Journal of
Renewable Energy Research and Applications (RERA)

Vol. 5, No 2, 2024, 229-241 *DOI: 10.22044/rera.2023.12880.1209*

LSTM and XGBoost Models for 24-hour Ahead Forecast of PV Power from Direct Irradiation

D. Kossoko Babatoundé Audace[1*], A. Richard Gilles[2] and A. Bienvenu Macaire[1]

1. Department of Electrical Engineering Polytechnic School of Abomey-Calavi (EPAC), Abomey-Calavi, Benin.
2. Department of ENSET-Lokossa National University of Science, Technology, Engineering and Mathematics of Abomey (UNSTIM), Abomey, Benin.

Received Date 30 March 2023; Revised Date 07 May 2023; Accepted Date 14 July 2023
**Corresponding author: didavia@gmail.com (D. Kossoko Babatoundé Audace)*

Abstract

In this work, the photo-voltaic power forecast for the next 24 hours by combining a time series forecasting model (LSTM) and a regression model (XGBoost) from direct irradiation only is performed. Several meteorological parameters such as irradiance, ambient temperature, wind speed, relative humidity, sun position, and dew point were identified as the influencing parameters of PV power variability. Thanks to the parameter extraction and selection techniques of the XGBoost model, only the direct irradiation could be kept as the input parameters. The LSTM model was used to predict the direct irradiation for the next 24 hours and the XGBoost model to estimate the future power from the predicted irradiation. These models were developed under Python 3, the exploited data was downloaded in the PVGIS database for the city of Abomey-Calavi in Benin, and the prediction was carried out on a panel of 1000W of peak power. An experimental validation was then performed by comparing the predicted irradiance values with the measured values on site. It was obtained for the LSTM model a root mean square error of 3.66 W/m^2 and for the XGBoost model a root mean square error and a regression coefficient of 1.72 W and 0.992129, respectively. These results were compared with the LSTM-XGBoost performances with irradiation, temperature, sun position, and wind speed as the inputs. It was found that the use of irradiation alone as input did not as such impair the forecast performance. The proposed method was also found to be more efficient than LSTM and CNN models used alone.

Keywords: *PV power forecasting, Direct irradiation, LSTM, XGBoost, Experimental validation.*

1. Introduction

The interest in renewable energy sources (RES) has increased considerably in the recent years. The main reason for this growth is the expected depletion of the world's conventional energy resources (oil, natural gas, coal, and even uranium), whereas RES can be considered inexhaustible on a human scale. Another reason for the boom is the non-uniform distribution of conventional energy sources on the planet, associated with non-uniform consumption. Thus a considerable part of the world's population does not have access to electricity. In sub-Saharan Africa, for example, one in two people, or more than 600 million people, 75% of whom live in rural areas, do not have access to electricity, according to the International Energy Agency's (IEA) Africa Energy Outlook 2019 report [1]. Yet the electrical needs of these people can be met by distributed generation provided by renewable energy systems. A third important reason is the fight against polluting emissions that are the cause of many harmful phenomena such as the greenhouse effect, the stratospheric ozone hole, and global warming. These harmful gases are emitted by various human activities including the production of electricity from conventional energy sources. These harmful gases are emitted by various human activities including the production of electricity from conventional energy sources. A final reason is the availability of renewable energy sources, especially in Africa, where the theoretical potentials of solar photovoltaic and wind energy are 1,449,742 TWh/year and 978,066 TWh/year [2].

Faced with this fact, many governments and regional and international organizations have considered investments in the renewable energy sector as a priority. As a result, many initiatives have been launched to support the development of the energy sector, without their effectiveness being, for the moment, fully satisfactory. Indeed, solar photo-voltaic and wind sources have two major flaws: their intermittency and their random variability. These defects lead to a random variation of the output power of the production systems and make them uncontrollable: they thus constitute an obstacle to the integration of the solar photovoltaic and wind sources. Faced with this intermittency and random variability, several solutions are proposed in the literature:

- International exchanges, which require the interconnection of countries in the region [3].
- Energy storage, which is very expensive on a large scale and has environmental impacts from its manufacture or installation to its recycling [4, 5].
- Prediction of electricity production, which is one of the advantages offered by smart grids [6].

It is on the latter solution that we focus globally in this work. For the prediction of photo-voltaic (PV) power, two main techniques are used: the direct prediction from PV power records [7-21] and the indirect prediction that combines the prediction of meteorological parameters and mathematical models of PV power estimation [22-33]. The direct forecasting technique is the most appropriate when the system is already installed and operational because it is based on the use of PV power data and meteorological parameters of the installation site such as direct irradiation, diffuse irradiation, reflected irradiation, ambient temperature, wind speed or humidity as input to the model in addition to the power itself. For this purpose, several methods are used in the literature:

PV power forecasting as a time series: this method predicts future power values from past values. It has the advantage of being very simple to implement but has the disadvantage of not directly considering the meteorological parameters of the site.

PV power prediction from meteorological data and power readings. The contribution of meteorological parameters makes the model more complex but presents better results as shown in [15,18,34,35].

The input parameters of these models are generally a performance criterion, and a compromise must be found between the complexity of the model, the size of the input data, and the performance. Moreover, for the model to be usable for forecasting, all input parameters must be easily measurable in real time. Therefore, it is necessary to find adequate forecasting techniques that use the least number of measurable parameters possible without altering the performance of the model. We, therefore, propose in this paper a multi-step PV power forecasting technique combining a parameter reduction technique, a time series forecasting technique and regression.

The article is organized as what follows. The methods and the material are presented in Section 2. In Section 3, the results obtained and the discussion are presented. In the last section, the conclusions of the study are presented.

2. Methods

As shown in the research flowchart in figure 1, the methodology used starts with the identification of the parameters influencing the variability of photovoltaic power. Once these parameters have been identified, the data related to them is acquired or downloaded from a meteorological database. This is followed by data processing to ensure that the models to be developed fully and hopefully understand the information we present to them. After data processing, the number of parameters is reduced by retaining only those that are highly correlated with PV power. Once the study parameters have been selected, the hyper-parameters of the models are optimized in order to obtain ones adapted to our study case. Then the models are developed (training, validation, and testing). Once the models have been developed, the parameters that have contributed most to their formation are retained, and the models are then re-trained with these parameters to check that their performance has not been impaired.

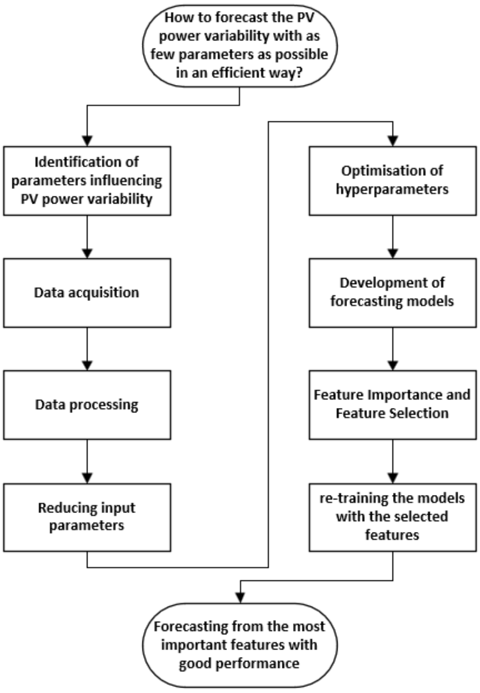

How to forecast the PV power variability with as few parameters as possible in an efficient way?
Identification of parameters influencing PV power variability
Data acquisition
Data processing
Reducing input parameters
Optimisation of hyperparameters
Development of forecasting models
Feature Importance and Feature Selection
re-training the models with the selected features
Forecasting from the most important features with good performance

Figure 1. Research flowchart.

