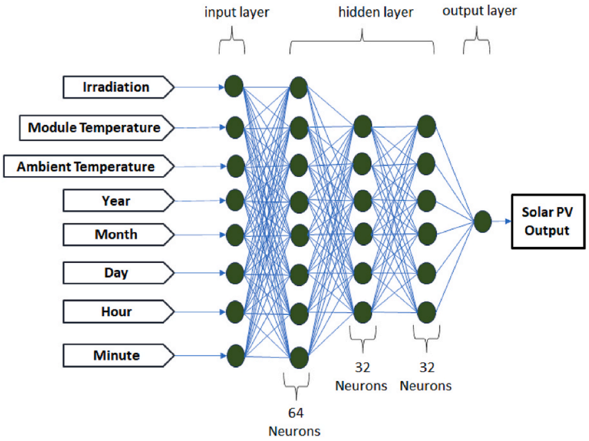


Fig. 2. Schematic diagram of the ANN.

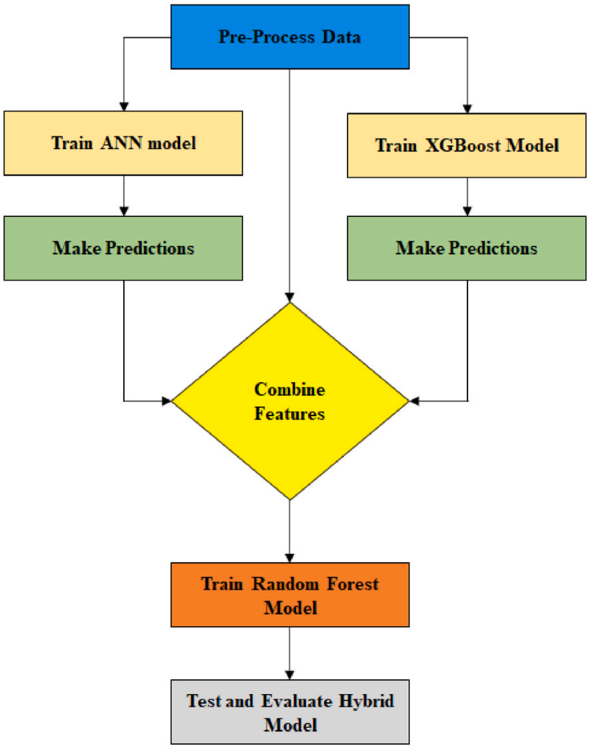


Fig. 3. Process flow of the hybrid models.

output and the desired value. One complete iteration of this process is known as an epoch, and several epochs are performed during a typical ANN training until the error stops reducing. An early stopping function terminates the learning process when the error function stops decreasing. The chosen ANN for this study has one input layer with eight neurons. This is followed by one hidden layer with 64 neurons, succeeded by two more hidden layers with 32 neurons each. The output layer finally generates the predicted value. The Rectified Linear Unit (ReLU) activated and trained the model using TensorFlow's Adam optimiser for 100 epochs. Fig. 2 presents the schematic diagram of the ANN used in this study.

2.6.7. Hybrid models

The stacking approach was used to build the hybrid machine-learning models for this study. The three hybrid models developed in this study are XGBoost-RandomForest Regressor, ANN-RandomForest Regressor and ANN-XGBoost-RandomForest Regressor. This was intended to improve the performance of the individual models.

The hybrid models in this study were built by combining the original features of the dataset with the predictions of ANN and

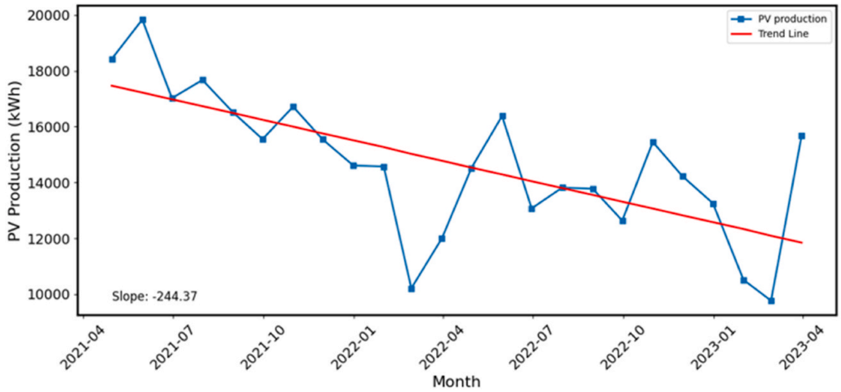


Fig. 4. Monthly Solar PV Production over 2 years.