2.1. Input parameters for PV power forecasting

The inputs to forecasting models have a direct influence on forecasting accuracy; it is a key factor in determining model performance. In general, careless selection of inputs can lead to forecasting errors that increase lead time, cost, and computational complexity. In the literature, several weather parameters have been identified as affecting the power output of a PV system. The use of meteorological variables is particularly useful when the data is irregular as it mitigates the effect of irregularity on the forecasting performance of the model. For example, in [14], Zhang *et al*. used only the historical PV power data as input to forecast the regular component, while for the irregular component, the historical PV power data in addition to solar radiation intensity and air temperature are input to generate the forecasting. According to [36], the variables that are strongly positively correlated with PV power output are solar irradiance, air temperature, and dew point, while relative humidity and cloud type have a rather negative correlation. In [31], the four most important features for forecasting global horizontal irradiance according to Pearson's correlation values are temperature, clear sky index, relative humidity, and time of day, while pressure, wind speed, and direction are less important.

In general, the most common meteorological parameters for PV power forecasting are irradiance (direct, diffuse, reflected), ambient temperature, relative humidity, wind speed, sky clarity index, sun position, and dew point in addition to the power itself. It should be remembered that our objective is to develop an efficient PV power prediction model based on as few easily measurable parameters as possible. For this purpose, we have chosen the most commonly used and measurable parameters: irradiation, ambient temperature, wind speed, and sun position or hours of the day.

2.2. Forecasting technique

In order to achieve the objective, the proposed forecasting technique is presented in figure 2. It combines a time series forecasting model to forecast for a given horizon h, the input meteorological parameters, and a regression model to estimate for each step of the forecast horizon the PV power.

Indeed, there are very powerful regression models but they are generally developed to perform one-step forecasts. The time series forecasting model is, therefore, used here to extend the number of forecasting steps, while taking advantage of the performance of regression models.

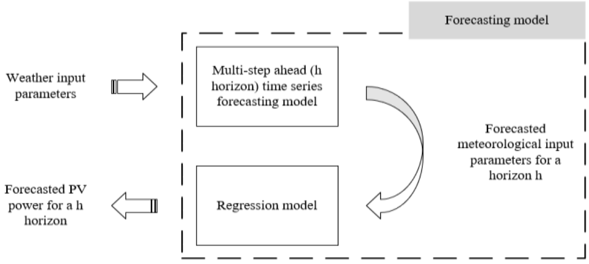

Forecasting model
Weather input parameters
Multi-step ahead (h horizon) time series forecasting model
Forecasted meteorological input parameters for a horizon h
Regression model
Forecasted PV power for a h horizon

Figure 2. Multi-step ahead forecasting concept implemented.

For the meteorological parameter forecasting, we opted for deep learning models because of their ability to learn large amounts of data. In the literature, several deep learning models are encountered in weather time series forecasting tasks: the CNN (Convolutional Neural Network) model [8, 9], the LSTM (Long Short-Term Memory) model [10-13], [22-27], [34], the GRU (Gated Recurrent Unit) model [24, 37] or the DBN (Deep Belief Network) model [28]. Similarly, for regression tasks for PV power prediction, the authors have made use of several models in the literature including the random forest [38-40], decision trees [40-42], XGBoost (eXtreme Gradient Boosting) [40], [43-47], neural networks [48-52].

Based on various research results from the literature, we select for this study the LSTM and CNN models for forecasting the input meteorological parameters. In one of our works [40] entitled "comparative study of decision tree, random forest, and XGBoost performance in forecasting the power output of a photo-voltaic system", we performed a comparative study of three models for forecasting the power output of a photovoltaic system based on meteorological data such as solar irradiance, air temperature at 2 m from the ground, wind speed and sun position. We concluded that the best performing prediction model is the one based on eXtreme Gradient Boosting (XGBoost). Therefore, we retain the XGBoost model for the regression model in this study.

2.3. Data acquisition and processing

2.3.1. Data acquisition and parameters

Once the influence parameters have been selected, they must be acquired over a substantial study period and then processed. The data we used for the forecasting concept application was taken for the city of Abomey-Calavi located in the South of

Benin. Their main characteristics are presented below:

- Geographic location of the site:
  - Latitude : 6.4842
  - Longitude : 2.3521
- Period: January 1st 2005 to December 31st 2020
- Database: PVGIS [53]
- Weather parameters and PV powers:
  - Direct Irradiation (W/m^2);
  - Wind speed at 10 m from the ground (m/s);
  - Ambient temperature at 2 m from the ground (°C);
  - Position of the sun (°);
  - Measured power from a 1 kW peak panel (W);
- PV Slope: 10°.
- Azimuth: 22°.
- Nominal power of the PV system (c-Si) (kWp): 1.0
- System losses (%): 14.0

2.3.2. Inspect and clean up outliers

It is important to eliminate all outliers and to complete those that are not available in the database, perhaps by default. Here, we can use statistical calculation methods such as mean, standard deviation, maximum, minimum to get an idea of a probable problem. We can also represent the data on time intervals to observe probable anomalies. Regarding missing data, there are several techniques applied depending on the importance and size of the missing data. The disadvantage of these approaches is that one can only interpolate those variables that have well distributed missing values. Interpolation gives very bad results when the gap of missing values is high. These data can also be regressed with a correlated variable. But when none of these methods seems to be efficient, we can keep these outliers and use an adequate model.

2.3.3. Feature creation

Before building a forecasting model, it is important to understand the data and to ensure that the model is given correctly formatted data and that it will learn what it wants to learn. As shown in figure 1, the first forecasting step in this work is the forecasting of the input data as time series. The data exploited here (solar irradiation, temperature, wind speed ...), are time series and it is important that they have or are given certain stochastic characteristics before exploiting them. In particular, they must be stationary, i.e. their statistics must not depend on time. In addition to the stationarity, it is necessary to be able to inculcate the notions of time in the model. Since it is weather data, it has a clear daily and annual periodicity. There are many ways to handle periodicity. We can obtain usable signals by using sine and cosine transforms as in equations (1) and (2) to derive new inputs for our dataset.

$$\boldsymbol{x_{sin}} = \mathbf{sin}\left(\boldsymbol{t} * \left(\frac{\mathbf{2} * \boldsymbol{\pi}}{\boldsymbol{T}}\right)\right) \tag{1}$$

$$\boldsymbol{x_{cos}} = \mathbf{cos}\left(\boldsymbol{t} * \left(\frac{\mathbf{2} * \boldsymbol{\pi}}{\boldsymbol{T}}\right)\right) \tag{2}$$

where t is the time in second, x_{sin}, x_{cos}, the new variables added to the dataset and T a significant period in the time series.