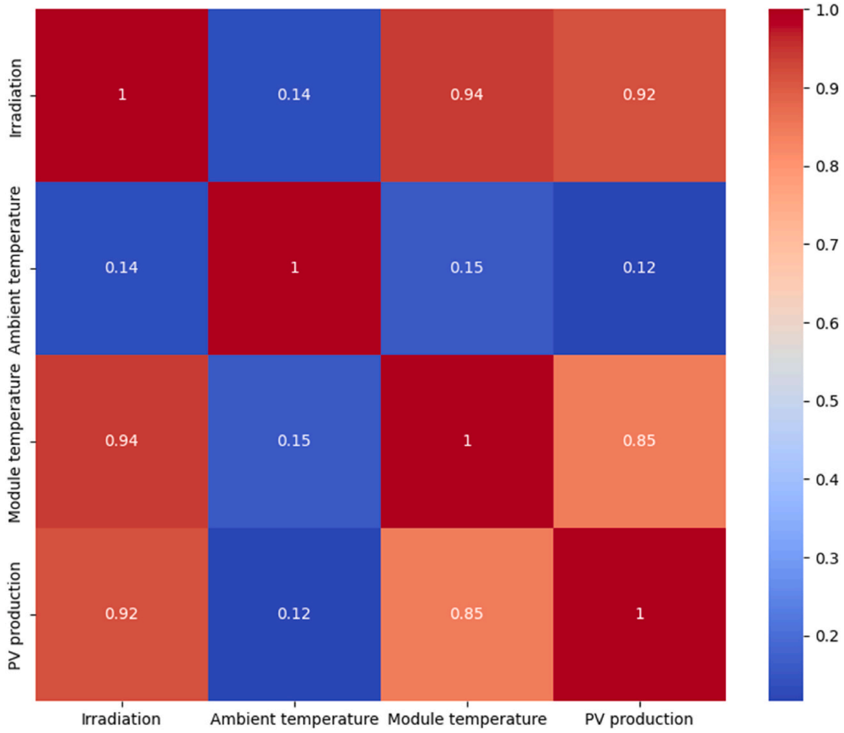


Fig. 5. Correlation heatmap.

XGBoost as input for the Random Forest Regressor. The parameters of ANN and XGBoost were maintained in the hybrid models just as during their model training. Fig. 3 presents the process flowchart for building the ANN-XGBoost-Random Forest hybrid model. The ANN-XGBoost and ANN-Random Forest models were built similarly but using only two individual models.

2.7. Evaluation of the models

The evaluation metrics used to evaluate the performance of the machine learning models in this study are Coefficient of Determination (R^2 score), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The value of R^2 was computed using equation (7).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left| y_{\text{forecasted}} - y_{\text{observed}} \right|}{\sum_{i=1}^{n} \left| y_{\text{forecasted}} - y_{\text{mean}} \right|} \tag{7}$$

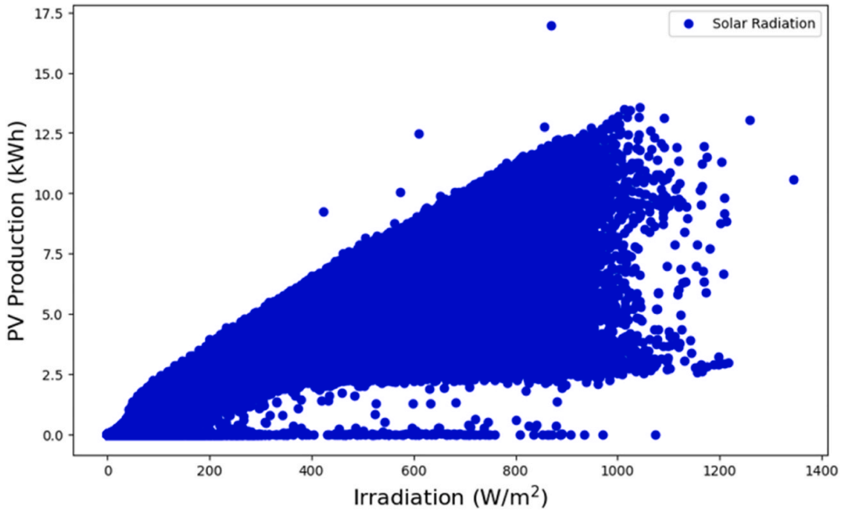


Fig. 6. Effect of Irradiation on solar PV production.

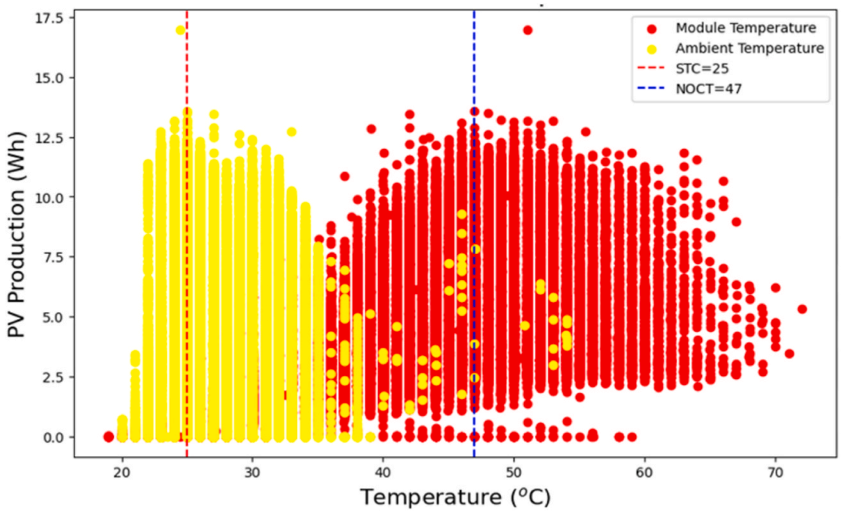
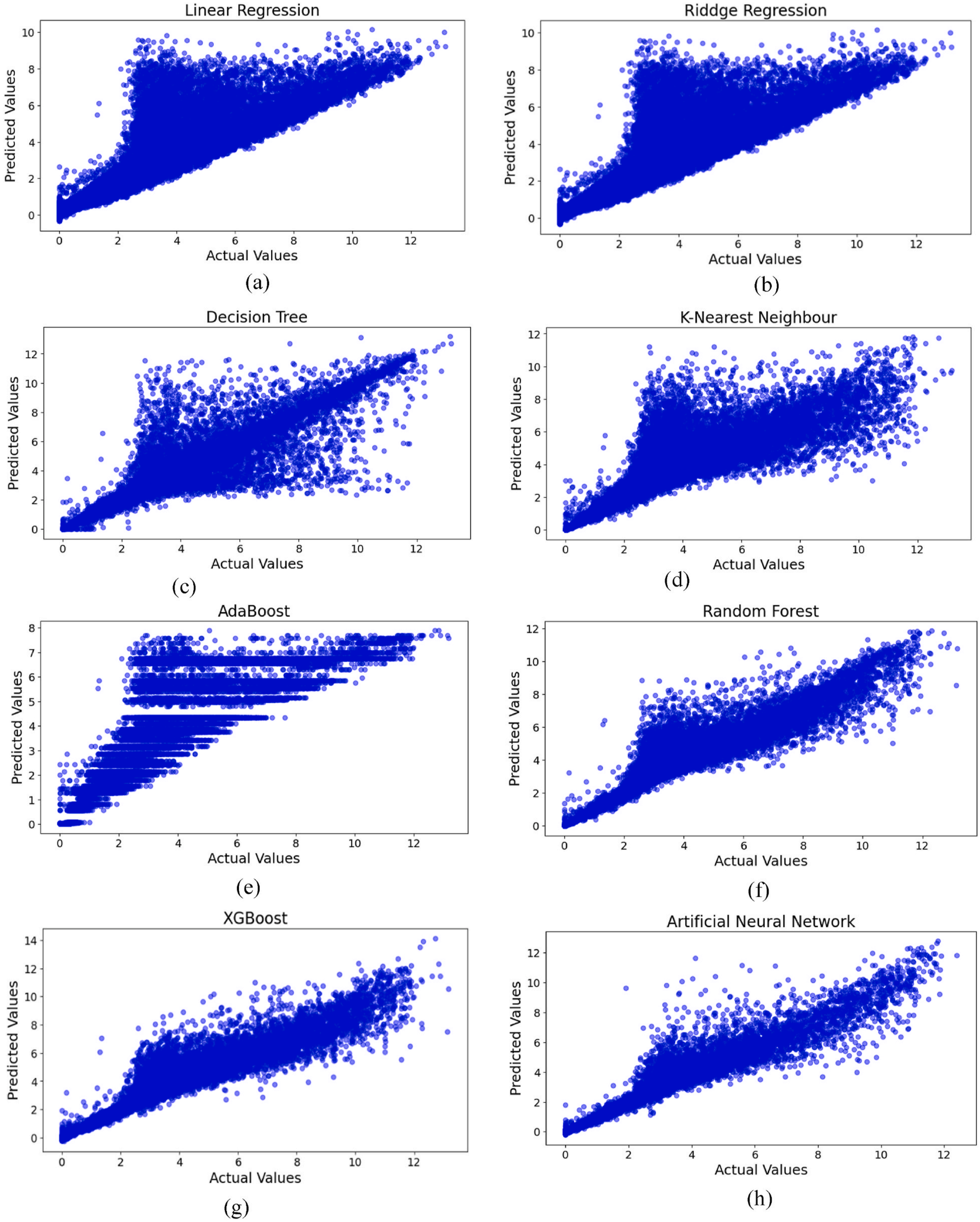


Fig. 7. Effect of Temperature on solar PV output.

Table 3
Performance of the machine learning models.

No.	Model (Regressors)	Evaluation Metrics		
		R^2 Score	MAE	RMSE
1	Ridge Regression	0.8376	0.5381	0.9999
2	Linear Regression	0.8443	0.5535	0.9792
3	Lasso Regression	0.8443	0.5535	0.9792
4	AdaBoost	0.8605	0.4545	0.9267
5	K-Nearest Neighbhour	0.8737	0.3678	0.8820
6	Decision Trees	0.9076	0.2492	0.7542
7	Random Forest	0.9525	0.2115	0.5410
8	XGBoost	0.9529	0.2311	0.5385
9	Artificial Neural Network (ANN)	0.9569	0.2246	0.5180



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 8. Regression plot of predicted and actual values using (a) Linear Regression (b) Riddge Regression (c) Decision Tree (d) K-Nearest Neighbour (e) AdaBoost (f) Random Forest (g) XGBoost and (h) Artificial Neural Network.

Table 4
Performance of Hybrid Machine Learning Models after training.

No.		Evaluation Metrics		
	Hybrid Model (Regressors)	R^2 Score	MAE	RMSE
1	ANN-Random Forest	0.9580	0.2010	0.5096
2	XGBoost-Random Forest	0.9608	0.1830	0.4916
3	ANN-XGBoost-Random Forest	0.9659	0.1682	0.4579