When no information is known in advance about the periodicity of the series, it is possible to determine which frequencies are important by extracting features with the fast Fourier transform. Adding these patterns allows the model to access the characteristics of the most important frequencies.

2.3.4. Dataset splitting

We divide the dataset into a training-set and a test-set. The training set is the fraction of the dataset we use to train the models, and the test set is the one we use to evaluate the performance of our models. Thus we used 70% of the data for training and the remaining 30% for testing.

2.3.5. Feature scaling

In most cases, we are working with datasets whose features are not on the same scale. Some features often have huge values, and others have small values. Thus it is better to scale them to the same scale. There are basically two ways to do this: standardization and normalization. In our work we have exploited normalization which consists in scaling the data so that they are bounded between $[a, b]$. The $max(x)$ will be equal to b and the $min(x)$ will be equal to a. It consists in subtracting, for each variable, the minimum value and dividing the result by the maximum deviation encountered:

$$\mathbf{x_{norm}} = \mathbf{a} + \frac{(\mathbf{x} - \mathbf{min}\ (\mathbf{x}))(\mathbf{b} - \mathbf{a})}{\mathbf{max}(\mathbf{x}) - \mathbf{min}\ (\mathbf{x})} \tag{3}$$

2.4. Xgboost feature importance and selection

In order to reduce the number of input parameters, reduction and selection techniques are used. For this study, we have essentially based ourselves on two methods of calculation which are:

- Feature importance built-in the Xgboost algorithm
- Feature importance computed with SHAP value

2.4.1. Feature importance built-in the Xgboost algorithm

The XGBoost model was developed in Python with version 1.3.3 of the XGBoost library. This library has the feature_importances_attribute, which is used here. The importance matrix is presented as a table with the first column containing the names of all the features effectively used in the boosted trees and in the other columns the resulting "importance" values calculated with different importance metrics. The different metrics available are [54]:

- the **gain** that implies the relative contribution of the corresponding feature to the model, calculated by taking the contribution of each feature for each tree in the model. A higher value of this metric relative to another feature implies that it is more important in the prediction;
- the **coverage** metric, which represents the relative number of observations related to each feature;
- **Frequency (R)/weight**, which is the percentage representing the relative number of times a particular feature appears in the model trees.

2.4.2. Feature importance computed with SHAP value

A second method for calculating the importance of features in Xgboost that we have exploited is the use of the Python’s SHAP package. It is model-independent and uses shapley values from game theory to estimate the contribution of each feature to the prediction.

2.5. Multi-steps ahead weather parameters forecasting as time series

Once the number of parameters was reduced and scaled, we moved on to the forecasting of the weather parameters selected for a h horizon. To do so, we tested two models of time series forecasting: convolutional neural networks (CNN) and long short-term memory (LSTM). Figure 3 presents the forecasting concept.

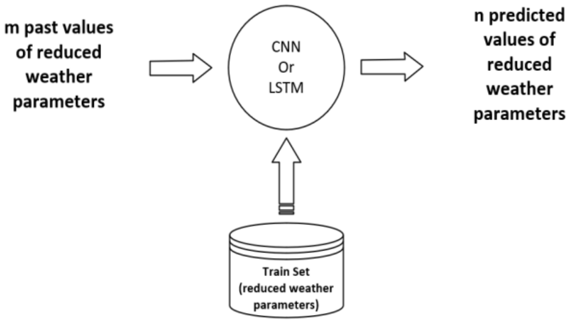

m past values of reduced weather parameters
CNN Or LSTM
n predicted values of reduced weather parameters
Train Set (reduced weather parameters)

Figure 3. Multi-step ahead time series forecasting.

The two tested models (CNN and LSTM) were trained with TensorFlow 2.8.0 under Python 3.9.7. In order to find the optimal parameters for each model, we exploit the KerasTuner library, which is a scalable and easy to use hyper-parameter optimization framework that solves hyperparameter search problems.

The parameters used for the training of the CNN are:

- Input size: 72
- Label width: ranged from 24 to 72 in steps of 24.
- Filters: ranged from 64 to 320 in steps of 64.
- Units: ranged from 500 to 1000 in steps of 100.
- Learning rate: {10^{-2}, 10^{-3}, 10^{-4}}.

The parameters used for the training of the LSTM are:

- Input size: 72.
- Label width: ranged from 24 to 72 in steps of 24.
- Units: ranged from 500 to 1000 in steps of 100.
- Learning rate: {10^{-2}, 10^{-3}, 10^{-4}}.

2.6. Multi-steps ahead PV power forecasting

Once the weather parameters are forecasted for an h-horizon, they can now be used to make the PV power forecast using the XGBoost model, as shown in figure 4.

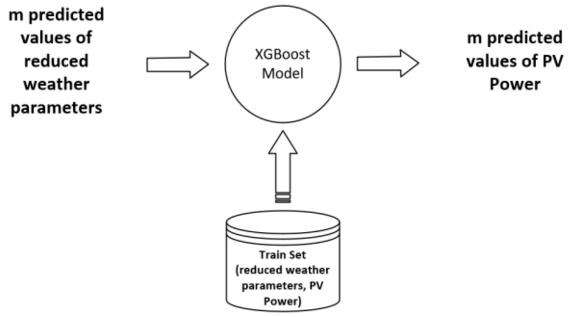

m predicted values of reduced weather parameters
XGBoost Model
m predicted values of PV Power
Train Set (reduced weather parameters, PV Power)

Figure 4. PV Power forecasting with XGBoost model.

The parameters of the XGBoost model are given below in table 1. This model is trained with the XGBRegressor library under Python.

Table 1. XGBoost model parameters.

Parameters	Values
base_score	None
booster	gbtree
colsample_bylevel	None
colsample_bynode	None
colsample_bytree	None
enable_categorical	False
gamma	None
gpu_id	None
importance_type	None
interaction_constraints	None
learning_rate	0.01
max_delta_step	None
max_depth	100
min_child_weight	None
verbosity	None
missing	nan
monotone_constraints	None
n_estimators	500
n_jobs	None
num_parallel_tree	None
objective	'reg:linear
predictor	None
random_state	None
reg_alpha	None
reg_lambda	None
scale_pos_weight	None
subsample	None
tree_method	None
validate_parameters	None

2.7. Experimental validation

Once the two models were trained, an experimental validation of the predicted meteorological parameters was carried out by comparing them to values measured for the same moments of the day on site. We performed four tests for different days. To do this we installed a weather station capable of measuring direct irradiation, ambient temperature and wind speed.

The irradiation, temperature and wind speed sensors provided by this station are shown in Figures 5, 6 and 7 respectively and their fundamental characteristics are presented in tables 2, 3, and 4, respectively.

Figure 5. Solar irradiance sensor RK200-04 [55].

Figure 6. Atmospheric RTD temperature, humidity, & pressure sensor RK330-01B [56].

Figure 7. Plastic wind speed sensor wind anemometer RK100-02 [57].

Table 2. RK04-200 specifications [55].