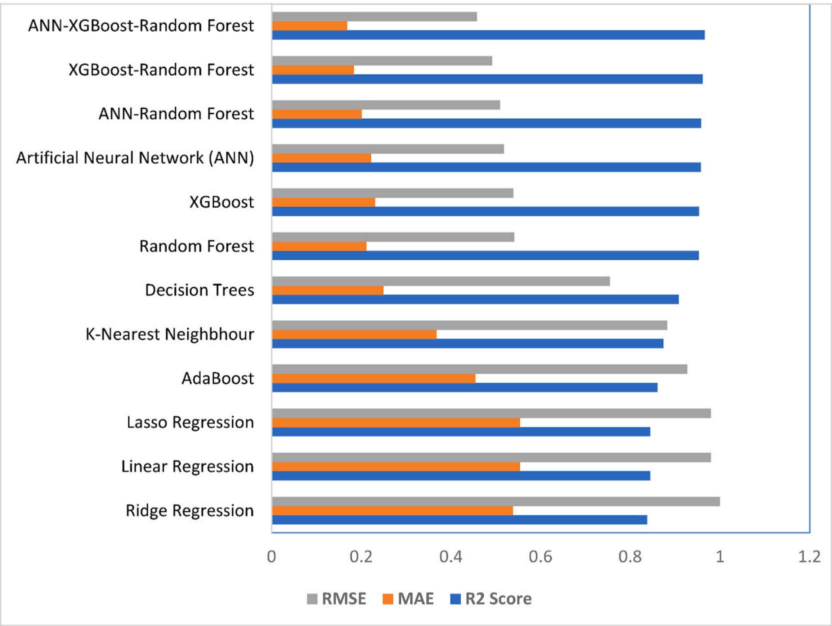


Fig. 9. Comparison of the training performance of hybrid and individual models.

Table 5
A day ahead and a week ahead forecast evaluation result.

		Evaluation Metrics					
		A Day Ahead Forecast			A Week Ahead Forecast		
No.	Model	R^2 Score	MAE	RMSE	R^2 Score	MAE	RMSE
1	Random Forest	-0.1450	1.5709	1.5897	0.8378	0.5505	1.1415
2	XGBoost	-0.3292	0.9543	1.3233	0.8548	0.5014	1.0616
3	ANN	**0.8702**	**0.3043**	**0.7477**	0.5311	0.9225	1.9411
4	ANN-Random Forest	0.3242	0.6944	0.7307	0.5819	0.8853	1.8328
5	XGBoost-Random Forest	-0.2737	0.8851	1.3039	**0.8556**	**0.4669**	**1.0771**
6	ANN-XGBoost-Random Forest	0.0473	0.7389	1.2875	0.8273	0.567	1.1781

The MAE was generated using equation (8) [30].

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_{\mathrm{forecasted}} - y_{\mathrm{observed}}\right| \tag{8}$$

And the RMSE was obtained using equation (9) [31].

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{\mathrm{forecasted}} - y_{\mathrm{observed}}\right)^2} \tag{9}$$

The hybrid models and their individual models were finally used to forecast the solar PV production for the next day, the next week,

Table 6
Two weeks ahead and a month ahead forecast evaluation result.

		Evaluation Metrics					
		Two Weeks Ahead Forecast			A Month Ahead Forecast		
No.	Model	R^2Score	MAE	RMSE	R^2 Score	MAE	RMSE
1	Random Forest	**0.8429**	**0.5029**	**1.0551**	**0.7681**	**0.5384**	**1.2319**
2	XGBoost	0.8338	0.5089	1.0852	0.7570	0.5339	1.2612
3	ANN	0.6062	0.7966	1.6703	0.6787	0.6780	1.4500
4	ANN-Random Forest	0.6303	0.7898	1.6185	0.7241	0.6699	1.3437
5	XGBoost-Random Forest	0.8300	0.4793	1.0976	0.7541	0.5095	1.2686
6	ANN-XGBoost-Random Forest	0.8339	0.5110	1.0847	0.74459	0.5179	1.2928

the next two weeks and the next month to evaluate their performance on actual predictions. The best performing model for each time horizon was finally selected using the evaluation metrics.

3. Results and discussion

3.1. Solar PV production trend

Fig. 4 shows a declining trend of the solar PV output of the 2-year period after its installation. A negative slope regression line confirms the theory of the degradation of solar PV system over time [32]. The estimated slope of the trend line of −244.37 reveals that the output production of the solar PV system reduces steadily by an average of 244.37 kWh per month. This effect can be attributed to factors such as soiling, dust, and the self-degrading effect of the solar PV modules [33].

3.2. Effect of the predictive variables on the solar PV output

As depicted in Fig. 5, the Pearson correlation matrix shows that all the chosen predictive variables positively correlate with solar PV Production. Solar irradiation, however, has the highest effect on the output power of the solar PV system, explaining 92% of the variation in the output power. Fig. 6 further shows that this relationship between solar irradiation and solar PV production is quasi-linear and positive.

The solar PV production also shows a significant dependency on the module temperature with a correlation coefficient (r) of 0.85. Though this coefficient of correlation is positive, Fig. 7 shows that solar PV production increases steadily with an increase in module temperature until it reaches a peak value. Then, a further increase in module temperature results in a decline in PV production. Fig. 7 shows this peak value to be 47 °C. This corresponds to the solar PV module's Nominal Operating Cell Temperature (NOCT) as specified in Table 2.

Finally, the ambient temperature shows the least correlation with solar PV production, explaining only 12% of the changes in solar PV production. However, like the module temperature, solar PV production increases marginally with a rise in the ambient temperature until it peaks at around 25°C. Afterwards, it begins to decline with a further increase in the ambient temperature. This peak temperature of 25 °C is the temperature at the Standard Test Condition (STC) [34]. This confirms that solar PV systems' optimal ambient temperature is 25 °C

3.3. Performance of machine learning models

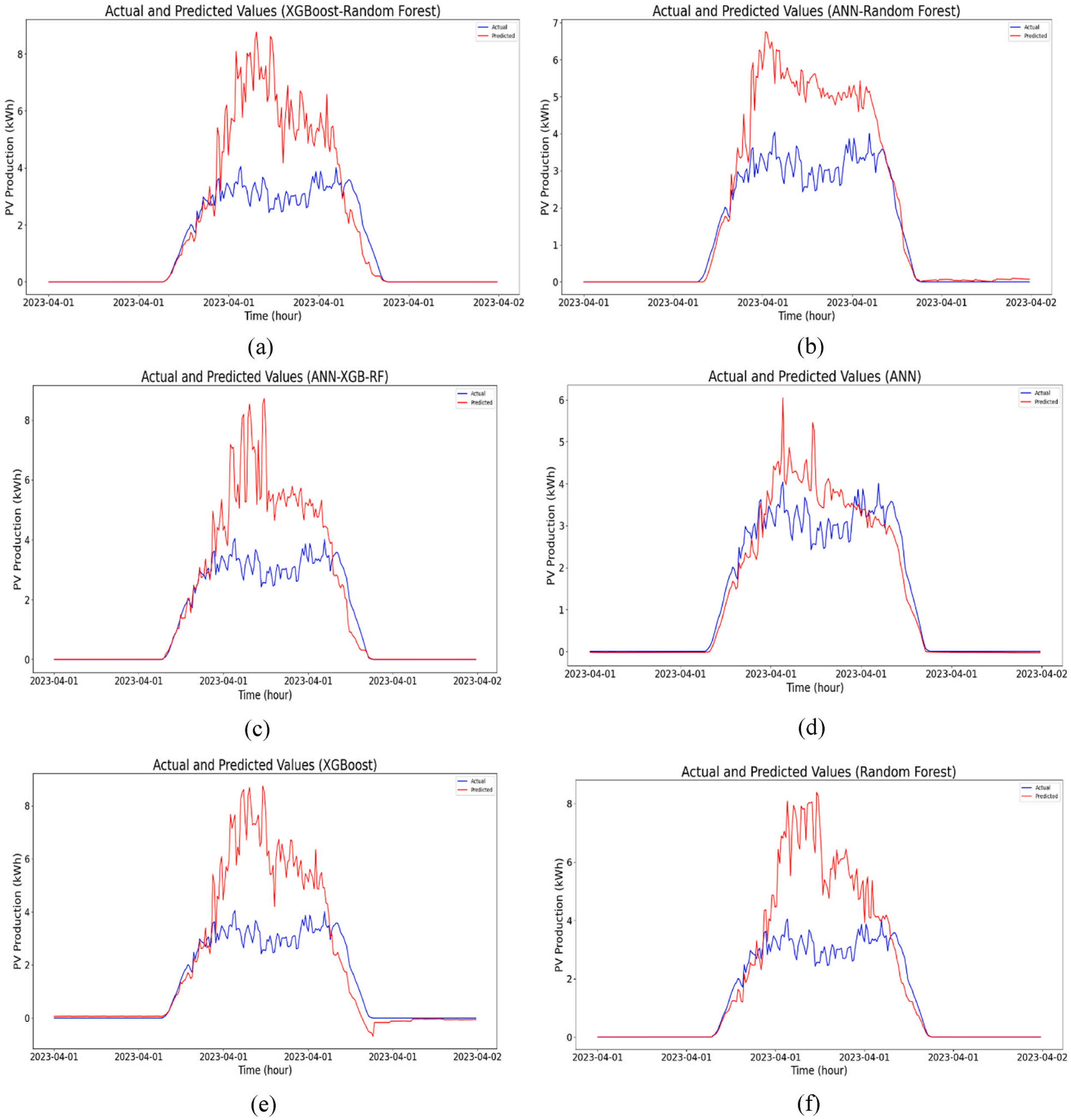
Table 3 displays the results of the training performance of the machine learning models. The results show that Random Forest, XGBoost and Artificial Neural Network outperform all the other models during the training phase. Artificial Neural Network emerged as the best performing model on the training phase. Random Forest however, performed better than Artificial Neural Network and XGBoost on the Mean Absolute Error (MAE) metrics, but when all metrics were considered, Artificial Neural Network demonstrated a superior performance. Linear Regression and its regularization models; Ridge and Lasso Regression showed a comparatively low performance at the training stage.

The regression plots of the actual solar PV production and the predicted values of the different machine learning models are shown in Fig. 8.

The regression plots reveal that all the models show significant errors in predicting solar PV production. Given the mean value of the solar PV production from Table 2 to be 1.67 kWh, the MAE of ANN of 0.2246 kWh represents an average error of 13.45% for every predicted value. XGBoost's predictions deviate by an average of 13.83%, and Random Forest records 12.66%. On the Mean Absolute Error metric, Random Forest showed the slightest error. Using the MAE as the benchmark, hybrid models are created to further reduce these observed error margins.