Item	Specifications
Spectral range	300～1100 nm
Range	0-1500 W/m^2
Resolution	1 W/m^2
Output	0-5 V,4-20 mA, RS485
Response time	≤5 s
Cosine correction	≤±10% (solar elevation angle = 10)
Non-linear	≤±3%
Temperature effect	±0.08%/℃
Stability	≤±2%/year
Operating temperature	-40 ° C - +80 ° C
Ingress protection	IP65

Table 3. RK330-01B specifications [56].

Item	Specifications
Temperature range	-40-60 ˚C
Resolution	0.1 ˚C
Accuracy	±0.1 ˚C
Output	4-20 mA, 0-5 V, 0-10 V, RS485(MODBUS), IIC
Operating temperature	-40 ° C - +80 ° C
Ingress protection	IP65

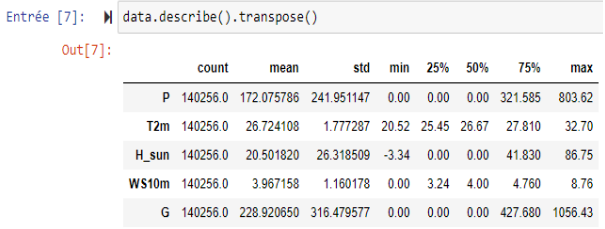
Table 4. RK330-01B specifications [57].

Item	Specifications
Speed range	0 – 45 m/s
Resolution	0.1 m/s
Limit wind speed	50 m/s
Accuracy	±(0.3+0.03V) m/s; (V is the current wind speed)
Starting threshold	<0.5 m/s
Operating temperature	-40 ° C - +50 ° C
Ingress Protection	IP65

3. Results and discussion

3.1. Data exploration

In order to make sure that there are no anomalies in the dataset, a very simple way to do it is to check the main characteristics of each variable. As can be seen in figure 8 where P, T2m, H_sun, WS10m, and G are, respectively, the panel power, the temperature at 2 m from the ground, the sun position, the wind speed at 10 m from the ground and the direct irradiation, the minimum and maximum values, the averages, the standard deviations show that there are no clear anomalies.



```
Entrée [7]:  data.describe().transpose()
Out[7]:
```

	count	mean	std	min	25%	50%	75%	max
P	140256.0	172.075786	241.951147	0.00	0.00	0.00	321.585	803.62
T2m	140256.0	26.724108	1.777287	20.52	25.45	26.67	27.810	32.70
H_sun	140256.0	20.501820	26.318509	-3.34	0.00	0.00	41.830	86.75
WS10m	140256.0	3.967158	1.160178	0.00	3.24	4.00	4.760	8.76
G	140256.0	228.920650	316.479577	0.00	0.00	0.00	427.680	1056.43

Figure 8. Dataset statistics.

3.2. Feature importance and selection

The histogram presented in figure 9 shows the participation of each variable in the formation of the XGBoost model obtained with the feature Importance built-in the XGboost algorithm.

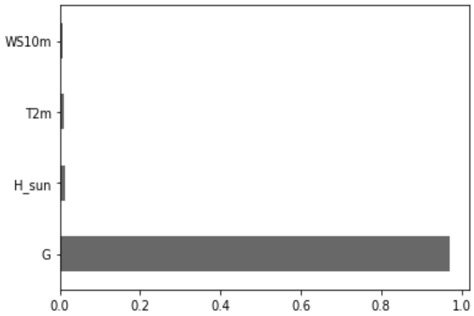


Figure 9. Feature importance.

It can be noted that the irradiation alone contributes to a very large extent to the formation of the model. We thus retain it for the rest of the work.

3.3. Feature creation

Based on the fast Fourier transform, we can highlight the most important frequencies of a time series, here the irradiation in our case.

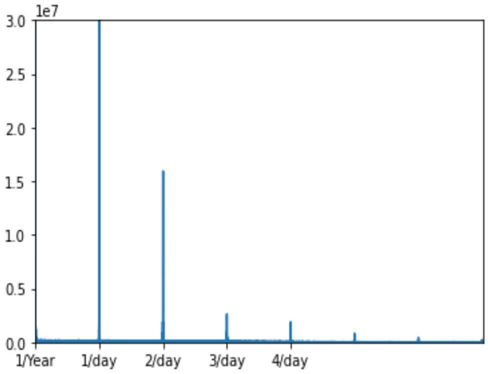
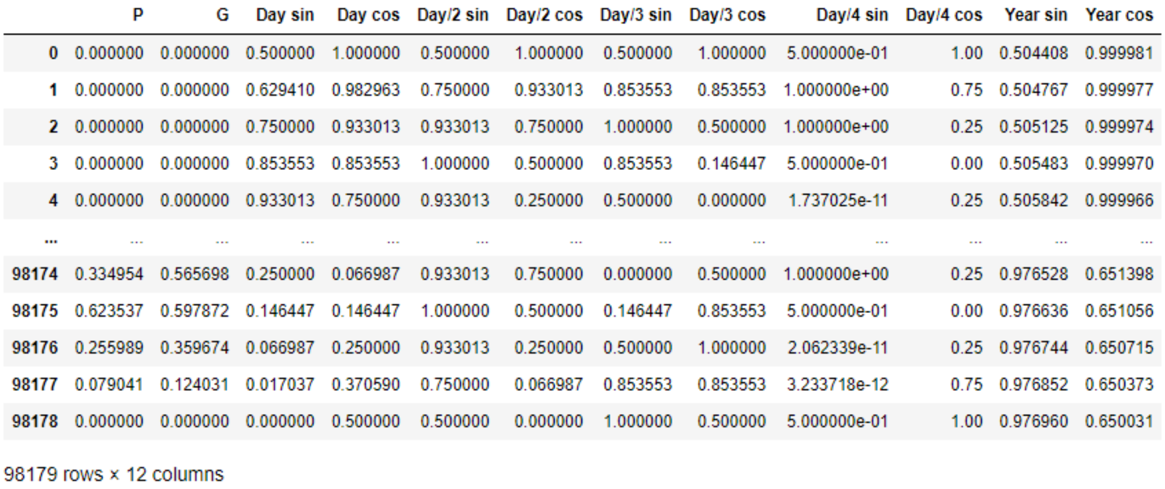


Figure 10. Irradiation fast Fourier transform.

As we can see in figure 10, in addition to the known daily and annual cycles for irradiation we also note some for the half day, third day and quarter day, which are as important as the annual cycle. As a result, we keep the daily, annual, half-day, third-day and quarter-day cycles for the creation of new variables. This gives the model access to the most important frequency characteristics.

3.4. Feature scaling

We have performed here the normalization to an interval [0; 1] by exploiting equation (3). We stress that the normalization parameters are determined with the train set and then applied to the test set. Figure 11 now shows the header of our dataset.