3.3.1. Performance of hybrid models

Table 4 shows the performance of the three hybrid models. The models after training were used to predict the solar PV production



(a) (b) (c) (d) (e) (f)

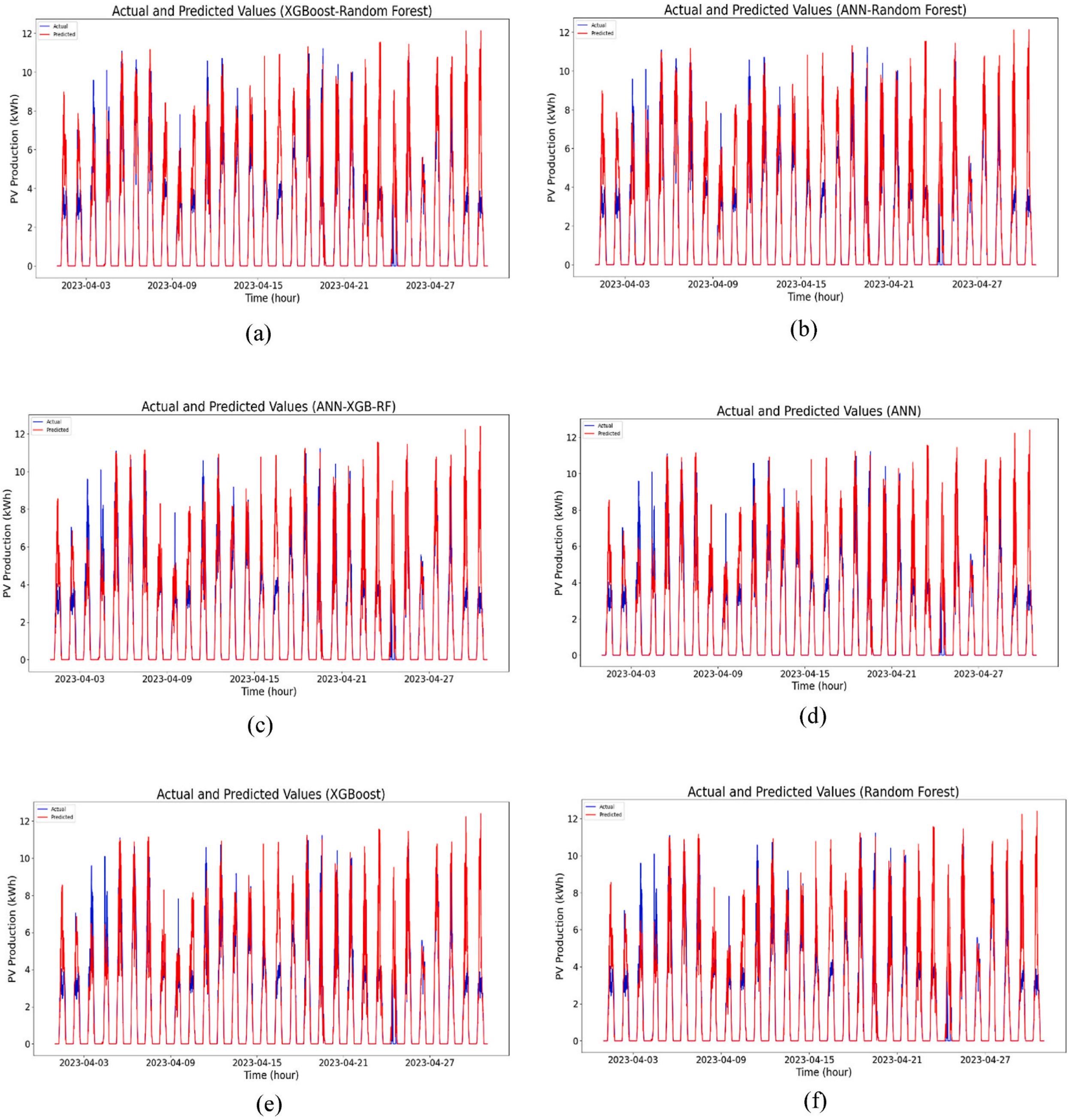
Fig. 10. Day ahead solar PV forecast using (a) XGBoost-Random Forest (b) ANN-Random Forest (c) ANN-XGB-RF (d) ANN (e) XGBoost (f) Random Forest.

for the month of April 2023, and their performances were evaluated using the same evaluation metrics. The first observation was that all the hybrid models performed better than the individual models as presented in Table 4. This shows that combining two or more machine learning models can improve their forecasting performance.

Among the three hybrid models, the stacking of Artificial Neural Network, XGBoost and Random Forest (ANN-XGB-RF) gave the best forecast performance. With its MAE of 0.1682, the average percentage forecast error of 12.66% obtained by training the individual models has been reduced to 10.10%. Fig. 9 compares the evaluation scores of the three hybrid models with the individual models.

3.3.2. *Final prediction and evaluation*

The performance of the hybrid and individual models is evaluated by making predictions on unseen datasets for different forecasting horizons. The predictions by the selected six models are made for a day-ahead, a week-ahead, two weeks-ahead, and finally, a



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 11. A month ahead solar PV forecast using (a) XGBoost-Random Forest (b) ANN-Random Forest (c) ANN-XGB-RF (d) ANN (e) XGBoost (f) Random Forest.

month-ahead. Tables 5 and 6 present the evaluation results of the hybrid models and ANN, XGBoost, and Random Forest models for the four selected forecasting horizons.

The results clearly show that all the models performed lower at the validation phase. However, satisfactory performance was recorded by some of the models. Artificial Neural Network performed better than all the other models with an R^2 score of 0.8702, RMSE of 0.5352 and an MAE of 0.3043 for the day ahead prediction. However, ANN did not perform well on the long-term predictions. This study shows that for this kind of dataset, the ANN model is more suitable for short-term forecasts than long-term forecasts.

Further, the hybrid models that performed best during the training phase failed to perform in the final prediction. This suggests that stacking models to form hybrids does not necessarily improve the actual prediction accuracy of the model in the short term.

The results further reveal that, though XGBoost and Random Forest did not perform better on short-term predictions, they did comparatively well on long-term predictions than ANN. Table 4 shows that XGBoost-Random Forest hybrid model performed better than all the other models on the one-week ahead prediction with an R^2 score of 0.8556 and MAE and RMSE of 0.4669 and 1.0771, respectively. The finding further shows that the hybrid model XGBoost-Random Forest improved the performance of its individual models' performance for the week ahead prediction.

Finally, the two-weeks and one-month predictions were dominated by Random Forest with R^2 scores as shown in Table 6. Following closely with a comparable performance was XGBoost and its Random Forest Ensemble. This further proves that Random Forest and XGboost perform better than Artificial Neural Network on long-term predictions.

Figs. 10 and 11 show the final predictions for the day ahead (April 1) and a month ahead (April 1–30) of solar PV production.

Figs. 10 and 11 show that Artificial Neural Network produced the best day-ahead prediction of solar PV production, and Random Forest produced the best performance for the month-ahead prediction. However, the models did not achieve perfect prediction, and significant errors could be observed. This requires further studies to improve upon the model's performance.

4. Conclusion

In this paper, a 180 kWp case study solar PV plant was used for the comparative assessment of different machine learning models in predicting solar PV production a day ahead, a week ahead, two weeks ahead and one month ahead. Hybrid models composed of Artificial Neural Network, Random Forest and XGBoost were developed to improve the forecasting performance of the models. The trend in solar PV production was also analysed alongside the effect of temperature on solar PV output power.

Observations indicated that solar irradiation, ambient temperature and module temperature can be used to predict solar PV production and that solar irradiation and module temperature have the most significant effect on the output power of the solar PV systems. Solar PV production decreased by an average of 244.37 kWh per month over the two years. It was also observed that solar PV production declines when the module temperature exceeds 47 $^\circ C$ (NOCT) and when the ambient temperature rises beyond 25 $^\circ C$ (the temperature at STC).

Furthermore, the study showed Artificial Neural Network gave the best performance for the day-ahead prediction with an R^2 Score of 0.8702 and an MAE and RMSE of 0.5352 and 0.3043, respectively. However, for the long-term solar PV forecast up to a month, Random Forest performed better with an R^2 score of 0.7681 and an MAE and RMSE of 0.5384 and 1.2319, respectively. The results show that the general performance of the models declines with increasing time horizons. The study also concludes that machine learning models perform differently over different time horizons.

The forecasts in this study showed significant errors, and further studies should target minimising the error. The study is also limited in scope as it focused on a single solar PV setup in Ghana. It is possible the models could perform differently on different dataset. Further research could incorporate spatial components into the models and explore their performance at different locations. Additionally, this study used irradiation, temperature, and time as the predictive variables. Future study could include other weather variables such as wind speed and relative humidity to improve the performance of the models.

Data availability statement

The authors confirm that the data supporting the findings of this study are available within the article.

CRediT authorship contribution statement

Shadrack T. Asiedu: Writing – original draft, Data curation. **Frank K.A. Nyarko:** Writing – review & editing, Conceptualization. **Samuel Boahen:** Writing – review & editing, Supervision. **Francis B. Effah:** Writing – review & editing, Formal analysis. **Benjamin A. Asaaga:** Writing – review & editing, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Ahmed, V. Sreeram, Y. Mishra, M.D. Arif, A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization, Renew. Sustain. Energy Rev. 124 (2020) 109792.

[2] A.S. Aziz, et al., Design and optimization of a grid-connected solar energy system: study in Iraq, Sustain. Times 14 (13) (2022) 1–29.

[3] A.M. Attia, A. Al Hanbali, H.H. Saleh, O.G. Alsawafy, A.M. Ghaithan, A. Mohammed, A multi-objective optimization model for sizing decisions of a grid-connected photovoltaic system, Energy 229 (2021) 120730.

[4] G.M. Shafiullah, A.M.T. Oo, A.B.M.S. Ali, P. Wolfs, A. Stojcevski, Experimental and simulation study of the impact of increased photovoltaic integration with the grid, J. Renew. Sustain. Energy 6 (3) (2014) 033144.

[5] V. Kushwaha, S. Member, N.M. Pindoriya, S. Member, Very short-term solar PV generation forecast using SARIMA model : A Case Study, in: 7th Int. Conf. Power Syst, Pune, India, 2017, pp. 430–435.

[6] E. Roumpakias, T. Stamatelos, Prediction of a grid-connected photovoltaic park's output with artificial neural networks trained by actual performance data, Appl. Sci. 12 (13) (2022) 6458.

[7] S.M. Babbar, C.Y. Lau, K.F. Thang, Long term solar power generation prediction using adaboost as a hybrid of linear and non-linear machine learning model, Int. J. Adv. Comput. Sci. Appl. 12 (11) (2021) 536–545.

[8] K. Opoku, S. Lucemo, Q. Zhou Sun, A. Dimitrovski, A bayesian approach to probabilistic solar irradiance forecasting, North American Power Symp, Salt Lake City, UT, USA (2023) 1–6.

[9] M. Al-Alddous, Z. Dalala, C.B. Class, F. Alawneh, H. Al-Taani, Performance analysis of off-grid PV systems in the Jordan valley, Renew. Energy 113 (2017) 930–941.

[10] R.A. Gupta, A. Bansal, K. Roy, Solar energy prediction using decision tree regressor, in: 5th Int. Conf. Intell. Comput. Control Syst. ICICCS, Madurai, India, 2021, pp. 489–495.