	P	G	Day sin	Day cos	Day/2 sin	Day/2 cos	Day/3 sin	Day/3 cos	Day/4 sin	Day/4 cos	Year sin	Year cos
0	0.000000	0.000000	0.500000	1.000000	0.500000	1.000000	0.500000	1.000000	5.000000e-01	1.00	0.504408	0.999981
1	0.000000	0.000000	0.629410	0.982963	0.750000	0.933013	0.853553	0.853553	1.000000e+00	0.75	0.504767	0.999977
2	0.000000	0.000000	0.750000	0.933013	0.933013	0.750000	1.000000	0.500000	1.000000e+00	0.25	0.505125	0.999974
3	0.000000	0.000000	0.853553	0.853553	1.000000	0.500000	0.853553	0.146447	5.000000e-01	0.00	0.505483	0.999970
4	0.000000	0.000000	0.933013	0.750000	0.933013	0.250000	0.500000	0.000000	1.737025e-11	0.25	0.505842	0.999966
...	...	...	...	...	...	...	...	...	...	...	...	...
98174	0.334954	0.565698	0.250000	0.066987	0.933013	0.750000	0.000000	0.500000	1.000000e+00	0.25	0.976528	0.651398
98175	0.623537	0.597872	0.146447	0.146447	1.000000	0.500000	0.146447	0.853553	5.000000e-01	0.00	0.976636	0.651056
98176	0.255989	0.359674	0.066987	0.250000	0.933013	0.250000	0.500000	1.000000	2.062339e-11	0.25	0.976744	0.650715
98177	0.079041	0.124031	0.017037	0.370590	0.750000	0.066987	0.853553	0.853553	3.233718e-12	0.75	0.976852	0.650373
98178	0.000000	0.000000	0.000000	0.500000	0.500000	0.000000	1.000000	0.500000	5.000000e-01	1.00	0.976960	0.650031

98179 rows × 12 columns

Figure 11. Dataset presentation after normalization and new features creation.

3.5. Performance of trained models

Table 5 summarizes the performance of the trained LSTM models as well as the optimum parameters for which this performance was obtained.

We can note that the best performing model among the LSTM models is model 1 for a 24 h forecast.

Table 5. Summary of LSTM trained model performance.

	Model 1	Model 2	Model 3
Input width	72	72	72
Label width	24	48	72
Units	800	700	500
Learning rate	0.001	0.001	0.01
Test RMSE (Root mean square error)	**3.66 W/m^2**	**3.85 W/m^2**	**3.88 W/m^2**

Table 6 summarizes the performance of the trained CNN models as well as the optimum parameters for which this performance was obtained.

Table 6. Summary of CNN trained model performance.

	Model 1	Model 2	Model 3
Input width	72	72	72
Label width	24	48	72
Units	700	500	600
Filters	192	256	256
Learning rate	0.001	0.01	0.01
Test RMSE (root mean square error)	**3.94 W/m^2**	**3.95 W/m^2**	**4.007 W/m^2**

We can note that the best performing model among the CNN models is model 2 for a 48 h forecast.

For the same forecasting range, the LSTM models perform better than the CNN models. The best performing models are those for which a 24-hour forecast has been made. Therefore, we retain the LSTM model with 24 hours of forecast output.

Figure 12 shows the prediction results of the selected model on three portions taken randomly from the Test set.

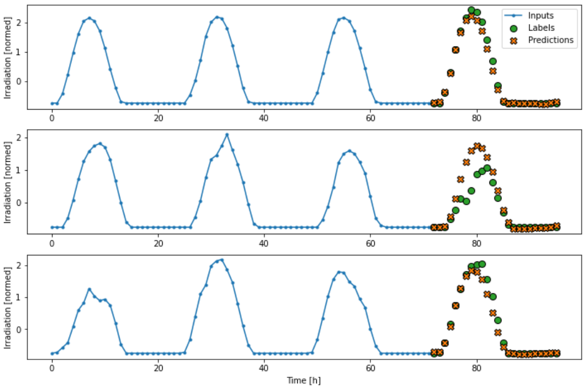


Figure 12. Forecasting results of the selected model.

3.6. Experimental validation

In order to test the effectiveness of the selected LSTM model for irradiation prediction, we performed on-site irradiation measurements from February 13, 2022 to February 28, 2022. Four different tests were performed as shown in table 7. For each test it is specified the three days (72 h) for which the data was used as input, the fourth day (24h) following for which the data are forecasted, the weather condition of the fourth day.

Table 7. Test day details and specifications.

	Input days	Output day	Weather conditions
Test 1	February 13, 14, 15	February 16	Sunny day
Test 2	February 16, 17, 18	February 19	Sunny day
Test 3	February 20, 21, 22	February 23	Partly cloudy sky
Test 4	February 24, 25, 26	February 27	Partly cloudy sky

Figure 13 shows the data collected over the test period.

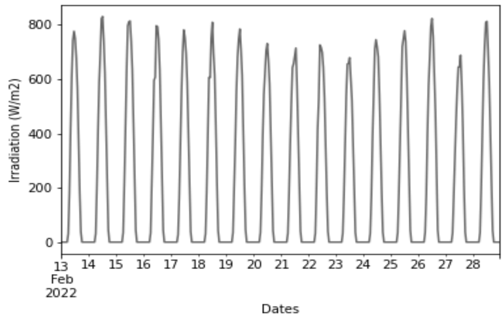


Figure 13. Irradiation data recorded from February 13 to 28, 2022.

Figures 14, 15, 16, and 17 show, respectively, for the days of February 16, 19, 23, and 27, 2022 the measured irradiations and the predicted irradiations.

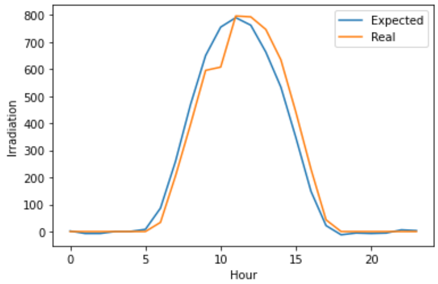


Figure 14. Measured and predicted irradiations for 16 February 2022.

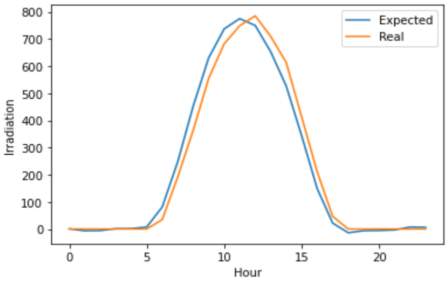


Figure 15. Measured and predicted irradiations for 19 February 2022.

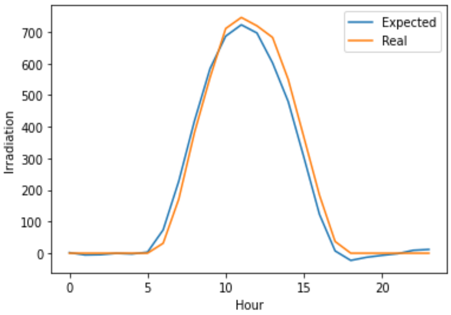


Figure 16. Measured and predicted irradiations for 23 February 2022.

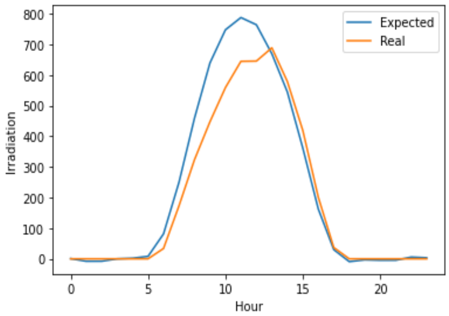


Figure 17. Measured and predicted irradiations for 27 February 2022.

Table 8 shows the root mean square errors (RMSE) as well as the regression coefficients (R) for the four trials.

Table 8. Forecast performance for the four days of testing.

	RMSE(W/m^2)	R
Forecast for February 16, 2022.	54.2	0.97
Forecast for February 19, 2022.	42.9	0.98
Forecast for February 23, 2022.	35.3	0.98
Forecast for February 27, 2022.	76.3	0.93

We can note that globally the predicted irradiations are quite close to those measured, which is confirmed by the root mean square errors and regression coefficients. Nevertheless, we can note that the largest deviations are obtained around noon and when the sky is cloudy as in the day of February 27.

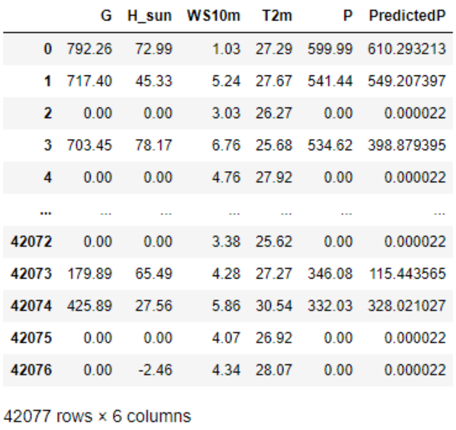
3.7. PV power estimation with XGBoost model

Table 9 shows the training performance of the XGBoost model with irradiation, wind speed, sun position, temperature as input, and then with irradiation alone as input.

From this table, we can notice that the performance of the model with only the input irradiation remains quite good although it has decreased. Figures 18 and 19 show some forecasts made with the two models.

Table 9. Test performance of the XGBoost model for the two tasks.

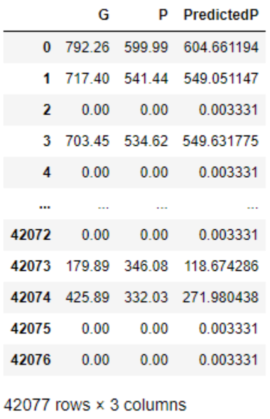
	Test RMSE (W)	Test R
XGBoost model with irradiation, wind speed, sun position, temperature as input	1.58	0.999998
XGBoost model with irradiation as input	1.72	0.992129



	G	H_sun	WS10m	T2m	P	PredictedP
0	792.26	72.99	1.03	27.29	599.99	610.293213
1	717.40	45.33	5.24	27.67	541.44	549.207397
2	0.00	0.00	3.03	26.27	0.00	0.000022
3	703.45	78.17	6.76	25.68	534.62	398.879395
4	0.00	0.00	4.76	27.92	0.00	0.000022
...	...	...	...	...	...	...
42072	0.00	0.00	3.38	25.62	0.00	0.000022
42073	179.89	65.49	4.28	27.27	346.08	115.443565
42074	425.89	27.56	5.86	30.54	332.03	328.021027
42075	0.00	0.00	4.07	26.92	0.00	0.000022
42076	0.00	-2.46	4.34	28.07	0.00	0.000022

42077 rows × 6 columns

Figure 18. Some power values predicted with the XGBoost with irradiation, wind speed, sun position, and temperature as input.



	G	P	PredictedP
0	792.26	599.99	604.661194
1	717.40	541.44	549.051147
2	0.00	0.00	0.003331
3	703.45	534.62	549.631775
4	0.00	0.00	0.003331
...	...	...	...
42072	0.00	0.00	0.003331
42073	179.89	346.08	118.674286
42074	425.89	332.03	271.980438
42075	0.00	0.00	0.003331
42076	0.00	0.00	0.003331

42077 rows × 3 columns

Figure 19. Some power values predicted with the XGBoost with irradiation as input.

Table 10 shows the comparison results of the developed forecasting technique with two forecasting techniques directly from the LSTM and CNN models with direct irradiation and PV power as inputs for past time points and PV power as output for future time points. The same dataset was used.

Table 10. Forecasting performance of LSTM-XGBoost, LSTM, and CNN models for PV power forecasting.

	Test RMSE (W)	Test R
LSTM-XGBoost model	**1.72**	0.9921
CNN model	2.15	0.96
LSTM model	1.82	**0.9934**

This comparison table shows us that the developed LSTM-XGBoost model commits globally less error than the LSTM and CNN models. Nevertheless, it should be noted that the LSTM model has a better coefficient of determination than the LSTM-XGBoost and CNN models: it would therefore produce more forecasts close to the true values but remains globally less accurate than the LSTM-XGBoost model.

The XGBoost model with only the irradiance as input was then used to estimate the PV power for the four days of testing with the irradiance predicted with the LSTM model as input. The prediction results are shown in figure 20.

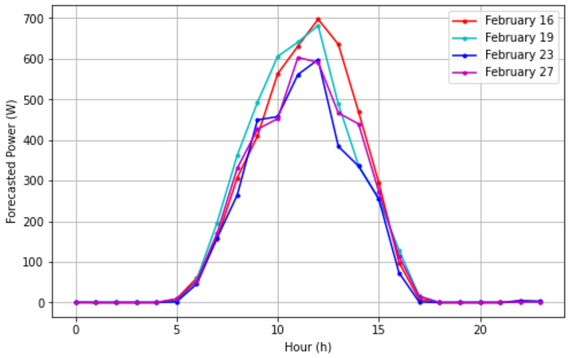


Figure 20. PV power forecast for the four test days.

4. Conclusion

The results of this work have shown that it is possible to extend the forecasting horizon of the XGBoost model by combining it with a time series forecasting model such as LSTM (1 h to 24 h) with good performance. Nevertheless, it should be noted that the forecast performance is biased towards noon when the sky is cloudy. This type of forecast can be performed anywhere by simply measuring the irradiance, but requires a history of photovoltaic power to drive the model. The proposed method, compared to most models found in the literature, uses fewer input parameters (only direct irradiation), which would facilitate its deployment at lower cost. The comparison of the proposed model with the LSTM and CNN models which are two of the most used models in PV power forecasting showed that the LSTM-XGBoost model was better in terms of accuracy (an RMSE of 1.72 W against 2.15 W for the CNN and 1.82 W for the LSTM). Nevertheless, it should also be noted that the developed LSTM-XGBoost model has a more complex architecture than the LSTM and CNN models used alone, which could increase its computation time. Furthermore, the LSTM-XGBoost model compared to the LSTM and CNN models used alone showed a lower regression coefficient than the LSTM (0.9921 versus 0.9934).

5. List of abbreviations

CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
R	Regression Coefficient
RES	Renewable Energy Sources
RMSE	Root Mean Square Error
PV	Photo-voltaic
SHAP	SHapley Additive exPlanations
XGBOOST	eXtreme Gradient Boosting

6. References

[1] IEA, “Africa Energy Outlook 2019 – Analysis,” IEA. https://www.iea.org/reports/africa-energy-outlook-2019 (accessed Feb. 07, 2022).