[11] L. Alhmoud, A.M. Al-Zoubi, I. Aljarah, Solar PV power forecasting at Yarmouk University using machine learning techniques, Open Eng. 12 (1) (2022) 1078–1088.

[12] M.Y. Erten, H. Aydilek, Solar power prediction using regression models, Int. Journal of Eng. Research and Devt. 14 (3) (2022) 333–342.

[13] C. Scott, M. Ahsan, A. Albarbar, Machine learning for forecasting a photovoltaic (PV) generation system, Energy 278 (2022) 127807.

[14] C.F. Yen, H.Y. Hsieh, K.W. Su, M.C. Yu, J.S. Leu, Solar power prediction via support vector machine and random forest, E3S Web Conf. 69 (2018) 01004.

[15] F. Mahia, A.R. Dey, A. Masud, M.S. Mahmud, Forecasting electricity consumption using ARIMA model, Int. Conf. Sustain. Technol. Ind., Dhaka, Bangladesh 4 (2019) 1–6.

[16] D.J. Bae, B.S. Kwon, K. Bin Song, XGboost-based day-ahead load forecasting algorithm considering behind-the-meter solar PV generation, Energies 15 (1) (2022) 128.

[17] Q.T. Phan, Y.K. Wu, Q.D. Phan, Short-term Solar power forecasting using XGBoost with numerical weather prediction, in: IEEE Int. Futur. Energy Electron. Conf. IFEEC, 2021, pp. 1–6. Taipei, Taiwan.

[18] Y. Essam, A.N. Ahmed, R. Ramli, K.W. Chau, M.S.I. Ibrahim, M. Sherif, A. Sefelnasr, A. El-Shafie, Investigating photovoltaic solar power output forecasting using machine learning algorithms, Eng. Appl. Comput. Fluid Mech. 16 (1) (2022) 2002–2034.

[19] O. Yadav, R. Kannan, S.T. Meraj, A. Masaoud, Machine learning based prediction of output PV power in India and Malaysia with the use of statistical regression,", Math. Probl Eng. (2022) 5680635.

[20] A. El Kounni, H. Radoine, H. Mastouri, H. Bahi, A. Outzourhit, Solar power output forecasting using artificial neural network, in: 9th Int. Renew. Sustain. Energy Conf. IRSEC, Morocco, 2021, pp. 1–7.

[21] M. Ding, L. Wang, R. Bi, An ANN-based approach for forecasting the power output of photovoltaic system, Procedia Environ. Sci. 11 (2011) 1308–1315.

[22] C.R. Chen, U.T. Kartini, K-nearest neighbor neural network models for very short-term global solar irradiance forecasting based on meteorological data, Energies 10 (2) (2017) 186.

[23] F. Kyeremeh, F. Zhi, Y. Yi, E. Gyamfi, I.K. Nti, Solar PV power forecasting with a hybrid LSTM-AdaBoost ensemble, IEEE/IET Int. Util. Conf, Greater Accra, Ghana (2022) 1–7.

[24] G. Li, S. Xie, B. Wang, J. Xin, Y. Li, S. Du, Photovoltaic power forecasting with a hybrid deep learning approach, IEEE Access 8 (2020) 175871–175880.

[25] B. Aboagye, S. Gyamfi, E.A. Ofosu, S. Djordjevic, Status of renewable energy resources for electricity supply in Ghana, Sci. African 11 (2021) e00660.

[26] Fronius Solar Web, https://www.solarweb.com, accessed April. 7, 2023.

[27] Z. Liu, Z. Zhang, Solar forecasting by K-Nearest Neighbors method with weather classification and physical model, North Am. Power Symp. (2016) 1–6. Denver, CO, USA.

[28] J. Sun, W. Du, N. Shi, A survey of kNN algorithm, Inf. Eng. Appl. Comput. 1 (2018) 1–10.

[29] R. Opoku, G. Mensah, E.A. Adjei, J.B. Dramani, O. Kornyo, R. Nijjhar, M. Addai, D. Marfo, F. Davis, G.Y. Obeng, Machine learning of redundant energy of a solar PV mini-grid system for cooking applications, Sol. Energy 262 (2023) 111790.

[30] Z. Liu, Z. Zhang, Solar forecasting by K-Nearest Neighbors method with weather classification and physical model, NAPS 2016 - 48th North Am, Power Symp. Proc. (2016) 1–6, https://doi.org/10.1109/NAPS.2016.7747859.

[31] M.Z. Mukaram, F. Yusof, Solar radiation forecast using hybrid SARIMA and ANN model, Malaysian J. Fundam. Appl. Sci. 13 (4) (2017) 4, 1.

[32] S. Begum, R. Banu, G.F. Ali Ahammed, B.D. Parameshachari, Rajashekarappa, Performance degradation issues of PV solar power plant, in: Int. Conf. on Electr, Electronics, Com., Comp., and Optm. Techniques (ICEECCOT), Mysuru, India, 2017, pp. 311–313.

[33] M.M. Fouad, L.A. Shihata, E.I. Morgan, An integrated review of factors influencing the performance of photovoltaic panels, Renew. Sustain. Energy Rev. 80 (2017) 1499–1511.

[34] F. Shaik, S.S. Lingala, P. Veeraboina, Effect of various parameters on the performance of solar PV power plant: a review and the experimental study, Sustain. Energy Res. 10 (1) (2023).