[2] IRENA, GIZ, and KFW, “La transition vers les énergies renouvelables en Afrique : Renforcer l’accès, la résilience et la prospérité,” 2021. [Online]. Available: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2013/Afrique_nergies_renouvelables.pdf

[3] D. Grand, A. Latrobe, C. Le Brun, and R. Vidil, “La transition énergétique sous contrainte de gestion de l’intermittence des énergies renouvelables”.

[4] A. FOPAH-LELE, “Technologie de stockage d’énergie pour les infrastructures énergétiques en Afrique subsaharienne: L’hydrogène comme perspective!,” Énerg. DURABLE EN Afr. Initiat., p. 68.

[5] C. Glaize, “Energies renouvelables et gestion du stockage de l’énergie: une necessité? Etat actuel et developpement futurs,” in Vol. 3éme conférence internationale DERBI, Perpignan juin, 2008.

[6] O. Ammar, “«Smart Grid» Réseau Electrique Intelligent,” 2017.

[7] M. Abarkan, N. K. M’Sirdi, and F. Errahimi, “MODELISATION ET SIMULATION D’UN SYSTEME DE PRODUCTION D’ENERGIE RENOUVELABLE MULTI-SOURCES ET MULTI-UTILISATEURS,” Rev. Méditerranéenne Télécommunications, Vol. 4, No. 2, 2014.

[8] Y. Sun, G. Sz\Hucs, and A. R. Brandt, “Solar PV output prediction from video streams using convolutional neural networks,” Energy Environ. Sci., Vol. 11, No. 7, pp. 1811–1818, 2018.

[9] C.-J. Huang and P.-H. Kuo, “Multiple-Input Deep Convolutional Neural Network Model for Short-Term Photovoltaic Power Forecasting,” IEEE Access, Vol. 7, pp. 74822–74834, 2019, doi: 10.1109/ACCESS.2019.2921238.

[10] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, and Y. Du, “Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention Mechanism,” IEEE Access, Vol. 7, pp. 78063–78074, 2019, doi: 10.1109/ACCESS.2019.2923006.

[11] L. Wen, K. Zhou, S. Yang, and X. Lu, “Optimal load dispatch of community microgrid with deep learning based solar power and load forecasting,” Energy, vol. 171, pp. 1053–1065, Mar. 2019, doi: 10.1016/j.energy.2019.01.075.

[12] M. Abdel-Nasser and K. Mahmoud, “Accurate photovoltaic power forecasting models using deep LSTM-RNN,” Neural Comput. Appl., Vol. 31, No. 7, pp. 2727–2740, Jul. 2019, doi: 10.1007/s00521-017-3225-z.

[13] H. Sharadga, S. Hajimirza, and R. S. Balog, “Time series forecasting of solar power generation for large-scale photovoltaic plants,” Renew. Energy, Vol. 150, pp. 797–807, May 2020, doi: 10.1016/j.renene.2019.12.131.

[14] J. Zhang, Z. Tan, and Y. Wei, “An adaptive hybrid model for day-ahead photovoltaic output power prediction,” J. Clean. Prod., Vol. 244, p. 118858, 2020.

[15] M. Gao, J. Li, F. Hong, and D. Long, “Day-ahead power forecasting in a large-scale photovoltaic plant based on weather classification using LSTM,” Energy, Vol. 187, p. 115838, Nov. 2019, doi: 10.1016/j.energy.2019.07.168.

[16] G. W. Chang and H.-J. Lu, “Integrating Gray Data Preprocessor and Deep Belief Network for Day-Ahead PV Power Output Forecast,” IEEE Trans. Sustain. Energy, Vol. 11, No. 1, pp. 185–194, Jan. 2020, doi: 10.1109/TSTE.2018.2888548.

[17] F. Wang, Z. Xuan, Z. Zhen, K. Li, T. Wang, and M. Shi, “A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework,” Energy Convers. Manag., Vol. 212, p. 112766, 2020.

[18] B. Ray, R. Shah, Md. R. Islam, and S. Islam, “A New Data Driven Long-Term Solar Yield Analysis Model of Photovoltaic Power Plants,” IEEE Access, Vol. 8, pp. 136223–136233, 2020, doi: 10.1109/ACCESS.2020.3011982.

[19] G. Li, S. Xie, B. Wang, J. Xin, Y. Li, and S. Du, “Photovoltaic Power Forecasting With a Hybrid Deep Learning Approach,” IEEE Access, Vol. 8, pp. 175871–175880, 2020, doi: 10.1109/ACCESS.2020.3025860.

[20] P. Li, K. Zhou, X. Lu, and S. Yang, “A hybrid deep learning model for short-term PV power forecasting,” Appl. Energy, vol. 259, p. 114216, 2020.

[21] J. Ospina, A. Newaz, and M. O. Faruque, “Forecasting of PV plant output using hybrid wavelet-based LSTM-DNN structure model,” IET Renew. Power Gener., Vol. 13, No. 7, pp. 1087–1095, 2019, doi: 10.1049/iet-rpg.2018.5779.

[22] A. Alzahrani, P. Shamsi, C. Dagli, and M. Ferdowsi, “Solar Irradiance Forecasting Using Deep Neural Networks,” Procedia Comput. Sci., Vol. 114, pp. 304–313, Jan. 2017, doi: 10.1016/j.procs.2017.09.045.

[23] Z. Pang, F. Niu, and Z. O’Neill, “Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons,” Renew. Energy, Vol. 156, pp. 279–289, Aug. 2020, doi: 10.1016/j.renene.2020.04.042.

[24] M. C. Sorkun, C. Paoli, and Ö. D. Incel, “Time series forecasting on solar irradiation using deep learning,” in 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), Nov. 2017, pp. 151–155.

[25] X. Qing and Y. Niu, “Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM,” Energy, Vol. 148, pp. 461–468, Apr. 2018, doi: 10.1016/j.energy.2018.01.177.

[26] M. A. F. B. Lima, P. C. M. Carvalho, L. M. Fernández-Ramírez, and A. P. S. Braga, “Improving solar forecasting using Deep Learning and Portfolio Theory integration,” Energy, Vol. 195, p. 117016, Mar. 2020, doi: 10.1016/j.energy.2020.117016.

[27] V. Suresh, P. Janik, J. M. Guerrero, Z. Leonowicz, and T. Sikorski, “Microgrid Energy Management System With Embedded Deep Learning Forecaster and Combined Optimizer,” IEEE Access, Vol. 8, pp. 202225–202239, 2020, doi: 10.1109/ACCESS.2020.3036131.

[28] Y. Q. Neo, T. T. Teo, W. L. Woo, T. Logenthiran, and A. Sharma, “Forecasting of photovoltaic power using deep belief network,” in TENCON 2017 - 2017 IEEE Region 10 Conference, Nov. 2017, pp. 1189–1194. doi: 10.1109/TENCON.2017.8228038.

[29] 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.

[30] B. Gao, X. Huang, J. Shi, Y. Tai, and J. Zhang, “Hourly forecasting of solar irradiance based on CEEMDAN and multi-strategy CNN-LSTM neural networks,” Renew. Energy, Vol. 162, pp. 1665–1683, Dec. 2020, doi: 10.1016/j.renene.2020.09.141.

[31] P. Kumari and D. Toshniwal, “Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance,” J. Clean. Prod., Vol. 279, p. 123285, Jan. 2021, doi: 10.1016/j.jclepro.2020.123285.

[32] Z. Zhen et al., “Deep learning based surface irradiance mapping model for solar PV power forecasting using sky image,” IEEE Trans. Ind. Appl., Vol. 56, No. 4, pp. 3385–3396, 2020.

[33] K. Wang, X. Qi, and H. Liu, “A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network,” Appl. Energy,

Vol. 251, p. 113315, Oct. 2019, doi: 10.1016/j.apenergy.2019.113315.

[34] M. S. Hossain and H. Mahmood, “Short-Term Photovoltaic Power Forecasting using an LSTM Neural Network and Synthetic Weather Forecast,” IEEE Access, Vol. 8, pp. 172524–172533, 2020, doi: 10.1109/ACCESS.2020.3024901.

[35] K. Wang, X. Qi, and H. Liu, “Photovoltaic power forecasting based LSTM-Convolutional Network,” Energy, Vol. 189, p. 116225, Dec. 2019, doi: 10.1016/j.energy.2019.116225.

[36] R. Ahmed, V. Sreeram, Y. Mishra, and 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., Vol. 124, p. 109792, 2020.

[37] Z. Niu, Z. Yu, W. Tang, Q. Wu, and M. Reformat, “Wind power forecasting using attention-based gated recurrent unit network,” Energy, Vol. 196, p. 117081, 2020.

[38] M. Massaoudi, I. Chihi, L. Sidhom, M. Trabelsi, S. S. Refaat, and F. S. Oueslati, “Enhanced Random Forest Model for Robust Short-Term Photovoltaic Power Forecasting Using Weather Measurements,” Energies, Vol. 14, No. 13, Art. No. 13, Jan. 2021, doi: 10.3390/en14133992.

[39] D. Liu and K. Sun, “Random forest solar power forecast based on classification optimization,” Energy, Vol. 187, p. 115940, Nov. 2019, doi: 10.1016/j.energy.2019.115940.

[40] A. B. K. Didavi, R. G. Agbokpanzo, and M. Agbomahena, “Comparative study of Decision Tree, Random Forest and XGBoost performance in forecasting the power output of a photovoltaic system,” in 2021 IEEE 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART), Dec. 2021, pp. 1–5. doi: 10.1109/BioSMART54244.2021.9677566.

[41] Rahul, A. Gupta, A. Bansal, and K. Roy, “Solar Energy Prediction using Decision Tree Regressor,” in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2021, pp. 489–495. doi: 10.1109/ICICCS51141.2021.9432322.

[42] N. Singh, S. Jena, and C. K. Panigrahi, “A novel application of Decision Tree classifier in solar irradiance prediction,” Mater. Today Proc., Vol. 58, pp. 316–323, Jan. 2022, doi: 10.1016/j.matpr.2022.02.198.

[43] C. N. Obiora, A. Ali, and A. N. Hasan, “Implementing Extreme Gradient Boosting (XGBoost) Algorithm in Predicting Solar Irradiance,” in 2021 IEEE PES/IAS PowerAfrica, Aug. 2021, pp. 1–5. doi: 10.1109/PowerAfrica52236.2021.9543159.

[44] D.-J. Bae, B.-S. Kwon, and K.-B. Song, “XGBoost-Based Day-Ahead Load Forecasting Algorithm Considering Behind-the-Meter Solar PV Generation,” Energies, Vol. 15, No. 1, Art. No. 1, Jan. 2022, doi: 10.3390/en15010128.

[45] Q.-T. Phan, Y.-K. Wu, and Q.-D. Phan, “Short-term Solar Power Forecasting Using XGBoost with Numerical Weather Prediction,” in 2021 IEEE International Future Energy Electronics Conference (IFEEC), Nov. 2021, pp. 1–6. doi: 10.1109/IFEEC53238.2021.9661874.

[46] R. Gupta, A. K. Yadav, S. Jha, and P. K. Pathak, “Time Series Forecasting of Solar Power Generation using Facebook Prophet and XG Boost,” in 2022 IEEE Delhi Section Conference (DELCON), Feb. 2022, pp. 1–5. doi: 10.1109/DELCON54057.2022.9752916.

[47] X. Li et al., “Probabilistic solar irradiance forecasting based on XGBoost,” Energy Rep., Vol. 8, pp. 1087–1095, Aug. 2022, doi: 10.1016/j.egyr.2022.02.251.

[48] M. Massaoudi et al., “An Effective Hybrid NARX-LSTM Model for Point and Interval PV Power Forecasting,” IEEE Access, Vol. 9, pp. 36571–36588, 2021, doi: 10.1109/ACCESS.2021.3062776.

[49] A. R. Gilles, D. Audace, H. Aristide, O. Arouna, and E. Christophe, “Evaluation of the photovoltaic power prediction performance of a neural network based on input data,” in 2020 IEEE 2nd International Conference on Smart Cities and Communities (SCCIC), Dec. 2020, pp. 1–6. doi: 10.1109/SCCIC51516.2020.9377334.

[50] H. Nazaripouya, B. Wang, Y. Wang, P. Chu, H. R. Pota, and R. Gadh, “Univariate time series prediction of solar power using a hybrid wavelet-ARMA-NARX prediction method,” in 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), May 2016, pp. 1–5. doi: 10.1109/TDC.2016.7519959.

[51] M. Massaoudi, I. Chihi, L. Sidhom, M. Trabelsi, S. S. Refaat, and F. S. Oueslati, “A Novel Approach Based Deep RNN Using Hybrid NARX-LSTM Model for Solar Power Forecasting.” arXiv, Oct. 21, 2019. doi: 10.48550/arXiv.1910.10064.

[52] Z. Boussaada, O. Curea, A. Remaci, H. Camblong, and N. Mrabet Bellaaj, “A Nonlinear Autoregressive Exogenous (NARX) Neural Network Model for the Prediction of the Daily Direct Solar Radiation,” Energies, Vol. 11, No. 3, Art. No. 3, Mar. 2018, doi: 10.3390/en11030620.

[53] “JRC Photovoltaic Geographical Information System (PVGIS) - European Commission.” https://re.jrc.ec.europa.eu/pvg_tools/en/ (accessed Jul. 08, 2022).

[54] A. Abu-Rmileh, “Be careful when interpreting your features importance in XGBoost!,” Medium, Sep. 02, 2021. https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7 (accessed May 16, 2022).

[55] "Rk200-04 Solar Radiation Sensor Solar Irradiance Sensor | Rika Sensors." https://www.rikasensor.com/rk200-04-solar-radiation-sensor-solar-irradiance-sensor.html (accessed Jul. 23, 2022).

[56] "Rk330-01b Atmospheric Temperature, Humidity & Pressure Sensor | Rika Sensors." https://www.rikasensor.com/rk330-01b-atmospheric-temperature-humidity-pressure-sensor.html (accessed Jul. 23, 2022).

[57] "Rk100-02 Cheap Plastic Wind Speed Sensor / Detector, 3 Cup Wind Anemometer | Rika." https://www.rikasensor.com/rk100-02-wind-speed-sensor-wind-speed-detector.html (accessed Jul. 23, 2022).